WATER RESOURCES RESEARCH, VOL. 33, NO. 3, PAGES 439–448, MARCH 1997

Simulation, Bayes, and bootstrap in statistical hydrology

Vincent Fortin, Jacques Bernier, and Bernard Bobée

NSERC/Hydro-Québec Chair on Statistical Hydrology, Institut National de la Recherche Scientifique, Université du Québec, Sainte-Foy, Quebec, Canada

Abstract. Statistical simulation in hydrology is discussed from a Bayesian perspective. The inherent difficulties in both parametric simulation, based on a parent distribution, and classical nonparametric simulation, based on the bootstrap, are discussed. As an alternative to these procedures, a nonparametric Bayesian simulation methodology, Pólya resampling, is introduced. It consists of simulating from a nonparametric predictive distribution obtained from the analysis of a reference sample, and it is asymptotically equivalent to the bootstrap. The method is generalized to take into account a prior hypothesis on the parametric distribution of a variable. A hybrid simulation model is then obtained that includes parametric and nonparametric simulation as particular cases. An extensive application is presented in a related paper [*Fortin et al.*, 1997], where Pólya resampling is used to compare statistical models for flood frequency analysis. In this paper an example is used to demonstrate how Pólya resampling can help assess the influence of a distribution hypothesis on simulation results.

1. Introduction

In statistical hydrology, parametric and nonparametric simulation procedures have been extensively used to compare statistical models of hydrological variables. The purpose of this paper is to present an alternative, based on Bayesian analysis, to classical parametric and nonparametric simulation techniques for independent and identically distributed (i.i.d.) variables. The proposed method, which we call Pólya resampling, is similar to the bootstrap [*Efron*, 1979], which consists in drawing observations with replacement from a reference sample. This simulation scheme, introduced by *Lo* [1988], is discussed in detail for binomial variables, then for multinomial variables, and finally in a completely nonparametric setting. The methodology is then generalized to take into account prior information independent from the reference sample. The present paper is mainly theoretical; an illustrative example which compares the GEV and Gumbel distributions for at-site flood frequency analysis is included. An extensive application to at-site flood frequency analysis is presented in a related paper [*Fortin et al.*, 1997].

1.1. Simulation in Statistical Hydrology

A common problem in statistical hydrology is to approximate the unknown statistical distribution F of a (hydrological) random variable X, given prior information and observed data. The usual approach consists in selecting a parametric distribution having probability density function (p.d.f.) $f(x; \underline{\theta})$ which approximates the empirical distribution of X. The parameters $\underline{\theta}$ of f are estimated from the data, and a model $f(x; \hat{\underline{\theta}})$ of the true distribution F is obtained. As different parametric distributions f and different estimation methods may lead to different models for F, simulation procedures have emerged to compare estimation models. Simulation procedures fall into two categories:

1. Parametric simulation: A parent distribution $g(x; \underline{\theta})$ assumed to be sufficiently flexible to model the observations is selected. Plausible sets of parameters are determined for the parent distribution. Samples $\{\mathbf{x}_i, i = 1, 2, \cdots, R\}$ of various sizes are randomly generated for each set of parameters.
2. Nonparametric simulation: Samples $\{\mathbf{x}_i, i = 1, 2, \cdots, R\}$ of various sizes are generated by sampling from the empirical frequency distribution of a reference sample $\mathbf{y} = \{y_j, j = 1, 2, \cdots, N\}$ of i.i.d. observations of the variable X, using, for example, the bootstrap [*Efron*, 1979].

In both cases, statistical distributions and methods for estimating the parameters may be compared in terms of their ability to approximate the distribution of the simulated samples. The main problem with parametric simulation is that the parametric distribution of most hydrological variables is unknown and that the choice of the parent distribution usually is arbitrary. Although the robustness of the conclusions to the hypothesis of a given parent distribution may be studied [*Slack et al.*, 1975; *Kuczera*, 1982; *Haktanir*, 1992], parametric simulation has been criticized for comparing statistical models in an artificial setting [*Klemes*, 1986; *Potter*, 1987; *Bobée et al.*, 1993]. Nonparametric simulation based on the bootstrap may seem more attractive, since no hypothesis on the statistical distribution of the variable appears necessary to simulate data. However, when resampling with replacement from a reference sample, it is implicitly assumed that the finite reference sample is equivalent to the whole population. When the reference sample is small, this hypothesis may lead to unreasonable conclusions [*Rubin*, 1981].

Most simulation studies in statistical hydrology are parametric [see *Matalas et al.*, 1975; *Slack et al.*, 1975; *Landwehr et al.*, 1978; *Wallis and Wood*, 1985; *Kuczera*, 1982; *Ahmad et al.*, 1988; *World Meteorological Organization* (*WMO*) 1989; *Haktanir*, 1992; *Lu and Stedinger*, 1992; *Haktanir and Horlacher*, 1993; *Moon et al.*, 1993], although some nonparametric simulation studies have been reported [see *Tasker*, 1987; *Potter and Lettenmaier*, 1990; *Ashkar et al.*, 1992; *Rasmussen et al.*, 1994]. However, the authors are unaware of any attempt to address the problem of statistical simulation in hydrology from a Bayesian point of view. The objective of this paper is to intro-

Paper number 96WR03355.
0043-1397/97/96WR-03355$09.00

1. Prior analysis:
   1.1. State the parametric model $f(\mathbf{x};\underline{\theta})$ and the possible values of $\mathbf{x}\in\mathcal{X}$ and $\underline{\theta}\in\Theta$:
   for a sample $\mathbf{x}=\{x_j, j=1,2,\ldots,n\}$ of i.i.d. observations, $f(\mathbf{x};\underline{\theta})=f(x_1;\underline{\theta})\cdot f(x_2;\underline{\theta})\cdots f(x_n;\underline{\theta})$.
   1.2. Using prior information, determine the prior distribution $b(\underline{\theta})$ of the parameters $\underline{\theta}$:
   with no prior information, use a non informative (locally uniform) prior.
2. Posterior analysis (on the basis of information $\mathbf{x}$):
   2.1. Determine the action $a\in A$ implied by a statistical procedure $p(\mathbf{x})=a$ for each $\mathbf{x}$.
   2.2. Determine the loss function $l(a,\underline{\theta})$ corresponding to each action a given $\underline{\theta}$.
   2.3. Compute the predictive distribution $f(\mathbf{x})$:
   $$f(\mathbf{x}) = \int_{\Theta} f(\mathbf{x};\underline{\theta})b(\underline{\theta})d\underline{\theta}$$
   2.4. Compute the posterior distribution $b(\underline{\theta}|\mathbf{x})$:
   $$b(\underline{\theta}|\mathbf{x}) = \frac{f(\mathbf{x};\underline{\theta})\cdot b(\underline{\theta})}{f(\mathbf{x})}$$
   2.5. Compute the posterior expected cost $c(a,\mathbf{x})$ of the action a selected by $p(\mathbf{x})$:
   $$c(a,\mathbf{x}) = \int_{\Theta} l(a,\underline{\theta})b(\underline{\theta}|\mathbf{x})d\underline{\theta}$$
3. Preposterior analysis (before observing $\mathbf{x}$, on the basis of information $\mathbf{y}$):
   3.1. Compute the updated predictive distribution $f(\mathbf{x}|\mathbf{y})$:
   $$f(\mathbf{x}|\mathbf{y}) = \int_{\Theta} f(\mathbf{x};\underline{\theta})b(\underline{\theta}|\mathbf{y})d\underline{\theta}$$
   3.2. Compute the preposterior expected cost $c(p|\mathbf{y})$ of the statistical procedure $p(\mathbf{x})=a$:
   $$c(p|\mathbf{y}) = \int_{\mathcal{X}} c\big(p(\mathbf{x}),\mathbf{y}\big)f(\mathbf{x}|\mathbf{y})d\mathbf{x}$$

Figure 1. Bayesian analysis: a summary.

duce a Bayesian nonparametric simulation procedure for i.i.d. variables, to discuss its attractive features and drawbacks, and to demonstrate its usefulness in statistical hydrology.

1.2. Simulation and Bayesian Analysis

From a Bayesian point of view, assessing the performance of a statistical model by simulation is a preposterior analysis problem which can be conducted on the basis of a known parametric distribution type or by using only the information contained in a reference sample. Figure 1 summarizes the main steps of Bayesian analysis: prior, posterior, and preposterior analysis. *Box and Tiao* [1973] and *Berger* [1985] provide a more thorough presentation of Bayesian analysis. Given a statistical procedure $p(\mathbf{x})$ and an information $\mathbf{y}$, a preposterior analysis consists in evaluating the expected cost of $p(\mathbf{x})$ with respect to the predictive distribution $f(\mathbf{x}|\mathbf{y})$.

The main difficulty with preposterior analysis is to determine the predictive distribution $f(\mathbf{x}|\mathbf{y})$, which requires knowledge of the distribution function $f(\mathbf{x};\underline{\theta})$. However, the basic problem is precisely that this function is not known. To obtain robust results, one may generalize $f(\mathbf{x};\underline{\theta})$ by adding extra parameters, as suggested by *Bernier* [1993]. An alternative procedure, which will be considered in this paper, consists in sampling from a nonparametric predictive distribution [*Lo*, 1988]. Using this procedure, no hypothesis on the parent distribution is made, although a hypothesis for the distribution $b(\underline{\theta})$ used to model prior information is still necessary.

2. Bayesian Simulation of Binomial Variables

Consider n i.i.d. Bernouilli trials represented by the random variables $X_1, X_2, \cdots, X_n$, with $X_j = 1$ if an event E occurs, and $X_j = 0$ otherwise ($\bar{E}$ occurs). The number of successes $K = \sum_{j=1}^{n} X_j$ in n trials has a binomial distribution with p.d.f. $\Pr[K = k] = f(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ where $\theta = \Pr[E]$ represents the probability of success in each individual trial and $\binom{n}{k} = n!/[k!(n-k)!]$ is the binomial coefficient. The results of a Bayesian analysis depend on the prior distribution $b(\theta)$. The usual choice for $b(\theta)$ is the beta distribution $\mathcal{B}(\alpha_1, \alpha_2)$ with p.d.f. $b(\theta) \propto \theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$, which is a natural conjugate of the binomial distribution, meaning that if the prior $b(\theta)$ is the beta distribution $\mathcal{B}(\alpha_1, \alpha_2)$, then the posterior distribution $b(\theta|k)$ after n trials is also a beta distribution with modified parameters $\mathcal{B}(\alpha_1 + k, \alpha_2 + n - k)$ [*Berger*, 1985]. For a beta prior $b(\theta)$ and a binomial model $f(k|\theta)$ the predictive probability, $f(\mathbf{x})$, of observing $k = \sum_{j=1}^{n} x_j$ successes in a sample $\mathbf{x} = \{x_j, j = 1, 2, \cdots, n\}$ of n i.i.d. trials is given by (compare Figure 1):

$$f(\mathbf{x}) = \binom{n}{k}\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1+\alpha_2+n)}\frac{\Gamma(\alpha_1+k)\Gamma(\alpha_2+n-k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}$$
$$= \binom{n}{k}\frac{\alpha_1(\alpha_1+1)\cdots(\alpha_1+k-1)\cdot\alpha_2(\alpha_2+1)\cdots(\alpha_2+n-k-1)}{(\alpha_1+\alpha_2)(\alpha_1+\alpha_2+1)\cdots(\alpha_1+\alpha_2+n-1)} \quad (1)$$

where $\Gamma(z) = (z-1)!$ is the gamma function. If a reference sample $\mathbf{y} = \{y_j, j = 1, 2, \cdots, N\}$ is observed which contains $r = \sum_{j=1}^{N} y_j$ successes in N trials, the predictive distribution can be updated by replacing α_1 by $\alpha_1' = \alpha_1 + r$ and α_2 by $\alpha_2' = \alpha_2 + N - r$. If a noninformative prior distribution $\mathcal{B}(\alpha_1 = 0, \alpha_2 = 0)$ is used to model the lack of prior information, the updated parameters are $\alpha_1' = r$ and $\alpha_2' = N - r$, and the updated predictive distribution $f(\mathbf{x}|\mathbf{y})$ is given by

$$f(\mathbf{x}|\mathbf{y}) = \binom{n}{k}\cdot\frac{[r(r+1)\cdots(r+k-1)][(N-r)(N-r+1)\cdots(N-r+n-k-1)]}{N(N+1)\cdots(N+n-1)} \quad (2)$$

The above expression is based on the somewhat arbitrary choice of $\mathcal{B}(0, 0)$ as the prior distribution. Another choice for $b(\theta)$ could be the uniform distribution, $b(\theta) = 1$, which corresponds to the beta distribution $\mathcal{B}(1, 1)$. This prior distri-

bution would lead to a different predictive distribution, but the difference would be small and would be negligible for large values of N [*Box and Tiao*, 1973].

Simulation of samples from $f(\mathbf{x}|\mathbf{y})$ can be done using a technique called Pólya resampling [*Lo*, 1988], since (2) is a Pólya urn model defined by the following experiment [*Feller*, 1971]: An urn contains r white balls and $N - r$ black balls, and balls are drawn sequentially from the urn. Each time a ball is drawn, two balls of the same color as the ball that was drawn are put back into the urn. If n balls are successively drawn in this way, the probability of observing k white balls is given by (2). Indeed, consider the probability $f(x_1 = 1, x_2 = 1, \cdots, x_k = 1, x_{k+1} = 0, x_{k+2} = 0, \cdots, x_n = 0)$ of drawing successively k white balls, and then $n - k$ black balls using Pólya resampling. For the first draw the probability $f(x_1 = 1)$ of drawing a white ball is r/N, the ratio of white balls to the total number of balls, but it increases after each successive drawing of a white ball, so that $f(x_j = 1|x_1 = 1, x_2 = 1, \cdots, x_{j-1} = 1) = (r + j - 1)/(N + j - 1)$. Therefore the probability of drawing k white balls is given by

$$f(x_1 = 1, x_2 = 1, \cdots, x_k = 1) = \frac{r(r+1)\cdots(r+k-1)}{N(N+1)\cdots(N+k-1)} \tag{3}$$

After k white balls have been drawn, the probability of drawing a black ball is $(N - r)/(N + k)$, the ratio of the initial number of black balls to the total number of balls, and it increases after each successive drawing of a black ball, so that $f(x_{k+j} = 0|x_1 = 1, x_2 = 1, \cdots, x_k = 1, x_{k+1} = 0, x_{k+2} = 0, \cdots, x_{k+j-1} = 0)$ is given by $(N - r + j - 1)/(N + k + j - 1)$. Therefore the probability of drawing $n - k$ black balls after having drawn k white balls is given by

$$f(x_{k+1} = 0, x_{k+2} = 0, \cdots, x_n = 0|x_1 = 1, x_2 = 1, \cdots, x_k = 1) = \frac{(N-r)(N-r+1)\cdots(N-r+n-k-1)}{(N+k)(N+k+1)\cdots(N+n-1)} \tag{4}$$

The probability of drawing k white balls followed by $n - k$ black balls is the product of (3) and (4):

$$f(x_1 = 1, x_2 = 1, \cdots, x_k = 1, x_{k+1} = 0, x_{k+2} = 0, \cdots, x_n = 0) = \frac{[r(r+1)\cdots(r+k-1)][(N-r)(N-r+1)\cdots(N-r+n-k-1)]}{N(N+1)\cdots(N+n-1)} \tag{5}$$

It can be shown in a similar way that the probability of drawing any sequence containing k white balls and $n - k$ black balls independently of the order in which they appear is given by (5). As there are $\binom{n}{k}$ such samples, the probability of drawing k white balls in a sample of n balls is given by (2). Notice that if $n/N \ll 1$ and $k/r \ll 1$, then adding balls to the urn does not greatly modify the initial probability r/N of drawing a white ball. Hence, under these conditions, drawing balls from the urn with simple replacement (i.e., bootstrapping) is approximately equivalent to Pólya resampling (*Lo* [1988] gives a rigorous demonstration).

3. Bayesian Simulation of Multinomial Variables

The discussion in the preceding section was based on the observation of two mutually exclusive events E and $\bar{E}$. The results obtained may be generalized to $m > 2$ mutually exclusive events $E_1, E_2, \cdots, E_m$. Consider a vector $\mathbf{N}$ whose m components are random variables $N_1, N_2, \cdots, N_m$ representing, respectively, the number of occurrences of the events $E_1, E_2, \cdots, E_m$ in n i.i.d. trials. Let $\mathbf{n} = \{n_j, j = 1, 2, \cdots, m\}$ denote a realization of the random variable $\mathbf{N} = \{N_j, j = 1, 2, \cdots, m\}$. The probability of observing $\mathbf{N} = \mathbf{n}$ is given by the multinomial distribution:

$$\Pr[\mathbf{N} = \mathbf{n}] = f(\mathbf{n}|\theta_1, \theta_2, \cdots, \theta_m) = \binom{n}{n_1, n_2, \cdots, n_m} \prod_{j=1}^{m} \theta_j^{n_j} \tag{6}$$

where

$$\binom{n}{n_1, n_2, \cdots, n_m} = n!/(n_1!n_2!\cdots n_m!)$$

is the multinomial coefficient and θ_j is the probability of occurrence of E_j in each individual trial, with $\sum_{j=1}^{m} \theta_j = 1$. In a Bayesian analysis, uncertainty related to the probabilities $\underline{\theta} = \{\theta_1, \theta_2, \cdots, \theta_m\}$ is modeled by a prior distribution $b(\underline{\theta})$. It is usually chosen to be the Dirichlet distribution, $\mathcal{D}(\alpha_1, \alpha_2, \cdots, \alpha_m)$, with p.d.f. given by [*Ferguson*, 1973]

$$b(\underline{\theta}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_m)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \tag{7}$$

where $\alpha = \sum_{j=1}^{m} \alpha_j$. Notice that the beta distribution is a particular case of the Dirichlet distribution, obtained for $m = 2$. By using Bayes' formula, it can be shown that when the prior distribution, $b(\underline{\theta})$, is a Dirichlet distribution $\mathcal{D}(\alpha_1, \alpha_2, \cdots, \alpha_m)$, the posterior distribution $b(\underline{\theta}|\mathbf{n})$ is also a Dirichlet distribution with parameters $\mathcal{D}(\alpha_1 + n_1, \alpha_2 + n_2, \cdots, \alpha_m + n_m)$. The predictive distribution $f(\mathbf{n})$ is obtained from (6) and (7):

$$f(\mathbf{n}) = \binom{n}{n_1, n_2, \cdots, n_m} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \frac{\prod_{j=1}^{m} \Gamma(\alpha_j + n_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} = \binom{n}{n_1, n_2, \cdots, n_m} \frac{\prod_{j=1}^{m} [\alpha_j(\alpha_j + 1)\cdots(\alpha_j + n_j - 1)]}{\alpha(\alpha + 1)\cdots(\alpha + n - 1)} \tag{8}$$

with $\alpha_j(\alpha_j + 1)\cdots(\alpha_j + n_j - 1) = 1$ if $n_j = 0$. For a noninformative prior distribution $\mathcal{D}(\alpha_1 = 0, \alpha_2 = 0, \cdots, \alpha_m = 0)$, the updated predictive distribution $f(\mathbf{n}|\mathbf{y})$, after observing, in a reference sample $\mathbf{y}$, $r_1, r_2, \cdots, r_m$ realizations of $E_1, E_2, \cdots, E_m$ respectively, is given by

$$f(\mathbf{n}|\mathbf{y}) = \binom{n}{n_1, n_2, \cdots, n_m} \frac{\prod_{j=1}^{m} [r_j(r_j + 1)\cdots(r_j + n_j - 1)]}{N(N+1)\cdots(N+n-1)} \tag{9}$$

with $r_j(r_j + 1)\cdots(r_j + n_j - 1) = 1$ if $n_j = 0$. This distribution is a generalization of (2), and it also corresponds

to a Pólya urn model. Here, instead of black and white balls, there are m colors of balls, and initially r_j balls of color j. If after each trial two balls whose color is the same as the one that was drawn are put back into the urn, then the probability of drawing $n_1, n_2, \cdots, n_m$ balls of each of the m colors is given by (9). Again, if $n/N \ll 1$, then the initial probability of drawing a ball of a given color s will not be significantly modified by the addition of balls to the urn and will correspond approximately to the empirical frequency r_s/N observed in the reference sample **y**. Thus, when $n/N \ll 1$, the bootstrap is approximately equivalent to Pólya resampling.

4. Pólya Resampling in a Nonparametric Setting

In the preceding section it was shown that Pólya resampling is a Bayesian counterpart to the bootstrap in the multinomial case. Since any continuous distribution can be approximated by a discrete multinomial distribution, one would expect that Pólya resampling also constitutes a Bayesian counterpart to the bootstrap in a nonparametric setting. This intuitive conjecture is demonstrated in Appendix A. A nonparametric Bayesian simulation of a sample **x** based on a reference sample **y** corresponds to a Pólya urn model. Consider an urn containing N balls of different colors, corresponding to each of the N observations of the reference sample. The first observation x_1 is drawn at random from the reference sample. Two observations identical to x_1 are put back into the reference sample **y**, which then consists of $N + 1$ observations. A second observation x_2 is drawn at random from these $N + 1$ observations, and two observations identical to x_2 are put back into **y**. This process is repeated until n observations have been drawn. When $n \ll N$, this simulation process is approximately equivalent to the bootstrap. However, preference should be given to Pólya resampling in the general case because it takes into account sampling variance in **y**, whereas in using the bootstrap one assumes that the reference sample **y** is equivalent to the whole population.

For Pólya resampling to be applicable the observations in **y** need not be real numbers. They could for instance belong to a k-dimensional euclidean space representing cross-correlated variables. Such a representation may be useful in nonparametric regionalization studies to preserve cross correlations between series at different sites.

5. Nonparametric Analysis and Extrapolation

With a parametric approach to simulation it is possible to simulate data outside the range of the reference sample **y**. This is not possible with the bootstrap, and with Pólya resampling it requires prior information independent from **y**, as will be shown in this section. While this may seem to be an important limitation of nonparametric methods, it must be emphasized that extrapolating a parent distribution chosen for its ability to fit a given reference sample may lead to worse results since nothing is known about the tails of the parent distribution. As shown in Appendix B, for a quadratic loss function $l(\hat{F}, F) = \int [\hat{F}(x) - F(x)]^2\, dx$ an estimate $\hat{F}$ of the cumulative distribution F of a random variable X based on a reference sample **y** of size N is given by a mixture of F_0, one's prior idea about the distribution of X, and $F_N(x|\mathbf{y})$, the empirical distribution of **y**:

$$\hat{F}(x) = p_N F_0(x) + (1 - p_N) F_N(x|\mathbf{y}) \tag{10}$$

where p_N represents the strength of one's belief in F_0 and is related to the weight α of the evidence used to determine F_0:

$$p_N = \alpha/(\alpha + N) \tag{11}$$

Notice that α must be measured in units homogeneous to a sample size. In practice, α could be assessed by evaluating how many observations one would be ready to exchange for the evidence modeled by F_0. If no evidence is available to determine F_0, then $p_N = 0$, and if the weight of the evidence α is large compared to N, then $p_N \cong 1$. For values larger than the maximum $y_{(N)}$ of **y** and values smaller than the minimum $y_{(1)}$ of **y**, (10) simplifies to

$$\text{if } x \geq y_{(N)} \quad F_N(x|\mathbf{y}) = 1 \Rightarrow \hat{F}(x) = p_N[F_0(x) - 1] + 1$$
$$\text{if } x < y_{(1)} \quad F_N(x|\mathbf{y}) = 0 \Rightarrow \hat{F}(x) = p_N F_0(x) \tag{12}$$

Hence $\hat{F}(x)$ depends entirely on the prior guess $F_0(x)$ whenever $x < y_{(1)}$ or $x \geq y_{(N)}$. Only the size N of the reference sample is taken into account through p_N. This result is a consequence of the discrete nature of the empirical distribution $F_N(x|\mathbf{y})$. By using kernel functions, it is possible to modify $F_N(x|\mathbf{y})$ so as to obtain an empirical distribution which is continuous and positive everywhere [*Adamowski*, 1985]. This would allow extrapolations of $\hat{F}(x)$ to depend also on $F_N(x|\mathbf{y})$. However, the objective of kernel methods is to provide a smooth interpolation inside the range of observed values. Their use for extrapolation is not justified. In fact, the tails of a nonparametric density function are sensitive both to the kernel function and the smoothing factor, not to the observed data [*Lall et al.*, 1993]. Therefore (12) suggests that extrapolation of the distribution function of a hydrological variable should be based on additional independent information (historical, physical, meteorological, etc.), which does not have to be numerical. For example, knowledge of an expert can be explicitly taken into account.

6. Incorporating Prior Information in Preposterior Analysis

Simulating observations outside the range of the reference sample **y** may be accomplished with nonparametric Bayesian analysis by including a prior hypothesis on the parametric distribution of the random variable [*Bernier*, 1997]. When there is evidence of weight α independent from **y** suggesting a prior distribution F_0, the observation x_i of a sample **x** should be simulated from $\hat{F}^{(i)}$:

$$\hat{F}^{(i)}(x_i) = p_N^{(i)} F_0(x_i) + (1 - p_N^{(i)}) F_N^{(i)}(x_i)$$
$$p_N^{(i)} = \alpha/(\alpha + N + i - 1) \tag{13}$$

where $F_N^{(i)}$ is the empirical distribution of the reference sample to which the observations $\{x_1, x_2, \cdots, x_{i-1}\}$ have been added. This is proven in Appendix C. The distribution $\hat{F}^{(i)}$ corresponds to a generalized Pólya urn model. To simulate x_i, a real number u is chosen at random between 0 and 1. If $u \leq p_N^{(i)}$, then x_i is obtained from $F_0(x)$. Otherwise, an integer k is randomly selected from between 1 and $N + i - 1$. If $k \leq N$, then x_i is set equal to y_k; otherwise x_i is set equal to x_{k-N}. Notice that if the weight α of the prior hypothesis is zero, then

$p_N^{(i)} = \alpha = 0$ and the predictive distribution $\hat{F}^{(i)}$ depends only on the reference sample **y**.

In practice, F_0 could be chosen from a family of parametric distributions whose parameters would have to be estimated. If the reference sample is used for that purpose, the prior information represented by F_0 will not be independent from the sample, and therefore the procedure described previously will not be rigorously applicable. It would be preferable to use independent physical, historical or meteorological information. However, estimating the parameters of F_0 using the reference sample can be an acceptable procedure to obtain a hybrid simulation model, combining both nonparametric resampling and parametric simulation. By changing the value of p_N (between 0 and 1), it is then possible to go from a completely nonparametric simulation model ($p_N = 0$) to a standard parametric simulation model ($p_N = 1$).

7. Comparing Statistical Models for Flood Frequency Analysis

To illustrate the Bayesian approach to simulation discussed in this paper, we consider the classical problem of estimating x_T, the design flood of return period T from a sample of flood data. This quantile can be estimated by fitting a probability distribution D using a method E for estimating the parameters. The choice of the combination D/E for return periods T larger than the sample size n has a significant influence on the result. Suppose that one has to choose between two D/E combinations, the generalized extreme value distribution (GEV) [*Jenkinson*, 1955] and the Gumbel distribution (EV1) [*Gumbel*, 1960], for fitting flood data of low coefficient of variation (CV) from the province of Ontario (Canada), and that parameters are to be estimated using the method of probability-weighted moment (PWM) [*Landwehr et al.*, 1978]. The p.d.f of the GEV distribution is given by

$$F(x) = \Pr[X \le x] = \exp\{-[1 - k(x-u)/\alpha]^{1/k}\} \qquad k \neq 0 \tag{14}$$

$$F(x) = \Pr[X \le x] = \exp\{-\exp[-(x-u)/\alpha]\} \qquad k = 0$$

where k, u, and $\alpha(>0)$ are respectively, the shape, location, and scale parameters of the distribution. The shape parameter k determines the coefficient of skewness (CS) of the distribution. The EV1 is a particular case of the GEV corresponding to $k = 0$, or CS = 1.14. The problem of choosing between the GEV/PWM and EV1/PWM models has been discussed in a parametric context by *Lu and Stedinger* [1992]. Nonparametric Bayesian analysis provides a complementary analysis.

7.1. Analysis of the Reference Sample

Fortin [1994] has shown that the sample of maximum annual flood data observed on the Black River near Washago is representative of series with low CV ($\le$0.3) in Ontario, meaning that it is possible to reproduce the variability of CS values observed in these flood series by resampling the Black River data. Seventy-five years of data are available. CV and CS, estimated from the sample, are, respectively, 0.24 and 0.19. Figure 2 shows the observations plotted on *Gumbel* [1960] probability paper, using *Cunnane*'s [1978] plotting position $p_j = (j - 0.4)/(N + 0.2)$, together with GEV/PWM and EV1/PWM models fitted to the data. Clearly, the EV1/PWM model does not fit the smallest observations satisfactorily. The GEV/PWM model, on the other hand, fits the lower observations well, but seems to underestimate x_T for large return periods. Apparently, neither distribution is flexible enough to model the underlying distribution of floods, but both could provide a satisfactory approximation.

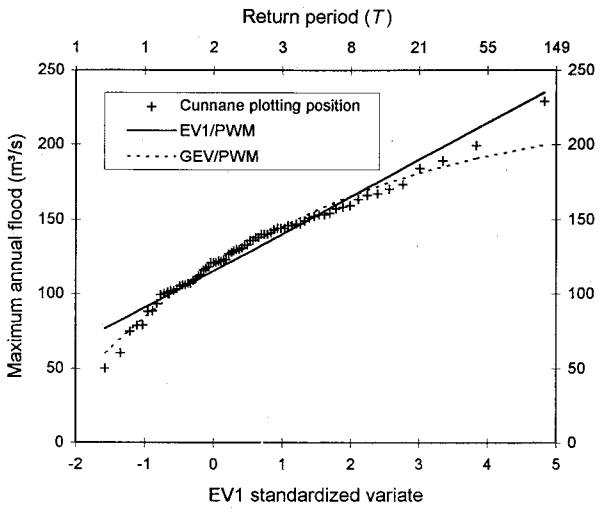

Figure 2. Probability plot of the maximum annual flood series recorded on the Black River near Washago (Ontario) until 1990.

7.2. Simulation Design

To compare the GEV/PWM and EV1/PWM models by simulation, three procedures are considered: parametric simulation, the bootstrap, and Pólya resampling. $R = 1000$ samples of size $n = 10(10)70$ were simulated using each simulation model. The parameters of the GEV parent distribution, obtained by the PWM method, are $k = 0.273$, $\alpha = 30.1$, and $u = 119$. For this population CV = 0.23 and CS = 0.01. The EV1/PWM and GEV/PWM models can be compared on the basis of the standard deviation $\sigma = E[(\hat{x}_T - E[\hat{x}_T])^2]^{1/2}$, bias $b = E[\hat{x}_T - x_T]$, and root mean square error RMSE = $E[(\hat{x}_T - x_T)^2]^{1/2} = (\sigma^2 + b^2)^{1/2}$ of design flood estimates $\hat{x}_T$ for various return periods $T = 2$, 5, 10, 20, and 50 years.

Standard deviation, bias, and RMSE are easily computed from the simulated samples using parametric simulation, but estimation of bias and RMSE using nonparametric simulation is more difficult. Indeed, it is not obvious how the reference value of x_T, to which the estimates $\hat{x}_T$ must be compared, should be computed from the reference sample **y**. *Ferguson* [1973] proposes a nonparametric Bayes estimate of a quantile under absolute error loss, but since the loss function is arbitrary there is no reason to prefer this estimate to others. It is also possible to obtain x_T by interpolating between the plotting positions, but this approach is sensitive both to the choice of the plotting position and to their variability. We obtained more stable reference values for x_T by using a nonparametric kernel estimation with a Cauchy kernel [*Adamowski*, 1985]. The smoothing factor h was estimated by minimizing the sum of squared differences $\sum_{j=1}^{N} [p_j - \hat{F}(x_{(j)})]^2$ between the nonparametric distribution function $\hat{F}$ and the plotting positions p_j [*Adamowski*, 1985]. This method of fitting a nonparametric distribution was chosen because it depends on the choice of a plotting position and therefore can be used to study the sensitivity of the computed reference values. The differences ob-

Table 1. Reference Values of x_T

T	Reference Value of x_T	
	GEV Parent	Kernel Method
2	129	130
5	156	152
10	170	166
20	180	184
50	191	199

GEV, generalized extreme value.

served by considering all plotting positions $p_j = (j - a)/(N + 1 - 2a)$ for $0 \leq a \leq 0.5$ were smaller than 1%. The values obtained with *Cunnane*'s [1978] plotting position are presented in Table 1.

7.3. Parametric and Nonparametric Simulation Results

Figure 3 shows that for $T = 50$ years, Pólya resampling is preferable to the bootstrap. Compared with the exact results obtained using Pólya resampling, the bootstrap gives a fair approximation of the mean value of $\hat{x}_T$ but significantly underestimates its standard deviation, especially for large sample sizes. As the use of the bootstrap will result in systematic underestimation of the standard deviation and RMSE, that simulation model will not be considered further in this example.

Figures 4a–4c show, respectively, the standard deviation, bias, and RMSE of the GEV/PWM and EV1/PWM models obtained using both Pólya resampling and a GEV parent distribution for $T = 50$ years. The use of a GEV parent leads to smaller standard deviations for both the GEV/PWM and EV1/PWM models (Figure 4a). Notice that under the hypothesis of a GEV parent, both models have similar standard deviations. This is somewhat surprising, since the EV1/PWM model has only two parameters compared with the three parameters of the GEV/PWM model. However, *Lu and Stedinger* [1992] have shown that for a GEV parent distrihution, the standard deviation of the GEV/PWM model approaches the standard deviation of the EV1/PWM model for large values of k.

The results for the bias (Figure 4b) are also interesting: under a GEV hypothesis, GEV/PWM estimates are almost bias-free, but the EVl/PWM estimates are highly biased (around 20 m^3/s) for all sample sizes (for $T = 50$ years). This was to be expected, since the parent distribution is almost symmetric, whereas the EV1 distribution always has a CS of 1.14, leading to a systematic overestimation of design floods. The bias obtained by Pólya resampling is smaller for both models and reflects the difference of 8 m^3/s between the parametric and the nonparametric reference value of x_{50}. Notice that for all sample sizes and for both simulation approaches, the bias of both models is found to be independent of the sample size.

The standard deviation and bias can be combined to obtain the RMSE (Figure 4c). With the hypothesis of a GEV parent the RMSE for EV1/PWM is higher than for GEV/PWM, especially for large sample sizes. Since both modeis have similar variance (Figure 4a), the lower RMSE of the GEV/PWM model is entirely due to its very low bias. The results obtained using Pólya resampling are reversed: the EV1/PWM does better than the GEV/PWM model, especially for small sample sizes. However, the difference between the two models is less important than with parametric simulation. When the RMSE is plotted as a function of the return period T for a fixed sample size $n = 20$ years (Figure 5), it is seen that similar conclusions would be drawn for $T = 20$ years. For return periods smaller than the sample size, it is difficult to discriminate between the estimation models, but since the difference between the models is small, it does not matter much which model is chosen. Given the contradictory results obtained using parametric and nonparametric simulation models, it is difficult to choose between the EV1/PWM and GEV/PWM models. The conclusions of the simulation study depend on the confidence one has in the hypothesis of the GEV distribution.

7.4. Combining Parametric and Nonparametric Simulation

The degree of confidence in the GEV parent distribution needed to accept the results of the parametric simulation may be estimated by incorporating this hypothesis in the nonparametric analysis. As previously explained (equation (13)), this is done by considering a mixture $\hat{F}$ of a prior hypothesis F_0 (in this case the GEV distribution) and the empirical distribution $F_N(x|\mathbf{y})$. To use this approach, it is necessary to specify p_N, which represents the degree of confidence (between 0 and 1) in the prior hypothesis F_0. A hybrid simulation model is then

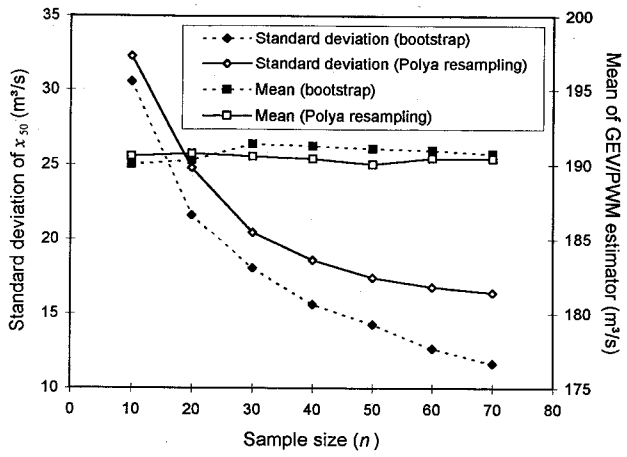

Figure 3. Mean and standard deviation of x_{50} for GEV/PWM model estimated by bootstrap and computed by Pólya resampling.

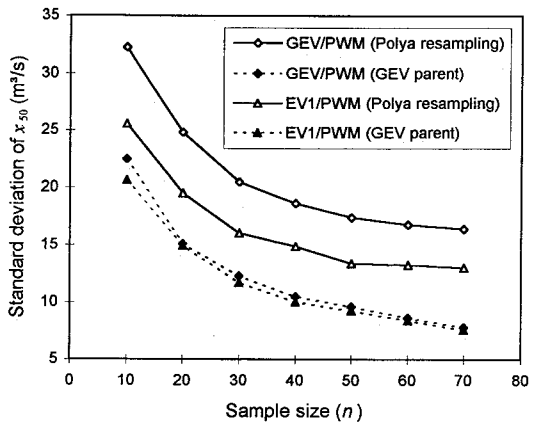

Figure 4a. Standard deviation of x_{50} for the GEV/PWM and EV1/PWM models.

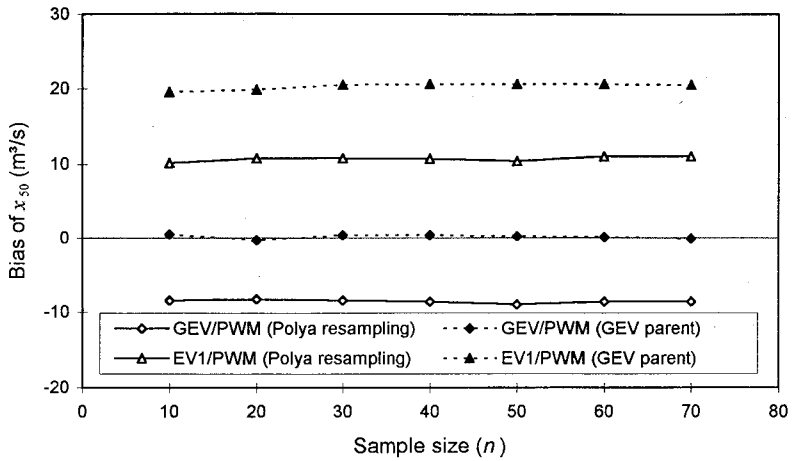

Figure 4b. Bias of x_{50} for the GEV/PWM and EV1/PWM models.

obtained which has parametric simulation (if $p_N = 1$) and Pólya resampling (if $p_N = 0$) as particular cases.

For a sample size of $n = 20$ and values of $p_N = 0(0.1)1$, the RMSE of the GEV/PWM and EV1/PWM models were compared on the basis of $R = 1000$ simulated samples. Figure 6 shows the results for return periods of $T = 20$ and $T = 50$ years. It is interesting to determine the threshold value of p_N for which the RMSE of the GEV/PWM model becomes lower than the RMSE of the EV1/PWM model. For values of p_N larger than this threshold, one should prefer the GEV/PWM model to the simpler EV1/PWM model; for values lower than the threshold, the prior hypothesis of the GEV parent is not strong enough to reach this conclusion. For both $T = 20$ and $T = 50$ years the threshold is about $p_N = 0.4$. The confidence in the GEV hypothesis necessary to reach a conclusion is relatively high. It is best understood by a change of scale: to each value of p_N corresponds a weight $\alpha = N \cdot p_N/(1 - p_N)$, measured in units homogeneous to a sample size. To conclude that the RMSE of the GEV/PWM model is lower, information independent from the reference sample equivalent to 50 years of data must support the hypothesis of a GEV parent distribution. This value is relatively large, when compared to the size of the reference sample $N = 75$ years, but depending on one's confidence in the GEV parent distribution, one might accept the conclusion that the GEV/PWM model is superior. If insufficient prior information is available to favor a GEV parent, then the conclusions of the nonparametric analysis are more attractive. In that case the low bias of the GEV/PWM model, although useful in the artificial GEV world, has less bearing on the decision, and the EV1/PWM model can be preferred.

8. Summary and Conclusion

A Bayesian procedure called Pólya resampling for simulating i.i.d. observations from a nonparametric predictive distribution has been presented. For large samples the procedure is approximately equivalent to the bootstrap, but Pólya resampling is preferable in all cases. Indeed, it takes into account the fact that the observed data is only a sample of an unknown population, whereas using the bootstrap one assumes that the observed sample is equivalent to the unknown population. The procedure was extended to take into account prior information, mainly in the form of a parametric distribution whose weight, relative to the reference sample size, may be adjusted depending on the confidence in this prior information. Conse-

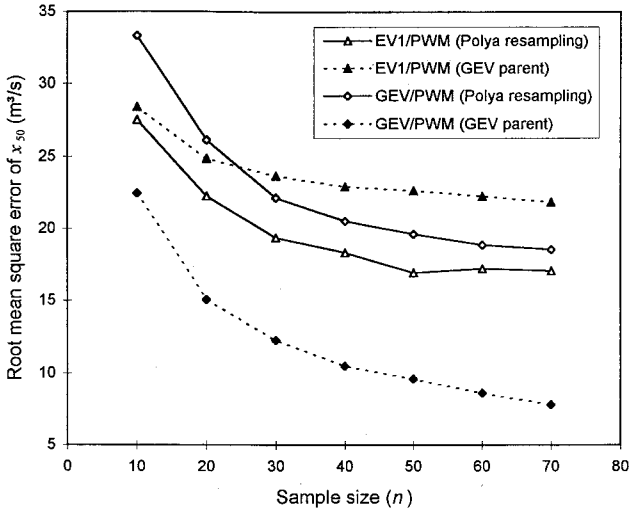

Figure 4c. RMSE of x_{50} for the GEV/PWM and EV1/PWM models.

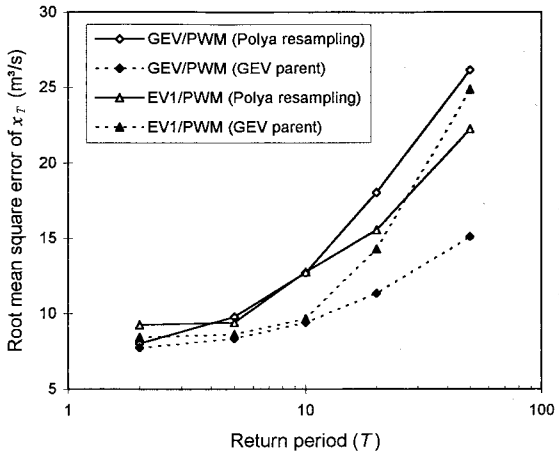

Figure 5. RMSE of x_T for the GEV/PWM and EV1/PWM models ($n = 20$).

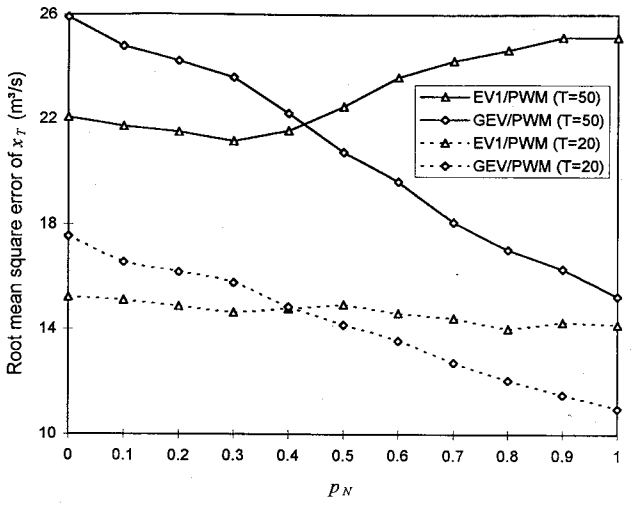

Figure 6. RMSE of x_{20} and x_{50} for the GEV/PWM and EV1/PWM models, computed by generalized Pólya resampling.

quently, the proposed simulation procedure generalizes both nonparametric and parametric simulation.

Although we propose a nonparametric simulation procedure, we do not advocate nonparametric extrapolation of quantiles. On the contrary, a sound parametric hypothesis is necessary for extrapolation. The nonparametric simulation procedure presented in this paper can, however, be used to compare the performance of various parametric models for extrapolating the distribution of a hydrological variable whose parametric distribution type is unknown.

Like parametric simulation, Pólya resampling is limited to stationary series and independent observations. A moving block bootstrap was presented by *Kunsch* [1989] and a nearest neighbor bootstrap was proposed by *Lall and Sharma* [1996] to preserve the dependence structure of time series in nonparametric simulation. A similar Bayesian methodology for resampling time series could be devised. When the hypothesis of stationarity is not respected, simulation based on past data is more problematic. In some cases a nonparametric analysis based on the theory of chaotic dynamical systems may be appropriate [*Lall et al.*, 1996].

In Bayesian analysis, however, the hypothesis of i.i.d. observations may be replaced by the weaker requirement that the observations be exchangeable [*Berger*, 1985]. Pólya resampling is appropriate if the same subjective probability applies to any permutation of the reference sample. A weakness of the Bayesian approach is the need to specify a parametric prior distribution for the parameters of the model. For Pólya resampling a complete Bayesian analysis should include a robustness study of the conclusions to the hypothesis of an improper Dirichlet prior.

Nonparametric simulation cannot entirely replace parametric simulation: a nonparametric analysis requires a large sample size N, and when N is large, goodness-of-fit tests may be sufficiently powerful to assist in identifying a parent distribution. However, nonparametric simulation may put the results of a parametric analysis into perspective. Indeed, as shown in the example, the two methods can give very different results. Nonparametric simulation provides a simple way of assessing the sensitivity of the results of a parametric simulation to the hypothesis of the parent distribution, which can either strengthen or relax the conclusions of a parametric analysis. Furthermore, the confidence in the hypothesis of a parent distribution needed for the conclusions of a parametric simulation to hold may be evaluated quantitatively by taking into account in the nonparametric analysis a hypothesis F_0 on the parametric distribution of a variable. In our example, the parameters of F_0 were estimated from the sample, which is not strictly correct, since F_0 must be based on information independent from the sample. Another way of estimating the parameters is to use independent regional information. This would be particularly useful at sites where the sample size is small.

Apart from the example presented here to illustrate the methodology, Pólya resampling has been used to compare at-site flood frequency analysis procedures for the provinces of Québec and Ontario [*Fortin*, 1994; *Fortin et al.*, 1997]. While the results obtained by this method were encouraging, additional applications are needed to determine the practical value of the method.

Appendix A: Nonparametric Preposterior Analysis

This appendix shows that a nonparametric predictive distribution based on the Dirichlet process corresponds to a Pólya urn model [*Lo*, 1988]. Let $\mathcal{X}$ be the sample space of a random variable X and let $\mathcal{A}$ be a σ field of $\mathcal{X}$, that is, a set of events closed under union, intersection, and complement. Let $\alpha(B)$ be a finite non-null measure on $(\mathcal{X}, \mathcal{A})$. A random process P is said to be a Dirichlet process $\mathcal{D}(\alpha)$ on $(\mathcal{X}, \mathcal{A})$ with parameter $\alpha(B)$ if, for any measurable partition $\mathbf{B} = \{B_1, B_2, \cdots, B_m\}$ of $\mathcal{X}$, the probabilities $P(B_1)$, $P(B_2)$, $\cdots$, $P(B_m)$ follow a Dirichlet distribution $\mathcal{D}[\alpha(B_1), \alpha(B_2), \cdots, \alpha(B_m)]$.

If additional information becomes available, the parameter $\alpha(B)$ can be updated through the use of *Ferguson*'s [1973] theorem. Given a sample $\mathbf{x} = \{x_j, j = 1, 2, \cdots, n\}$ of size n from P, the updated distribution of the random variables $P(B_1)$, $P(B_2)$, $\cdots$, $P(B_m)$ is $\mathcal{D}(\alpha^{\mathbf{x}}(B_1), \alpha^{\mathbf{x}}(B_2), \cdots, \alpha^{\mathbf{x}}(B_m))$ with $\alpha^{\mathbf{x}}(B) = \alpha(B) + \sum_{j=1}^{n} \delta_{x_j}(B)$, where $\delta_x(B)$ denotes the measure giving mass one to the point x if $x \in B$ and zero otherwise ($x \notin B$). Let $n_k = \sum_{j=1}^{n} \delta_{x_j}(B_k)$ denote the number of occurrences of event B_k in $\mathbf{x}$. The predictive distribution $f(\mathbf{x})$ is given by (15), which is similar to (8):

$$f(\mathbf{x}) = \binom{n}{n_1, n_2, \cdots, n_m} \frac{\Gamma(\alpha(\mathcal{X}))}{\Gamma(\alpha^{\mathbf{x}}(\mathcal{X}))} \cdot \left[\prod_{j=1}^{m} \Gamma(\alpha^{\mathbf{x}}(B_j)) \Big/ \prod_{j=1}^{m} \Gamma(\alpha(B_j)) \right] \tag{15}$$

The updated predictive distribution $f(\mathbf{x}|\mathbf{y})$ for a reference sample $\mathbf{y} = \{y_j, j = 1, 2, \cdots, N\}$ of size N containing, respectively, $r_1, r_2, \cdots, r_m$ occurrences of $B_1, B_2, \cdots, B_m$ is given by

$$f(\mathbf{x}|\mathbf{y}) = \binom{n}{n_1, n_2, \cdots, n_m} \frac{\Gamma(\alpha^{\mathbf{y}}(\mathcal{X}))}{\Gamma(\alpha^{\mathbf{x},\mathbf{y}}(\mathcal{X}))} \cdot \left[\prod_{j=1}^{m} \Gamma(\alpha^{\mathbf{x},\mathbf{y}}(B_j)) \Big/ \prod_{j=1}^{m} \Gamma(\alpha^{\mathbf{y}}(B_j)) \right] \tag{16}$$

where $\alpha^{\mathbf{y}}(B)$ and $\alpha^{\mathbf{x},\mathbf{y}}(B)$ are obtained using Ferguson's theorem:

$$\alpha^{\mathbf{y}}(B) = \alpha(B) + \sum_{j=1}^{N} \delta_{y_j}(B) \tag{17}$$

$$\alpha^{\mathbf{x},\mathbf{y}}(B) = \alpha^{\mathbf{y}}(B) + \sum_{j=1}^{n} \delta_{x_j}(B) = \alpha(B) + \sum_{j=1}^{N} \delta_{y_j}(B) + \sum_{j=1}^{n} \delta_{x_j}(B) \tag{18}$$

When no other information than a reference sample $\mathbf{y}$ is available, $\alpha(B) = 0$; therefore $\alpha^{\mathbf{y}}(\mathcal{X}) = N$, $\alpha^{\mathbf{y}}(B_j) = r_j$, $\alpha^{\mathbf{x},\mathbf{y}}(\mathcal{X}) = N + n$, and $\alpha^{\mathbf{x},\mathbf{y}}(B_j) = r_j + n_j$. Consequently, (16) simplifies to

$$f(\mathbf{x}|\mathbf{y}) = \binom{n}{n_1, n_2, \cdots, n_m} \frac{\Gamma(N)}{\Gamma(N+n)} \cdot \left[\prod_{j=1}^{m} \Gamma(r_j + n_j) \Big/ \prod_{j=1}^{m} \Gamma(r_j) \right] \tag{19}$$

Equation (19), valid for all measurable partitions $\mathbf{B}$ of $\mathcal{X}$, simplifies further for specific partitions. Consider a partition $\mathbf{B} = \{B_1, B_2, \cdots, B_{N+1}\}$ of $\mathcal{X}$ into $N + 1$ subsets such that $y_j \in B_j$, $j = 1, 2, \cdots, N$; that is, only one observation of the reference sample falls into each subset B_j, $j = 1, 2, \cdots, N$ and no observation falls into B_{N+1}. Of course, all observations in $\mathbf{y}$ must be different for this to be possible, but it is not essential for the final result to hold. If r_k represents the number of occurrences of the event B_k in $\mathbf{y}$, then $r_1 = 1$, $r_2 = 1, \cdots, r_N = 1$, $r_{N+1} = 0$. Therefore (19) simplifies to

$$f(\mathbf{x}|\mathbf{y}) = \binom{n}{n_1, n_2, \cdots, n_N} \frac{n_1! n_2! \cdots n_N!}{N(N+1)\cdots(N+n-1)} \tag{20}$$

which is a Pólya urn model, since (20) is a particular case of (9) obtained for $r_1 = 1, r_2 = 1, \cdots, r_N = 1$.

Appendix B: Posterior Estimation of a Distribution Function

A nonparametric estimation of a distribution function can be based on the Dirichlet process [*Ferguson*, 1973]. Let X be a random variable with distribution $F(x) = \Pr[X \leq x]$ defined on the real line $\mathbb{R}$. Let the space of actions A include all distributions $\hat{F}$ defined on $\mathbb{R}$, and suppose a quadratic loss function $l(\hat{F}, F) = \int [\hat{F}(x) - F(x)]^2 \, dx$. Suppose that apart from a random sample $\mathbf{y} = \{y_j, j = 1, 2, \cdots, N\}$, all other prior information on F can be represented by a Dirichlet process $\mathcal{D}(\alpha)$ on $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the σ field of Borel sets. *Ferguson* [1973] shows that the Bayesian estimate $\hat{F}(x)$ minimizes the posterior expectation of $F(x|\mathbf{y})$ which is given by

$$\hat{F}(x) = E[F(x|\mathbf{y})] = \frac{\alpha((-\infty, x]) + \sum_{j=1}^{N} \delta_{y_j}((-\infty, x])}{\alpha(\mathbb{R}) + N} \tag{21}$$

Equation (21) can alternatively be written as a mixture of the prior guess $F_0(x)$ and of the empirical distribution function $F_N(x|\mathbf{y})$:

$$\begin{aligned} \hat{F}(x) &= E[F(x|\mathbf{y})] = p_N F_0(x) + (1 - p_N) F_N(x|\mathbf{y}) \\ p_N &= \alpha(\mathbb{R})/(\alpha(\mathbb{R}) + N) \\ F_0(x) &= \alpha((-\infty, x])/\alpha(\mathbb{R}) \\ F_N(x|\mathbf{y}) &= 1/N \sum_{j=1}^{N} \delta_{y_j}((-\infty, x]) \end{aligned} \tag{22}$$

Appendix C: Using Prior Information With Pólya Resampling

Pólya resampling may be generalized to include a prior hypothesis of the distribution of a variable X, restricted to $\mathbb{R}$ to simplify the notation. Consider the updated predictive distribution $f(x_1|\mathbf{y})$ of a single observation x_1, obtained from (16), (17), and (18), by letting $n = 1$:

$$f(x_1|\mathbf{y}) = \frac{\Gamma(\alpha(\mathbb{R}) + N)}{\Gamma(\alpha(\mathbb{R}) + N + 1)} \cdot \frac{\prod_{j=1}^{m} \Gamma\left(\alpha(B_j) + \sum_{i=1}^{N} \delta_{y_i}(B_j) + \delta_{x_1}(B_j) \right)}{\prod_{j=1}^{m} \Gamma\left(\alpha(B_j) + \sum_{i=1}^{N} \delta_{y_i}(B_j) \right)} \tag{23}$$

Equation (23) is valid for all measurable partitions $\mathbf{B} = \{B_1, B_2, \cdots, B_m\}$ of $\mathcal{X}$ and in particular for $\mathbf{B} = \{B_1, B_2\}$ with $B_1 = (-\infty, x_1]$ and $B_2 = (x_1, \infty)$. For this simple partition, (23) simplifies to

$$f(x_1|\mathbf{y}) = \frac{\alpha((-\infty, x_1]) + \sum_{j=1}^{N} \delta_{y_j}((-\infty, x_1])}{\alpha(\mathbb{R}) + N} \tag{24}$$

which, according to (22), can be rewritten as

$$f(x_1|\mathbf{y}) = \hat{F}(x_1) = p_N F_0(x_1) + (1 - p_N) F_N(x_1|\mathbf{y}) \tag{25}$$

where $F_0(x)$ is an hypothesis on the distribution of X, p_N is the strength of this hypothesis, and $F_N(x|\mathbf{y})$ is the empirical distribution function based on the observed reference sample $\mathbf{y}$. The predictive distribution of a second observation x_2, given x_1 and $\mathbf{y}$, can be obtained in a similar manner by applying Ferguson's theorem:

$$f(x_2|x_1, \mathbf{y}) = \frac{\Gamma(\alpha(\mathbb{R}) + N + 1)}{\Gamma(\alpha(\mathbb{R}) + N + 2)} \cdot \frac{\prod_{j=1}^{m} \Gamma\left[\alpha(B_j) + \sum_{i=1}^{N} \delta_{y_i}(B_j) + \delta_{x_1}(B_j) + \delta_{x_2}(B_j) \right]}{\prod_{j=1}^{m} \Gamma\left[\alpha(B_j) + \sum_{i=1}^{N} \delta_{y_i}(B_j) + \delta_{x_1}(B_j) \right]} \tag{26}$$

Recalling that (26) is valid for all measurable partitions, and considering the simple partition $\mathbf{B}' = \{B'_1, B'_2\}$, with $B'_1 = (-\infty, x_2]$ and $B'_2 = (x_2, \infty)$, a simpler distribution is obtained:

$$f(x_2|x_1, \mathbf{y}) = \frac{\alpha((-\infty, x_2]) + \sum_{j=1}^{N} \delta_{y_j}((-\infty, x_2]) + \delta_{x_1}((-\infty, x_2])}{\alpha(\mathbb{R}) + N + 1} \quad (27)$$

which can be rewritten as

$$f(x_2|x_1, \mathbf{y}) = \hat{F}^{(2)}(x_2) = p_N^{(2)} F_0(x_2) + (1 - p_N^{(2)}) F_N^{(2)}(x_2) \quad (28)$$

Here $p_N^{(2)} = \alpha(\mathbb{R})/(\alpha(\mathbb{R}) + N + 1)$ and $F_N^{(2)}(x) = [\sum_{j=1}^{N} \delta_{y_j}((-\infty, x]) + \delta_{x_1}((-\infty, x])]/(N + 1)$, which corresponds to the empirical distribution function of the reference sample to which observation x_1 has been added. It can be proven by induction that observation x_i is obtained from a mixture of $F_0(x)$ and $F_N^{(i)}(x)$:

$$f(x_i|x_1, x_2, \cdots, x_{i-1}, \mathbf{y}) = \hat{F}^{(i)}(x_i) = p_N^{(i)} F_0(x_i) + (1 - p_N^{(i)}) F_N^{(i)}(x_i)$$

$$p_N^{(i)} = \alpha(\mathbb{R})/(\alpha(\mathbb{R}) + N + i - 1) \quad (29)$$

$$F_N^{(i)}(x_i) = \left[\sum_{j=1}^{N} \delta_{y_j}((-\infty, x_i]) + \sum_{j=1}^{i-1} \delta_{x_j}((-\infty, x_i]) \right] \Big/ (N + i - 1)$$

Acknowledgments. The financial support of the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged. Discussions with Peter F. Rasmussen and Jery R. Stedinger, as well as the comments of an anonymous reviewer, resulted in a significant improvement of the manuscript.

References

Adamowski, K., Nonparametric kernel estimation of flood frequencies, *Water Resour. Res.*, *21*, 1585–1590, 1985.

Ahmad, M. I., C. D. Sinclair, and A. Werritty, Log-logistic flood frequency analysis, *J. Hydrol.*, *98*, 205–224, 1988.

Ashkar, F., B. Bobée, and J. Bernier, Separation of skewness: Reality or regional artifact?, *J. Hydraul. Eng.*, *118*, 460–475, 1992.

Berger, J. O., *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer-Verlag, New York, 1985.

Bernier, J., Robustness of models and estimation methods in flood analysis, in *Stochastic and Statistical Methods in Hydrology and Environmental Engineering*, vol. 1, edited by K. Hipel, pp. 187–197, Kluwer Acad., Norwell, Mass., 1993.

Bernier, J., Information, modèles, risques et hydrologie statistique, in *Méthodes Statistiques et Bayésiennes en Hydrologie*, Unesco Press, Paris, in press, 1997.

Bobée, B., G. Cavadias, F. Ashkar, J. Bernier, and P. Rasmussen, Towards a systematic approach to comparing distributions used in flood frequency analysis, *J. Hydrol.*, *142*, 121–136, 1993.

Box, G. E. P., and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Mass., 1973.

Cunnane, C., Unbiased plotting position—a review, *J. Hydrol.*, *37*, 205–222, 1973.

Efron, B., Computers and the theory of statistics: Thinking the unthinkable, *Soc. Ind. Appl. Math.*, *21*, 460–480, 1979.

Feller, W., *An Introduction to Probability Theory and Its Applications*, vol. 2, 2nd ed., John Wiley, New York, 1971.

Ferguson, T. S., A Bayesian analysis of some nonparametric problems, *Ann. Stat.*, *1*, 209–230, 1973.

Fortin, V., Une méthode rationnelle de comparaison des distributions de crue, M.Sc. dissertation, 105 pp., Inst. Natl. de la Rech. Sci., Univ. du Quebec, Sainte-Foy, Canada, 1994.

Fortin, V., B. Bobée, and J. Bernier, A rational approach to the comparison of flood distributions by simulation, *J. Hydrol. Eng.*, in press, 1997.

Gumbel, E. J., *Statistics of Extremes*, Columbia Univ. Press, New York, 1960.

Haktanir, T., Comparison of various flood frequency distributions using annual flood peaks data of rivers in Anatolia, *J. Hydrol.*, *136*, 1–31, 1992.

Haktanir, T., and H. B. Horlacher, Evaluation of various distributions for flood frequency analysis, *Hydrol. Sci. J.*, *38*, 15–32, 1993.

Jenkinson, A. F., The frequency distribution of the annual maximum (or minimum) of meteorological elements, *Q. J. R. Meteorol. Soc.*, *81*, 158–171, 1955.

Klemes, V., Dilettantism in hydrology: Transition or destiny?, *Water Resour. Res.*, *22*, 177S–188S, 1986.

Kuczera, G., Robust flood frequency models, *Water Resour. Res.*, *18*, 315–324, 1982.

Kunsch, H. R., The Jackknife and the Bootstrap for general stationary observations, *Ann. Stat.*, *17*, 1217–1241, 1989.

Lall, U., and A. Sharma, A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res.*, *32*, 679–693, 1996.

Lall, U., Y.-I. Moon, and K. Bosworth, Kernel flood frequency estimators: Bandwidth selection and kernel choice, *Water Resour. Res.*, *29*, 1003–1015, 1993.

Lall, U., T. Sangoyomi, and H. D. I. Abarbanel, Nonlinear dynamics of the Great Salt Lake: Nonparametric short term forecasting, *Water Resour. Res.*, *32*, 975–985, 1996.

Landwehr, J. M., N. C. Matalas, and J. R. Wallis, Some comparisons of flood statistics in real and log space, *Water Resour. Res.*, *14*, 902–920, 1978.

Lo, A. Y., A Bayesian bootstrap for a finite population, *Ann. Stat.*, *4*, 1684–1695, 1988.

Lu, L.-H., and J. R. Stedinger, Variance of two- and three-parameter GEV/PWM quantile estimators: Formulae, confidence intervals, and a comparison, *J. Hydrol.*, *138*, 247–267, 1992.

Matalas, N. C., J. R. Slack, and J. R. Wallis, Regional skew in search for a parent, *Water Resour. Res.*, *11*, 815–826, 1975.

Moon, Y.-I., U. Lall, and K. Bosworth, A comparison of tail probability estimators for flood frequency analysis, *J. Hydrol.*, *151*, 343–363, 1993.

Potter, K. W., Research in flood frequency analysis: 1983–86, *Rev. Geophys.*, *25*, 113–118, 1987.

Potter, K. W., and D. P. Lettenmaier, A comparison of regional flood frequency estimation methods using a resampling method, *Water Resour. Res.*, *26*, 415–424, 1990.

Rasmussen, P. F., B. Bobée, and J. Bernier, Une méthodologie générale de comparaison de modèles d'estimation régionale de crue, *Rev. Sci. Eau*, *7*, 23–41, 1994.

Rubin, D. B., The Bayesian bootstrap, *Ann. Stat.*, *9*, 130–134, 1981.

Slack, J. R., J. R. Wallis, and N. C. Matalas, On the value of information to flood frequency analysis, *Water Resour. Res.*, *11*, 629–647, 1975.

Tasker, G. D., Comparison of methods for estimating low flow characteristics of streams, *Water Resour. Bull.*, *23*, 1077–1083, 1987.

Wallis, J. R., and E. F. Wood, Relative accuracy of log-Pearson type III procedures, *J. Hydraul. Eng.*, *111*, 1043–1056, 1985.

World Meteorological Organization (WMO), Statistical Distributions for Flood Frequency Analysis, *Oper. Hydrol. Rep. 33*, Geneva, 1989.

J. Bernier, B. Bobée, and V. Fortin, Institut National de la Recherche Scientifique, Université du Québec, P. O. Box 7500, Sainte-Foy, Quebec, Canada G1V 4C7. (e-mail: bobee@uquebec.ca)

(Received February 27, 1996; revised September 9, 1996; accepted October 29, 1996.)