WATER RESOURCES RESEARCH, VOL. 42, W01419, doi:10.1029/2004WR003772, 2006

# Identification of periodic autoregressive moving average models and their application to the modeling of river flows

Yonas Gebeyehu Tesfaye

Graduate Program of Hydrologic Sciences, University of Nevada, Reno, Nevada, USA

Mark M. Meerschaert

Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand

Paul L. Anderson

Department of Mathematics and Computer Science, Albion College, Albion, Michigan, USA

Received 28 October 2004; revised 19 September 2005; accepted 18 October 2005; published 31 January 2006.

[1] The generation of synthetic river flow samples that can reproduce the essential statistical features of historical river flows is useful for the planning, design, and operation of water resource systems. Most river flow series are periodically stationary; that is, their mean and covariance functions are periodic with respect to time. This article develops model identification and simulation techniques based on a periodic autoregressive moving average (PARMA) model to capture the seasonal variations in river flow statistics. The innovations algorithm is used to obtain parameter estimates. An application to monthly flow data for the Fraser River in British Columbia is included. A careful statistical analysis of the PARMA model residuals, including a truncated Pareto model for the extreme tails, produces a realistic simulation of these river flows.

**Citation:** Tesfaye, Y. G., M. M. Meerschaert, and P. L. Anderson (2006), Identification of periodic autoregressive moving average models and their application to the modeling of river flows, *Water Resour. Res.*, *42*, W01419, doi:10.1029/2004WR003772.

## 1. Introduction

[2] Time series analysis and modeling is an important tool in hydrology and water resources. It is used for building mathematical models to generate synthetic hydrologic records, to determine the likelihood of extreme events, to forecast hydrologic events, to detect trends and shifts in hydrologic records, and to interpolate missing data and extend records. The research reported in this article relates more directly to river flow data generation. Generation of synthetic river flow series may be useful for determining the dimensions of hydraulic works, for risk assessment in urban water supply and irrigation, for optimal operation of reservoir systems, for determining the risk of failure of dependable capacities of hydroelectric systems, for planning capacity expansion of water supply systems, and others [see *Salas*, 1993].

[3] The statistical characteristics of hydrologic series are important deciding factors in the selection of the type of model. For example, in most cases known in nature, river flows have significant periodic behavior in the mean, standard deviation and skewness. In addition to these periodicities, they show a time correlation structure which may be either constant or periodic. Such serial dependence or autocorrelation in river flow series usually arises from the effect of storage, such as surface, soil, and ground storages, which cause the water to remain in the system through subsequent time periods. The common procedure in modeling such periodic river flow series is first to standardize or filter the series and then fit an appropriate stationary stochastic model to the reduced series [*Salas et al.*, 1980; *Thompstone et al.*, 1985; *Vecchia*, 1985a, 1985b; *Salas*, 1993; *Chen and Rao*, 2002]. However, standardizing or filtering most river flow series may not yield stationary residuals due to periodic autocorrelations. In these cases, the resulting model is misspecified [*Tiao and Grupe*, 1980]. Periodic models can therefore be employed to remove the periodic correlation structure. An important class of periodic models useful in such situations consists of periodic autoregressive moving average (PARMA) models, which are extensions of commonly used ARMA models that allow periodic parameters. PARMA models explicitly represent the seasonal fluctuations in mean flow, flow standard deviation, and flow autocorrelation, resulting in a more realistic time series model that leads to more reliable simulations of natural river flows.

[4] There have been many discussions about periodic time series models [*Jones and Brelsford*, 1967; *Pagano*, 1978; *Troutman*, 1979; *Tjøstheim and Paulsen*, 1982; *Salas et al.*, 1981, 1982, 1985; *Vecchia*, 1985a, 1985b; *Vecchia and Ballerini*, 1991; *Salas and Obeysekera*, 1992; *Anderson and Vecchia*, 1993; *Ula*, 1990, 1993; *Ula and Smadi*, 1997, 2003; *Adams and Goodwin*, 1995; *Anderson and Meerschaert*, 1997, 1998; *Lund and Basawa*, 1999, 2000; *Shao and Lund*, 2004]. Time series analysis of data sequences usually involves three main steps: model identification, parameter estimation and diagnostic checking. Parameter

0043-1397/06/2004WR003772$09.00

**Table 1.** Moving Average Parameter Estimates and $p$ Values After $k$ = 15 Iterations of the Innovations Algorithm Applied to $N_y$ = 500 Years of Simulated PARMA$_4$(1,1) Data[a]

| Lag $\ell$ | $\hat{\psi}_0(\ell)$ | $p$ for $\hat{\psi}_0(\ell)$ | $\hat{\psi}_1(\ell)$ | $p$ for $\hat{\psi}_1(\ell)$ | $\hat{\psi}_2(\ell)$ | $p$ for $\hat{\psi}_2(\ell)$ | $\hat{\psi}_3(\ell)$ | $p$ for $\hat{\psi}_3(\ell)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | −0.606 | 0.00 | 1.231 | 0.00 | 1.710 | 0.00 | 0.617 | 0.00 |
| 2 | −0.620 | 0.00 | −0.360 | 0.00 | 1.047 | 0.00 | 0.387 | 0.00 |
| 3 | −0.329 | 0.00 | −0.250 | 0.20 | −0.346 | 0.00 | 0.329 | 0.00 |
| 4 | −0.322 | 0.00 | −0.037 | 0.48 | −0.120 | 0.69 | −0.009 | 0.87 |
| 5 | 0.041 | 0.40 | −0.055 | 0.62 | 0.058 | 0.47 | 0.087 | 0.52 |
| 6 | −0.004 | 0.98 | 0.136 | 0.08 | 0.128 | 0.46 | 0.001 | 0.98 |
| 7 | 0.017 | 0.61 | 0.148 | 0.45 | 0.187 | 0.11 | 0.086 | 0.27 |
| 8 | −0.024 | 0.74 | 0.042 | 0.42 | 0.199 | 0.50 | −0.003 | 0.95 |
| 9 | −0.017 | 0.71 | −0.006 | 0.96 | 0.049 | 0.55 | 0.034 | 0.80 |

[a]Note that the parameter estimates continue past lag 9.

estimation for PARMA models is more difficult than for stationary ARMA models because of the higher number of parameters to be estimated. *Anderson et al.* [1999] developed the innovations algorithm for parameter estimation of an infinite moving average representation of PARMA models. *Anderson and Meerschaert* [2005] also provided an asymptotic distribution for these estimates. These results can be used for the identification of PARMA models for periodic processes but heretofore have not been put to such a task. Model identification and simulation for river flows with periodic autocorrelations is the main thrust of this article. Hence, in order to create realistic synthetic river flows, one more important step is necessary. The model residuals approximate the fundamental noise sequence from which the PARMA process is built. Therefore it is necessary to estimate the statistical distribution of these random variables, in order to accurately simulate them. Failure to perform this step will lead to distorted simulation results, particularly in terms of extreme values that are important in the analysis of floods and droughts.

[5] This article has two objectives. The first is to demonstrate the effectiveness of the technique by using simulated data from different PARMA models, and the other is to describe an application with monthly flow data for the Fraser River at Hope, British Columbia.

## 2. Mathematical Formulation of PARMA Model

[6] A stochastic process $\tilde{X}_t$ is periodically stationary if its mean $\mu_t = E\tilde{X}_t$ and covariance function $\gamma_t(h) = \mathrm{Cov}(\tilde{X}_t, \tilde{X}_{t+h})$ for $h = 0, \pm 1, \pm 2, \ldots$ are periodic functions of time $t$ with the same period $S$ (that is, for some integer $S$, for $i = 0, 1, \ldots, S - 1$, and for all integers $k$ and $h$, $\mu_i = \mu_{i+kS}$ and $\gamma_i(h) = \gamma_{i+kS}(h)$).

[7] The periodic ARMA process $\tilde{X}_t$ with period $S$ (denoted by PARMA$_S(p, q)$) has representation

$$X_t - \sum_{j=1}^{p} \phi_t(j) X_{t-j} = \varepsilon_t - \sum_{j=1}^{q} \theta_t(j) \varepsilon_{t-j} \tag{1}$$

where $X_t = \tilde{X}_t - \mu_t$ and $\{\varepsilon_t\}$ is a sequence of random variables with mean zero and scale $\sigma_t$ such that $\{\delta_t = \sigma_t^{-1}\varepsilon_t\}$ is independent and identically distributed (iid). The notation in (1) is consistent with that of *Box and Jenkins* [1976]. The autoregressive parameters $\phi_t(j)$, the moving average parameters $\theta_t(j)$, and the residual standard deviations $\sigma_t$ are all periodic functions of $t$ with the same period $S \geq 1$. The standard deviations $\sigma_t$ of the noise process $\{\varepsilon_t\}$ are assumed to be strictly positive. We also assume that (1) the model admits a causal representation

$$X_t = \sum_{j=0}^{\infty} \psi_t(j) \varepsilon_{t-j} \tag{2}$$

where $\psi_t(0) = 1$ and $\sum_{j=0}^{\infty} |\psi_t(j)| < \infty$ for all $t$. Note that $\psi_t(j) = \psi_{t+kS}(j)$ for all $j$ and (2) the model also satisfies an invertibility condition

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_t(j) X_{t-j} \tag{3}$$

where $\pi_t(0) = 1$ and $\sum_{j=0}^{\infty} |\pi_t(j)| < \infty$ for all $t$. Again, $\pi_t(j) = \pi_{t+kS}(j)$ for all $j$.

## 3. Identification and Estimation for PARMA Models

[8] In this section we document the essential results and ideas regarding the identification and parameter estimation for PARMA models. A narrative is included prior to each result allowing the reader to gain insight into the methodology. We include references that furnish the detailed proofs. Given $N_y$ years of data, consisting of $N = N_y S$ data points, define the sample mean

$$\hat{\mu}_i = N_y^{-1} \sum_{k=0}^{N_y - 1} \tilde{X}_{kS+i} \tag{4}$$

**Table 2.** Model Parameters and Estimates for Simulated PARMA$_4$(1,1) Data

| Parameter | Season 0 | Season 1 | Season 2 | Season 3 |
|---|---|---|---|---|
| $\theta$ | 0.25 | 0.65 | 0.90 | 0.35 |
| $\hat{\theta}$ | 0.400 | 0.636 | 0.859 | 0.391 |
| $\phi$ | −0.90 | 0.50 | 0.80 | 0.25 |
| $\hat{\phi}$ | −1.005 | 0.595 | 0.850 | 0.226 |
| $\sigma$ | 0.90 | 1.90 | 0.50 | 1.20 |
| $\hat{\sigma}$ | 0.832 | 1.771 | 0.482 | 1.215 |

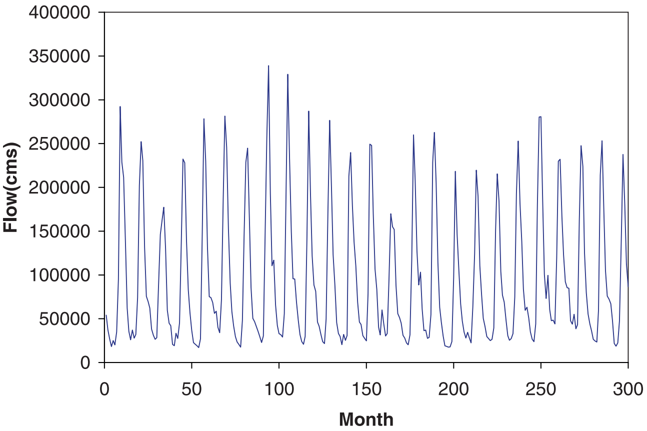


**Figure 1.** Average monthly flows ($m^3$ $s^{-1}$) for the Fraser River at Hope, British Columbia, indicate a seasonal pattern.

the sample autocovariance

$$\hat{\gamma}_i(\ell) = N_y^{-1} \sum_{j=0}^{N_y-1-m} X_{jS+i} X_{jS+i+\ell} \tag{5}$$

and the sample autocorrelation

$$\hat{\rho}_i(\ell) = \frac{\hat{\gamma}_i(\ell)}{\sqrt{\hat{\gamma}_i(0)\hat{\gamma}_{i+\ell}(0)}} \tag{6}$$

where $X_t = \tilde{X}_t - \hat{\mu}_t$, $m = [(i + \ell)/S]$ and $[\,\cdot\,]$ is the greatest integer function. The innovations algorithm developed by *Anderson et al.* [1999] gives us a practical method for estimating the parameters in our PARMA model. Since the parameters are seasonally dependent, there is a notational difference between the innovations algorithm for PARMA processes and that for ARMA processes [cf. *Brockwell and Davis*, 1991]. We introduce this difference through the "season," $i$. For monthly data we have $S = 12$ seasons and our convention is to let $i = 0$ represent the first month, $i = 1$ represent the second, . . ., and $i = S - 1 = 11$ represent the last.

[9] The idea of the innovations algorithm is to estimate the moving average in (2) by a finite approximation involving the $k$ most recent observations. Let

$$\hat{X}^{(i)}_{i+k} = \phi^{(i)}_{k,1} X_{i+k-1} + \cdots + \phi^{(i)}_{k,k} X_i, \quad k \geq 1, \tag{7}$$

represent the best linear predictor of $X_{i+k}$ based on the data $X_i$, . . ., $X_{i+k-1}$, i.e., the one which minimizes the mean squared error

$$v_{k,i} = E\left(X_{i+k} - \hat{X}^{(i)}_{i+k}\right)^2 \tag{8}$$

by choice of the coefficients $\phi^{(i)}_{k,1}$, . . ., $\phi^{(i)}_{k,k}$. We can rewrite (7) as

$$\hat{X}^{(i)}_{i+k} = \sum_{j=1}^{k} \theta^{(i)}_{k,j}\left(X_{i+k-j} - \hat{X}^{(i)}_{i+k-j}\right) \tag{9}$$

where $X_{i+k-j} - \hat{X}^{(i)}_{i+k-j}$, $j = 1$, . . ., $k$ are uncorrelated elements, which estimate the noise sequence (also called the "innovations") $\varepsilon_{i+k-j}$ in (2). Hence the parameters $\theta^{(i)}_{k,j}$ estimate the moving average coefficients $\psi_{i+k}(j)$ in that equation. *Anderson et al.* [1999] show that these coefficients (along with the corresponding mean squared errors) can be recursively computed by the following variant of the innovations algorithm:

$$\begin{aligned} v_{0,i} &= \gamma_i(0) \\ \theta^{(i)}_{k,k-\ell} &= \left(v_{\ell,i}\right)^{-1}\left[\gamma_{i+\ell}(k-\ell) - \sum\nolimits_{j=0}^{\ell-1} \theta^{(i)}_{\ell,\ell-j}\theta^{(i)}_{k,k-j} v_{j,i}\right] \\ v_{k,i} &= \gamma_{i+k}(0) - \sum\nolimits_{j=0}^{k-1} \left(\theta^{(i)}_{k,k-j}\right)^2 v_{j,i} \end{aligned} \tag{10}$$

where (10) is solved in the order $v_{0,i}$, $\theta^{(i)}_{1,1}$, $v_{1,i}$, $\theta^{(i)}_{2,2}$, $\theta^{(i)}_{2,1}$, $v_{2,i}$, $\theta^{(i)}_{3,3}$, $\theta^{(i)}_{3,2}$, $\theta^{(i)}_{3,1}$, $v_{3,i}$, . . . and so forth. As $k$ increases to infinity, these parameter estimates converge to the moving average model parameters in equation (2) as follows:

$$\begin{aligned} \theta^{(\langle i-k\rangle)}_{k,j} &\to \psi_i(j) \\ v_{k,\langle i-k\rangle} &\to \sigma_i^2 \end{aligned} \tag{11}$$

for all $i, j$ where $\langle t\rangle$ is the season corresponding to index $t$, so that $\langle jS + i\rangle = i$. If we replace the autocovariances in (10) with the corresponding sample autocovariances (5), we obtain the innovations estimates $\hat{\theta}^{(i)}_{k,l}$ and $\hat{v}_{k,i}$ based on the time series data. *Anderson et al.* [1999] show that these quantities converge (in probability) to give a consistent estimate of the moving average model parameters from data. Furthermore, *Anderson and Meerschaert* [2005] show that

$$N_y^{1/2}\left(\hat{\theta}^{(\langle i-k\rangle)}_{k,j} - \psi_i(j)\right) \Rightarrow \mathcal{N}\left(0, \sum_{n=0}^{j-1} \frac{\sigma^2_{i-n}}{\sigma^2_{i-j}} \psi_i^2(n)\right) \tag{12}$$

as $N_y \to \infty$ and $k \to \infty$ for any fixed $i = 0, 1, \ldots, S - 1$, where "⇒" indicates convergence in distribution, and $\mathcal{N}(m, v)$ is a normal random variable with mean $m$ and variance $v$. The main technical condition for the conver-

**Table 3.** Sample Mean, Standard Deviation, and Autocorrelation at Lag 1 and 2 of Average Monthly Flow Series for the Fraser River Near Hope, British Columbia

| Month | $\hat{\mu}$ | $\hat{\sigma}$ | $\hat{\rho}(1)$ | $\hat{\rho}(2)$ |
|---|---|---|---|---|
| Oct | 69763 | 19997 | 0.688 | 0.517 |
| Nov | 56000 | 17698 | 0.731 | 0.581 |
| Dec | 40352 | 12817 | 0.715 | 0.531 |
| Jan | 33135 | 9252 | 0.787 | 0.691 |
| Feb | 30861 | 8845 | 0.779 | 0.385 |
| Mar | 29709 | 8834 | 0.510 | 0.224 |
| Apr | 59293 | 20268 | 0.302 | −0.294 |
| May | 171907 | 40200 | 0.272 | −0.047 |
| Jun | 248728 | 45120 | 0.568 | 0.496 |
| Jul | 199118 | 42543 | 0.779 | 0.462 |
| Aug | 127157 | 28070 | 0.718 | 0.320 |
| Sep | 86552 | 20052 | 0.635 | 0.454 |

**Table 4.** Moving Average Parameter Estimates $\hat{\psi}_i(\ell)$ at Season $i$ and Lag $\ell = 1, 2, \ldots, 5$ and $p$ Values After $k = 20$ Iterations of the Innovations Algorithm Applied to Average Monthly Flow Series for the Fraser River Near Hope, British Columbia[a]

| $i$ | $\hat{\psi}_i(1)$ | $p$ for $\hat{\psi}_i(1)$ | $\hat{\psi}_i(2)$ | $p$ for $\hat{\psi}_i(2)$ | $\hat{\psi}_i(3)$ | $p$ for $\hat{\psi}_i(3)$ | $\hat{\psi}_i(4)$ | $p$ for $\hat{\psi}_i(4)$ | $\hat{\psi}_i(5)$ | $p$ for $\hat{\psi}_i(5)$ | $\hat{\psi}_i(6)$ | $p$ for $\hat{\psi}_i(6)$ | $\hat{\psi}_i(7)$ | $p$ for $\hat{\psi}_i(7)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.885 | 0.00 | 0.134 | 0.28 | 0.105 | 0.10 | 0.163 | 0.01 | 0.006 | 0.93 | 0.038 | 0.78 | −0.044 | 0.92 |
| 1 | 0.625 | 0.00 | 0.625 | 0.00 | 0.085 | 0.46 | 0.140 | 0.02 | 0.077 | 0.17 | −0.004 | 0.94 | 0.029 | 0.81 |
| 2 | 0.508 | 0.00 | 0.350 | 0.00 | 0.419 | 0.00 | 0.032 | 0.72 | 0.097 | 0.03 | 0.019 | 0.65 | 0.063 | 0.16 |
| 3 | 0.515 | 0.00 | 0.287 | 0.00 | 0.140 | 0.07 | 0.239 | 0.00 | 0.034 | .60 | 0.030 | 0.37 | 0.043 | 0.16 |
| 4 | 0.791 | 0.00 | 0.165 | 0.10 | 0.295 | 0.00 | 0.112 | 0.12 | 0.160 | 0.03 | 0.045 | 0.43 | 0.010 | 0.75 |
| 5 | 0.567 | 0.00 | 0.757 | 0.00 | 0.057 | 0.61 | 0.250 | 0.00 | 0.062 | 0.40 | 0.139 | 0.06 | 0.044 | 0.45 |
| 6 | 1.076 | 0.01 | 0.711 | 0.11 | 0.856 | 0.01 | 0.415 | 0.13 | 0.241 | 0.17 | 0.112 | 0.52 | 0.277 | 0.11 |
| 7 | 0.522 | 0.03 | 0.684 | 0.41 | 0.988 | 0.28 | 1.095 | 0.09 | 0.350 | 0.51 | 0.198 | 0.56 | 0.325 | 0.33 |
| 8 | 0.451 | 0.00 | −1.014 | 0.00 | −0.062 | 0.66 | −0.745 | 0.50 | 0.128 | .87 | −0.635 | 0.31 | 0.076 | 0.85 |
| 9 | 0.618 | 0.00 | −0.041 | 0.77 | −0.746 | 0.01 | −1.083 | 0.26 | −0.047 | .97 | 0.514 | 0.50 | −0.031 | 0.96 |
| 10 | 0.448 | 0.00 | 0.409 | 0.00 | 0.026 | 0.78 | −0.241 | 0.20 | −1.125 | 0.08 | 0.799 | 0.26 | 0.146 | 0.77 |
| 11 | 0.677 | 0.00 | 0.159 | 0.01 | 0.194 | 0.00 | 0.050 | 0.46 | −0.190 | 0.17 | −0.402 | 0.38 | 0.461 | 0.37 |

[a]Note that the parameter estimates continue past lag 7.

gence (12) to hold is that the noise sequence $\varepsilon_t$ has a finite fourth moment. In practical applications, $N_y$ is the number of years of data, $k$ is the number of iterations of the innovations algorithm (typically on the order of $k = 10$ or 15, see the discussion in section 4), and the convergence in distribution is used to approximate the quantity on the left-hand side of (12) by a normal random variable. Equation (12) can be used to produce confidence intervals and hypothesis tests for the moving average parameters in (2). For example, an $\alpha$-level test statistic rejects the null hypothesis ($H_0$: $\psi_i(u) = 0$) in favor of the alternative ($H_a$: $\psi_i(u) \neq 0$, indicating that the model parameter is statistically significantly different from zero) if $|Z| > z_{\alpha/2}$. The $p$ value for this test is

$$p = P(|Z| > |z|),$$
$$Z \sim \mathcal{N}(0,1), \qquad z = \frac{N_y^{1/2}\hat{\theta}_{k,u}^{(\langle i-k\rangle)}}{W}, W^2 = \frac{\sum_{n=0}^{u-1} \hat{v}_{k,\langle i-k-n\rangle}\left(\hat{\theta}_{k,n}^{(\langle i-k\rangle)}\right)^2}{\hat{v}_{k,\langle i-k-u\rangle}}. \tag{13}$$

The innovations algorithm allows us to identify an appropriate model for the periodic time series at hand, and the $p$ value formula gives us a way to determine which coefficients in the identified PARMA model are statistically significantly different from zero (those with a small $p$ value, say, $p < 0.05$). We illustrate the practical application of these formulae in sections 4 and 5.

## 4. Simulation Study

[10] A detailed simulation study was conducted to investigate the practical utility of the innovations algorithm for model identification in the presence of seasonally correlated data. Data for several different PARMA$_S(p, q)$ models were simulated. For each model, individual realizations of $N_y$ = 50, 100, 300, and 500 years of data were simulated and the innovations algorithm was used to obtain parameter estimates for each realization. In each case, estimates were obtained for $k = 10$, $k = 15$ and $k = 20$ iterations in order to examine the convergence, and $p$ values were computed using (13) to identify those estimates that were statistically significant ($p < 0.05$). Some general conclusions can be drawn from this study. We found that 10 to 15 iterations of the innovations algorithm are usually sufficient to obtain reasonable estimates of the model parameters. We also found that $N_y = 50$ years of monthly or quarterly data give only rough estimates of the model parameters, while $N_y$ = 100 years generally is enough to give good estimates. For the data between 50 and 100 years, the estimates are less accurate but generally adequate for practical applications. In order to illustrate the general quality of those results, we summarize here one particular case of a PARMA$_4$(1,1) model

$$X_{kS+i} = \phi_i X_{kS+i-1} + \varepsilon_{kS+i} + \theta_i \varepsilon_{kS+i-1} \tag{14}$$

where $\{\delta_{kS+i} = \sigma_i^{-1}\varepsilon_{kS+i}\}$ is an iid sequence of normal random variables with mean zero and standard deviation one. The periodic notation $X_{kS+i}$ refers to the (mean zero) simulated data for season $i$ of year $k$. From the above model, a single realization with $N_y$ = 500 years of quarterly data (sample size of $N = N_yS = 500 \cdot 4 = 2000$) was generated.

[11] Table 1 shows the results after $k = 15$ iterations of the innovations algorithm. For season 0 the first four lags are statistically significant, for season 2 and 3 the first three lags are significant, while for season 1 only the first two are

**Table 5.** Parameter Estimates for PARMA Model (18) of Average Monthly Flow Series for the Fraser River Near Hope, British Columbia

| Month | $\hat{\phi}$ | $\hat{\theta}$ | $\hat{\sigma}$ |
|---|---|---|---|
| Oct | 0.198 | 0.687 | 11875.479 |
| Nov | 0.568 | 0.056 | 11598.254 |
| Dec | 0.560 | −0.052 | 7311.452 |
| Jan | 0.565 | −0.050 | 5940.845 |
| Feb | 0.321 | 0.470 | 4160.214 |
| Mar | 0.956 | −0.389 | 4610.209 |
| Apr | 1.254 | −0.178 | 15232.867 |
| May | 0.636 | −0.114 | 31114.514 |
| Jun | −1.942 | 2.393 | 32824.370 |
| Jul | −0.092 | 0.710 | 29712.190 |
| Aug | 0.662 | −0.213 | 15511.187 |
| Sep | 0.355 | 0.322 | 12077.991 |

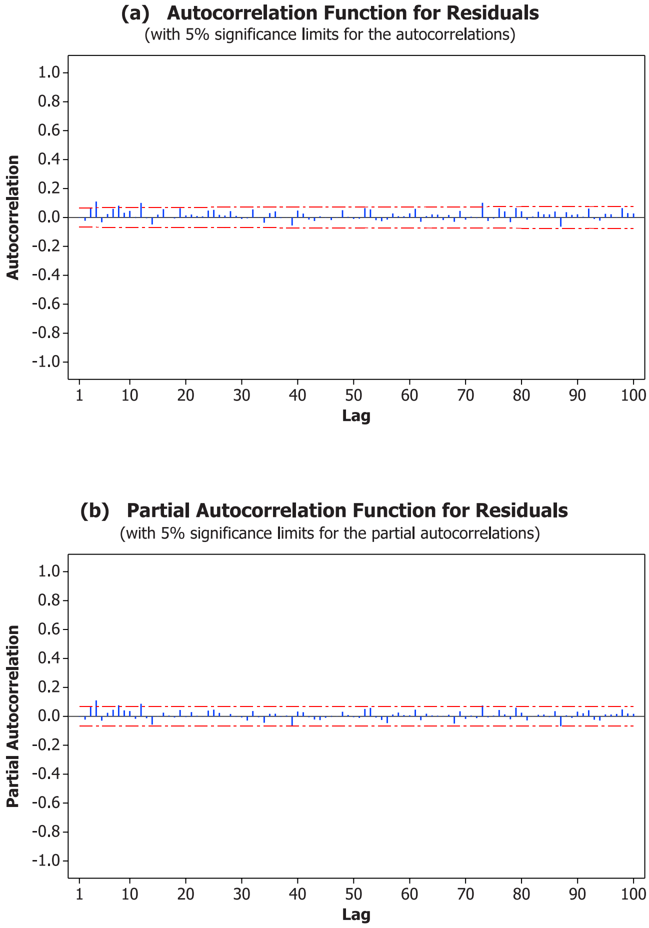


**Figure 2.** (a) ACF for model residuals, showing the bounds $\pm 1.96/\sqrt{N}$, indicate no serial dependence. (b) PACF for model residuals, showing the bounds $\pm 1.96/\sqrt{N}$, indicate no serial dependence.

significant. Since parameter estimates do not generally cut off to (statistically) zero at a certain lag, it is advantageous to seek a parsimonious mixed moving average/autoregressive model. To fit a mixed model (1) using the innovations estimates, we substitute (2) into (1) and then equate the coefficients of the noise $\varepsilon_t$ on both sides to determine the parameters $\theta_t$ and $\phi_t$. This requires that statistically significant values of $\psi_i(j)$ are available for each season $i$ for lags $1 \le \ell \le p + q$. For $p = q = 1$ the resulting equation takes a simplified form

$$\sum_{j=0}^{\infty} \psi_t(j)\varepsilon_{t-j} - \phi_t \sum_{j=0}^{\infty} \psi_{t-1}(j)\varepsilon_{t-1-j} = \varepsilon_t + \theta_t\varepsilon_{t-1} \qquad (15)$$

which leads to

$$\phi_t = \psi_t(2)/\psi_{t-1}(1) \quad \text{and} \quad \theta_t = \psi_t(1) - \phi_t \qquad (16)$$

The actual values of the autoregressive parameters $\phi_i$ and moving average parameters $\theta_i$ for season $i$ along with the estimated parameters obtained by substituting the innovations estimates from Table 1 into (16) are displayed in Table 2. Standardized residuals, $\delta_t$, for this PARMA$_4$(1,1) model can be computed using the equation

$$\hat{\sigma}_t\hat{\delta}_t = X_t - \left(\hat{\phi}_t + \hat{\theta}_t\right)X_{t-1} + \sum_{j=2}^{\infty} (-1)^j \left(\hat{\phi}_{t-j+1} + \hat{\theta}_{t-j+1}\right) \cdot \hat{\theta}_t\hat{\theta}_{t-1}\ldots\hat{\theta}_{t-j+2}X_{t-j} \qquad (17)$$

which was obtained by solving (1) for the innovations and substituting the estimated model parameters for their true values. The PARMA$_S$(1,1) model is the simplest mixed model, and thus is preferred so long as diagnostic plots of the residual autocorrelation (ACF) and/or partial autocorrelation (PACF) indicate no significant serial dependence. For the simulation reported here, this was the case, and hence a PARMA$_4$(1,1) model was judged adequate. The ACF and PACF plots were similar to those in section 5 (see Figure 2).

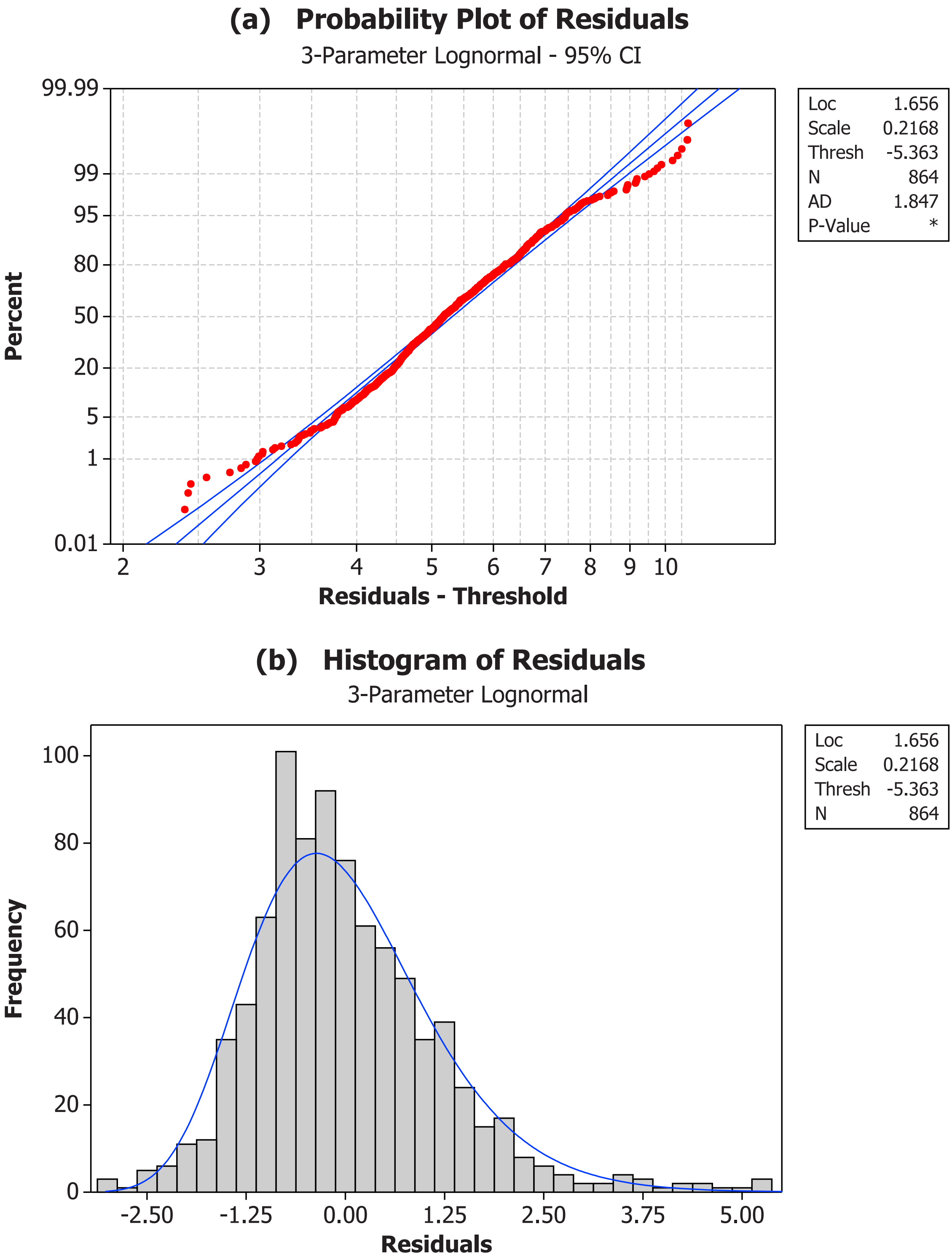
## 5. Application to Modeling of Natural River Flows

[12] Next we model a monthly river flow time series from the Fraser River at Hope, British Columbia. The Fraser River is the longest river in British Columbia, travelling almost 1400 km and sustained by a drainage area covering 220,000 km$^2$. It rises in the Rocky Mountains, at Yellowhead Pass, near the British Columbia-Alta. line and flows northwest through the Rocky Mountain Trench to Prince George, thence south and west to the Strait of Georgia at Vancouver. Its main tributaries are the Nechako, Quesnel, Chilcotin, and Thompson rivers. See http://scitech.pyr.ec.gc.ca/waterweb/ for maps and flow data downloads.

[13] The data are obtained from daily discharge measurements, in cubic meter per second, averaged over each of the respective months to obtain the monthly series. The series contains 72 years of data from October 1912 to September 1984. In the following analysis, $S = 0$ corresponds to October and $S = 11$ corresponds to September. Using the "water year" starting on 1 October is customary for stationary ARMA modeling of river flows, because of low correlation between Fall monthly flows. We adopt the same notation for ease of comparison with those models, but in our case any starting month is equally appropriate, since we explicitly model the seasonal variations. A partial plot of the original data, given Figure 1, shows the cyclic behavior of the monthly flows. The sample mean, standard deviation and autocorrelations at lag 1 and lag 2 are given in Table 3 (see also Figure 8). The nonstationarity of the series is apparent since the mean, standard deviation and correlation functions vary significantly from month to month. Removing the periodicity in mean and variance will not yield a stationary series. Therefore a periodically stationary series model is appropriate. After $k = 20$ iterations, the innovations algorithm yields the innovations estimates and associated $p$ values found in Table 4.

[14] Since the $\hat{\psi}_i$ weights do not generally cut off to (statistically) zero at a certain lag, we choose a parsimonious mixed model that captures the periodic behavior as well as the exponential decay evidenced in the autocorrelation function. We find that a PARMA$_{12}$(1,1) model

$$X_{kS+i} - \phi_i X_{kS+i-1} = \varepsilon_{kS+i} + \theta_i \varepsilon_{kS+i-1} \qquad (18)$$



(a) Probability Plot of Residuals

3-Parameter Lognormal - 95% CI

(b) Histogram of Residuals

3-Parameter Lognormal

**Figure 3.** (a) Lognormal probability plot for model residuals, Fraser River at Hope, British Columbia. (b) Histogram for model residuals, Fraser River at Hope, British Columbia.

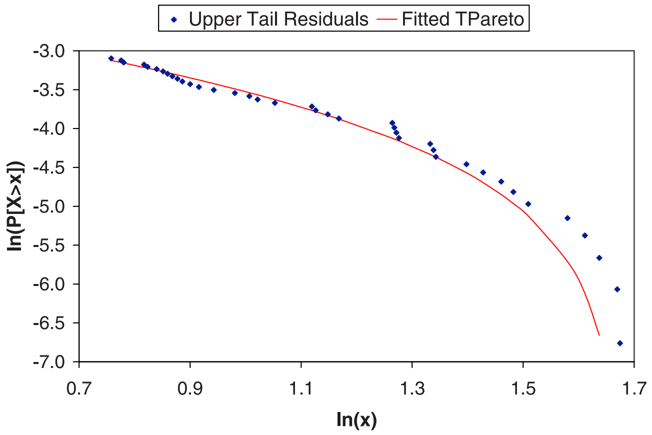


**Figure 4.** Log-log plot of upper residual tail and fitted truncated Pareto distribution, Fraser River at Hope, British Columbia.

is sufficient in adequately capturing the series autocorrelation structure. The physical basis of the river flow process could also be helpful in choosing the appropriate model [*Salas et al.*, 1981; *Salas and Obeysekera*, 1992]. The parameter estimates for this model, obtained using equations (16), are summarized in Table 5. Model residuals were estimated using equation (17). Although the model is periodically stationary, the residuals should be stationary, so the standard 95% confidence limits (that is, $1.96/\sqrt{N}$) still apply. Figure 2 shows the ACF and PACF of the model residuals, respectively. Although a few values lie slightly outside of the 95% confidence bands, there is no apparent pattern, providing some evidence that the PARMA$_{12}$(1,1) model is adequate.

[15] One reason for carefully modeling the river flow time series is to develop the ability to generate synthetic river flows for further analysis. This requires a realistic distributional model for the residuals that can be used to simulate the noise sequence. After exploring a number of possible distributions, we found that a three-parameter lognormal fits the residuals fairly well. A histogram of the residuals showing the best fitting lognormal density curve (scale = 0.217, location = 1.656 and threshold = −5.363), as well as the corresponding probability plot, are shown in Figures 3a and 3b, respectively. On the probability plot, points along the diagonal line (model percentiles equal data percentiles) indicate a good fit. According to this lognormal model, residuals follow the distribution of a random variable $R = -5.363 + e^{(1.656+0.217Z)}$ where $Z \sim \mathcal{N}(0,1)$.

[16] The histogram in Figure 3b shows that the three parameter lognormal gives an acceptable overall fit, but the probability plot in Figure 3a reveals a lack of fit at both tails. This is important for practical applications, since tail behavior of the residuals (or the noise sequence) determines the extreme values of the times series, which govern both droughts and floods. None of the standard probability plots we tried (normal, lognormal, Weibull, gamma, etc.) gave an adequate fit at the tails. To check for a power law probability tail we constructed a Mandelbrot plot of each tail (Figures 4 and 5) as described by *Mandelbrot* [1963] and *Anderson and Meerschaert* [1998]. Suppose that $X_1, \ldots, X_n$ are iid Pareto with distribution function $F(x) = Cx^{-\alpha}$. Then $F(x) = P[X > x] = Cx^{-\alpha}$ and so $\ln F(x) = \ln C - \alpha \ln x$. Sorting the data in decreasing order so that $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ (order statistics) we should have approximately that $x = X_{(r)}$ when $F(x) = r/n$. Then a plot of $\ln X_{(r)}$ versus $\ln(r/n)$ should be approximately linear with slope $-\alpha$. In Figures 4 and 5, the downward curve indicating that a simple power law model for the tail (Pareto, GEV Frechet, $\alpha$ stable) is not appropriate. However, the shape of the plots is consistent with many examples of truncated Pareto distributions found in the geophysics literature [see, e.g., *Aban et al.*, 2006; *Burroughs and Tebbens*, 2001a, 2001b, 2002]. This distribution is appropriate when a power law model is affected by an upper bound on the observations.

[17] In hydrology it is commonly believed that there is an upper bound on precipitation and therefore river flows [see, e.g., *Maidment*, 1993]. A truncated Pareto random variable $X$ has distribution function

$$F_X(x) = P(X \leq x) = \frac{1 - (\gamma/x)^{\alpha}}{1 - (\gamma/\beta)^{\alpha}} \tag{19}$$

and density

$$f_X(x) = \frac{\alpha\gamma^{\alpha}x^{-\alpha-1}}{1 - (\gamma/\beta)^{\alpha}} \tag{20}$$

with $0 < \gamma \leq x \leq \beta < \infty$ and $\gamma < \beta$. *Aban et al.* [2006] develop maximum likelihood estimators (MLE) for the parameters of the truncated Pareto distribution. When a truncated Pareto is fit to the tail of the data, the parameters are estimated by obtaining the conditional maximum likelihood estimate based on the largest-order statistics, representing only the portion of the tail where the truncated Pareto model holds. When $X_{(r)} > X_{(r+1)}$, the conditional maximum likelihood estimator for the parameters of the upper truncated Pareto in (19) based on the $r + 1$ largest-order statistics is given by

$$\hat{\beta} = X_{(1)} \tag{21}$$

$$\hat{\gamma} = r^{1/\hat{\alpha}} X_{(r+1)} \left[ n - (n-r)\left(X_{(r+1)}/X_{(1)}\right)^{\hat{\alpha}} \right]^{-1/\hat{\alpha}} \tag{22}$$

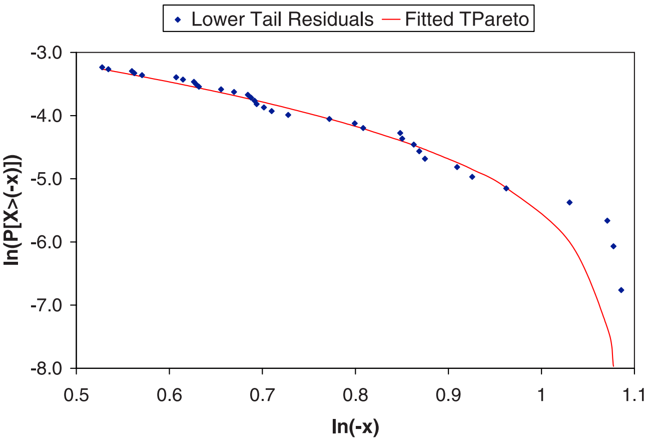


**Figure 5.** Log-log plot of lower residual tail and fitted truncated Pareto distribution, Fraser River at Hope, British Columbia.

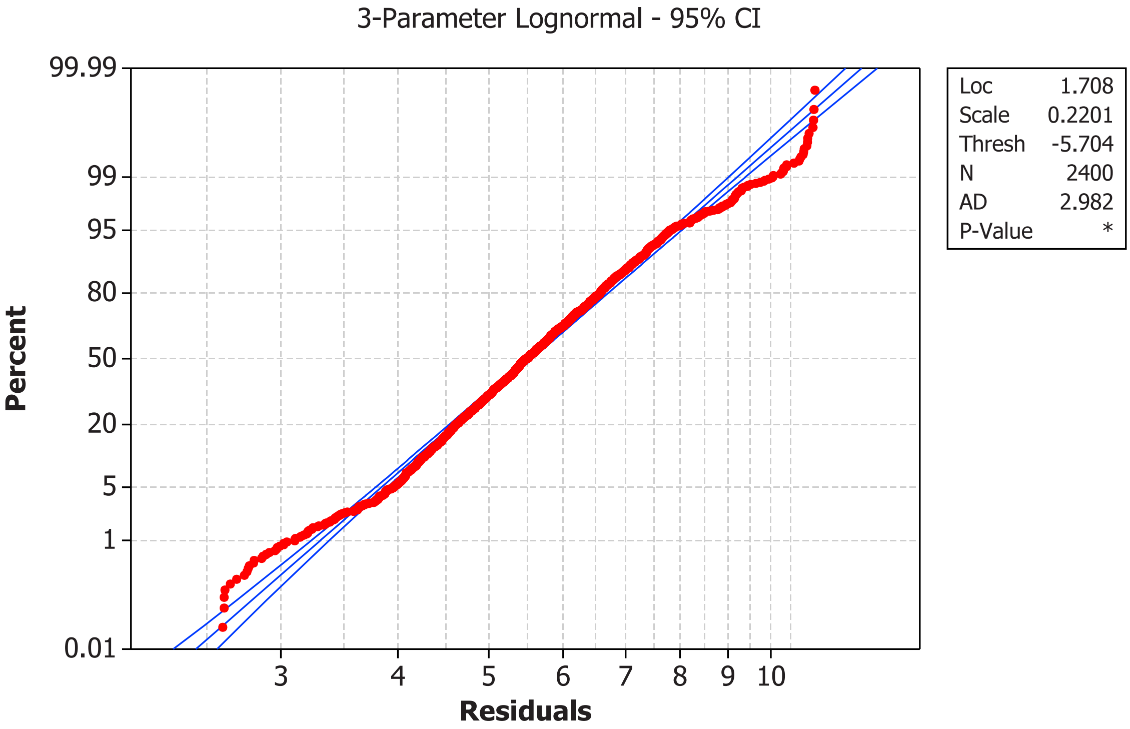


**Figure 6.** Probability plot of simulated noise sequence using the mixed three parameter lognormal and truncated Pareto distributions. Compare to Figure 3a.

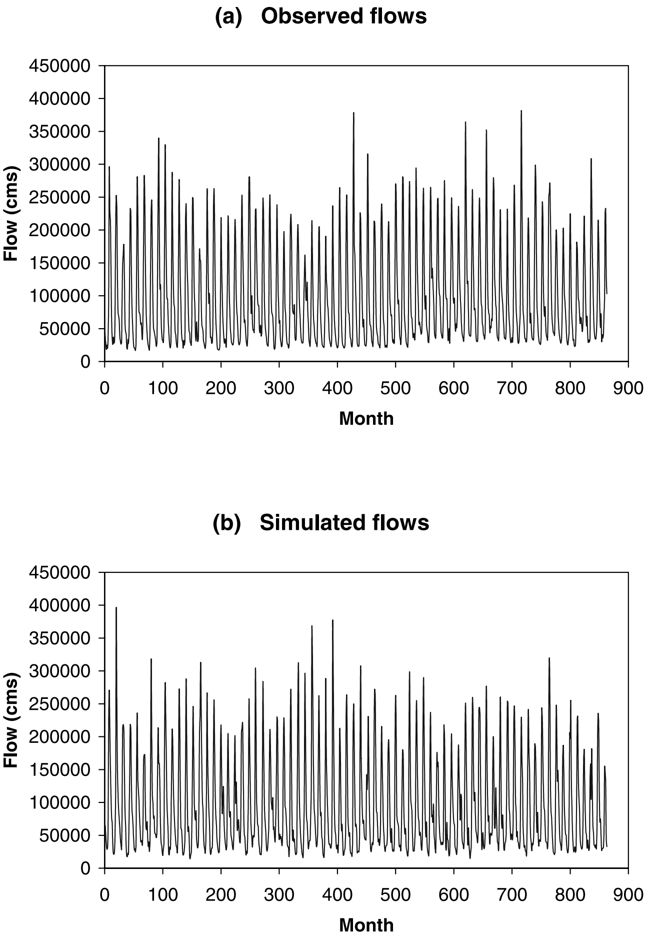


**Figure 7.** (a) Plot of observed monthly river flows for the Fraser River at Hope, British Columbia. (b) Plot of simulated monthly river flows for the Fraser River at Hope, British Columbia. Compare to Figure 7a.

and $\hat{\alpha}$ solves the equation

$$\frac{r}{\hat{\alpha}} + \frac{r\left(X_{(r+1)}/X_{(1)}\right)^{\hat{\alpha}} \ln\left(X_{(r+1)}/X_{(1)}\right)}{1 - \left(X_{(r+1)}/X_{(1)}\right)^{\hat{\alpha}}} - \sum_{i=1}^{r} \left[\ln X_{(i)} - \ln X_{(r+1)}\right] = 0 \qquad (23)$$

[18] The probability plot in Figure 3a shows that a lognormal distribution fits well except for the upper and lower 5% of the residuals. Hence a truncated Pareto was fitted to the upper 5% of the residuals, and another truncated Pareto was fitted to the lower 5% of the residuals after a change of sign. Using the computed values of $-1.697$ = the 5th percentile and 2.122 = the 95th percentile of the three parameter lognormal distribution we fit to the body of the residuals, we determined the cutoff for each fitted distribution. Next we determined that $r = 39$ residuals exceed the 95th percentile, and $r = 34$ residuals fall below the 5th percentile. Then the MLE was used to estimate the parameters ($\hat{\beta} = 5.336$, $\hat{\gamma} = 0.072$, $\hat{\alpha} = 0.722$) of the best fitting truncated Pareto distribution, and the theoretical distribution tail $P(R > r)$ was plotted over the 39 largest positive residuals in Figure 4. In Figure 5 we used the same method to fit a truncated Pareto ($\hat{\beta} = 2.961$, $\hat{\gamma} = 0.291$, $\hat{\alpha} = 1.560$) to the 34 largest negative residuals, after a change of sign. Both of the plots in Figures 4 and 5 indicate an adequate fit.

[19] A mixture distribution with lognormal body and truncated Pareto tails was used to simulate the noise sequence. The mixture has cumulative distribution function (cdf)

$$P(\delta \le r) = \begin{cases} F_-(r) & \text{if } r < -1.697 \\ F_0(r) & \text{if } -1.697 \le r \le 2.122 \\ F_+(r) & \text{if } r > 2.122 \end{cases} \qquad (24)$$

where $F_0$ is the cdf of the lognormal, and $F_+$, $F_-$ are truncated Pareto cdfs of the positive and negative tails, respectively. Recall that the cutoffs in (24) are the 5% and 95% quantiles of the lognormal distribution, which were determined to be the range in which the lognormal provides an adequate fit. The truncated Pareto distributions in (24) were shifted (by $s = 0.172$ on the positive tail and $s = 0.174$ on the negative tail) to make the mixture cdf continuous. Now the noise sequence could be simulated by the inverse cumulative distribution function method $\delta = F^{-1}(U)$ where $U$ is a pseudorandom number uniformly distributed on the unit interval (0,1). However, this is impractical in the present case since the lognormal cdf is not analytically invertible. Instead, we used the Box-Müller method to generate standard normal random variates $Z$ [see *Gentle*, 2003]. Then lognormal random variates were calculated using $\delta = -5.363 + \exp(1.656 + 0.217Z)$. If $R > 2.122$, the 95th percentile of the lognormal, we generated another uniform (0,1) random variate $U$ and substituted $\delta = F_+^{-1}(0.95 + 0.05U)$. If $R < -1.697$, the 5th percentile of the lognormal, we substituted $\delta = F_-^{-1}(0.05U)$. This gives simulated noise sequence $\delta$ with the mixture distribution (24).

[20] Figure 6 shows a probability plot for $N = N_yS$ simulated noise sequence (for $S = 12$ months and $N_y = 100$ years) from the mixture distribution (24). Comparison with Figure 3a shows that the simulated noise sequence is statistically identical to the computed model residuals in terms of distribution. Substituting the simulated noise sequence into the model (18) generates $N_y$ years of simulated river flow. Figure 7 compares a typical synthetic flow, obtained by this simulation procedure, to the original data (Figure 7a). It is apparent that the two time series are statistically similar. In performing this type of autoregressive simulation, it is advantageous to simulate several extra years of river flows and throw out the initial years (100 years in this case), since we did not simulate $X_t$ for $t < 0$. This ensures that the simulated series is periodically stationary. For an alternative approach, one could adapt exercise 8.17 of *Brockwell and Davis* [1991] to the PARMA model. Figure 8 shows the main statistical characteristics (mean, standard deviation and autocorrelations) of a typical synthetic river flow time series obtained by this method, as well as the same statistical measures for the observed time series. It is apparent that this procedure closely reproduces the main statistical characteristics, indicating that the modeling procedure is trustworthy for generating synthetic river flows. Such synthetic river flows

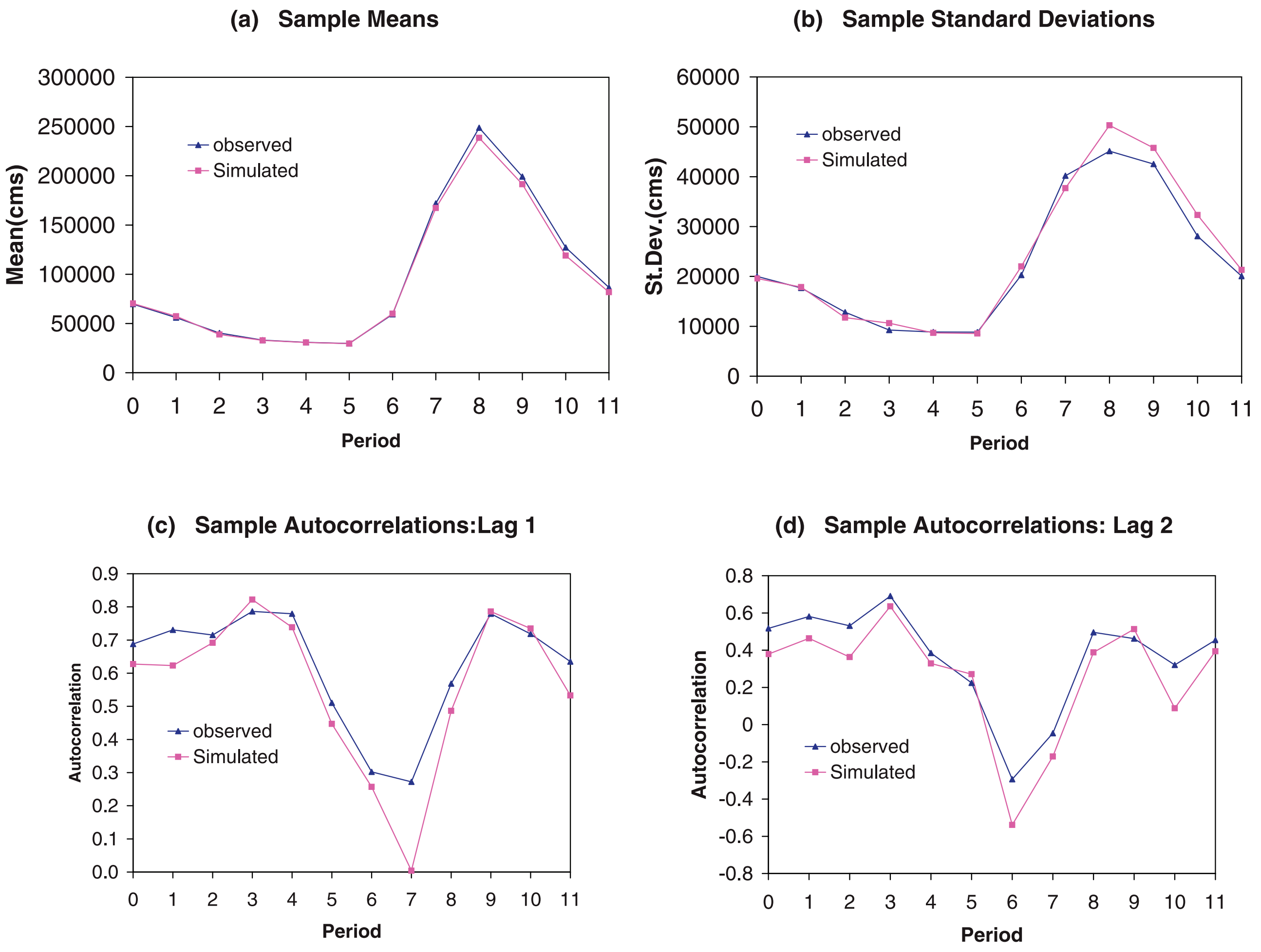


**Figure 8.** (a) Comparison of mean for simulated versus observed monthly river flow data for the Fraser River at Hope, British Columbia. (b) Comparison of standard deviation for simulated versus observed monthly river flow data for the Fraser River at Hope, British Columbia. (c) Comparison of autocorrelation at lag 1 for simulated versus observed monthly river flow data for the Fraser River at Hope, British Columbia. (d) Comparison of autocorrelation at lag 2 for simulated versus observed monthly river flow data for the Fraser River at Hope, British Columbia.

are useful for design of hydraulic structures, for optimal operation of reservoir systems, for calculating the risk of failure of water supply systems, etc. Another calibration test (not performed during this analysis) would be to construct a PARMA model for a subset of the data, and compare the resulting simulated flow to the remainder of the time series. This can be useful to illustrate the effects of parameter/model uncertainty.

## 6. Conclusions

[21] Generation of synthetic river flow data is important in planning, design and operation of water resources systems. PARMA models provides a powerful tool for the modeling of periodic hydrologic series in general and river flow series in particular. In this article, the innovations algorithm estimation procedure, as well as model identification using simulated data from different Gaussian PARMA$_S$($p$, $q$) models, is discussed in detail. A simulation study demonstrates that the innovations algorithm is an efficient and reliable technique for parameter estimation and model identification of PARMA models. In the case of a mixed PARMA process, this model identification technique can be supplemented by modeler experience. Our sample application illustrates that the innovations algorithm is useful for modeling river flow time series. For monthly average river flow data for the Fraser River at Hope, British Columbia, a first-order periodic autoregressive moving average model is adequate to capture the essential features. A mixture of three parameter lognormal body and truncated Pareto tails fits the model residuals nicely. This mixture model is then applied with satisfactory results to generate synthetic monthly river flow records. The methodology presented in this article provides a useful tool for river flow modeling and synthetic river flow data generation. The results allow practitioners and planners to explore realistic decision-making scenarios for a given water resource system.

[22] **Acknowledgments.** We wish to thank the Associate Editor and two anonymous reviewers for comments and suggestions that led to

improvements in the final manuscript. This research is partially supported by the National Science Foundation grant DMS-0139927.

## References

Aban, I. B., M. M. Meerschaert, and A. K. Panorska (2006), Parameter estimation for the truncated Pareto distribution, *J. Am. Stat. Assoc.*, in press.

Adams, G. J., and G. C. Goodwin (1995), Parameter estimation for periodically ARMA models, *J. Time Ser. Anal.*, *16*, 127–145.

Anderson, P. L., and M. M. Meerschaert (1997), Periodic moving averages of random variables with regularly varying tails, *Ann. Stat.*, *25*, 771–785.

Anderson, P. L., and M. M. Meerschaert (1998), Modeling river flows with heavy tails, *Water Resour. Res.*, *34*(9), 2271–2280.

Anderson, P. L., and M. M. Meerschaert (2005), Parameter estimates for periodically stationary time series, *J. Time Ser. Anal.*, *26*, 489–518.

Anderson, P. L., and A. V. Vecchia (1993), Asymptotic results for periodic autoregressive moving-average processes, *J. Time Ser. Anal.*, *14*, 1–18.

Anderson, P. L., M. M. Meerschaert, and A. V. Veccia (1999), Innovations algorithm for periodically stationary time series, *Stochastic Processes Their Appl.*, *83*, 149–169.

Box, G. E., and G. M. Jenkins (1976), *Time Series Analysis: Forcasting and Control*, 2nd ed., Holden-Day, Boca-Raton, Fla.

Brockwell, P. J., and R. A. Davis (1991), *Time Series: Theory and Methods*, 2nd ed., Springer, New York.

Burroughs, S. M., and S. F. Tebbens (2001a), Upper-truncated power law distributions, *Fractals*, *9*, 209–222.

Burroughs, S. M., and S. F. Tebbens (2001b), Upper-truncated power laws in natural systems, *J. Pure Appl. Geophys.*, *158*, 741–757.

Burroughs, S. M., and S. F. Tebbens (2002), The upper-truncated power law applied to earthquake cumulative frequency-magnitude distributions, *Bull. Seismol. Soc. Am.*, *92*, 2983–2993.

Chen, H.-L., and A. R. Rao (2002), Testing hydrologic time series for stationarity, *J. Hydrol. Eng.*, *7*(2), 129–136.

Gentle, J. E. (2003), *Random Number Generation and Monte Carlo Methods*, 2nd ed., Springer, New York.

Jones, R. H., and W. M. Brelsford (1967), Times series with periodic structure, *Biometrika*, *54*, 403–408.

Lund, R. B., and I. V. Basawa (1999), Modeling and inference for periodically correlated time series, in *Asymptotics, Nonparameterics, and Time Series*, edited by S. Ghosh, 37–62, Marcel Dekker, New York.

Lund, R. B., and I. V. Basawa (2000), Recursive prediction and likelihood evaluation for periodic ARMA models, *J. Time Ser. Anal.*, *20*(1), 75–93.

Maidment, D. R. (1993), *Handbook of Hydrology*, McGraw-Hill, New York.

Mandelbrot, B. (1963), The variation of certain speculative prices, *J. Bus.*, *36*, 394–419.

Pagano, M. (1978), On periodic and multiple autoregressions, *Ann. Stat.*, *6*(6), 1310–1317.

Salas, J. D. (1993), Analysis and modeling of hydrologic time series, in *Handbook of Hydrology*, edited by D. R. Maidment, pp. 19.5–19.9, McGraw-Hill, New York.

Salas, J. D., and J. T. B. Obeysekera (1992), Conceptual basis of seasonal streamflow time series models, *J. Hydraul. Eng.*, *118*(8), 1186–1194.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane (1980), *Applied Modeling of Hydrologic Time Series*, Water Resour. Publ., Highlands Ranch, Colo.

Salas, J. D., J. T. B. Obeysekera, and R. A. Smith (1981), Identification of streamflow stochastic models, *J. Hydraul. Div. Am. Soc. Civ. Eng.*, *107*(7), 853–866.

Salas, J. D., D. C. Boes, and R. A. Smith (1982), Estimation of ARMA models with seasonal parameters, *Water Resour. Res.*, *18*, 1006–1010.

Salas, J. D., G. Q. Tabios, III and P. Bartolini (1985), Approaches to multivariate modeling of water resources time series, *Water Resour. Bull.*, *21*, 683–708.

Shao, Q., and R. B. Lund (2004), Computation and characterization of autocorrelations and partial autocorrelations in periodic ARMA models, *J. Time Ser. Anal.*, *25*(3), 359–372.

Thompstone, R. M., K. W. Hipel, and A. I. McLeod (1985), Forecasting quarter-monthly riverflow, *Water Resour. Bull.*, *25*(5), 731–741.

Tiao, G. C., and M. R. Grupe (1980), Hidden periodic autoregressive moving average models in time series data, *Biometrika*, *67*, 365–373.

Tjøstheim, D., and J. Paulsen (1982), Empirical identification of multiple time series, *J. Time Ser. Anal.*, *3*, 265–282.

Troutman, B. M. (1979), Some results in periodic autoregression, *Biometrika*, *6*, 219–228.

Ula, T. A. (1990), Periodic covariance stationarity of multivariate periodic autoregressive moving average processes, *Water Resour. Res.*, *26*(5), 855–861.

Ula, T. A. (1993), Forcasting of multivariate periodic autoregressive moving average processes, *J. Time Ser. Anal.*, *14*, 645–657.

Ula, T. A., and A. A. Smadi (1997), Periodic stationarity conditions for periodic autoregressive moving average processes as eigenvalue problems, *Water Resour. Res.*, *33*(8), 1929–1934.

Ula, T. A., and A. A. Smadi (2003), Identification of periodic moving average models, *Commun. Stat. Theory Methods*, *32*(12), 2465–2475.

Vecchia, A. V. (1985a), Periodic autoregressive-moving average (PARMA) modelling with applications to water resources, *Water Resour. Bull.*, *21*, 721–730.

Vecchia, A. V. (1985b), Maximum likelihood estimation for periodic moving average models, *Technometrics*, *27*(4), 375–384.

Vecchia, A. V., and R. Ballerini (1991), Testing for periodic autocorrelations in seasonal time series data, *Biometrika*, *78*(1), 53–63.

P. L. Anderson, Department of Mathematics and Computer Science, Albion College, Albion, MI 49224, USA. (panderson@albion.edu)

M. M. Meerschaert, Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand. (mcubed@maths.otago.ac.nz)

Y. G. Tesfaye, Graduate Program of Hydrologic Sciences, University of Nevada, Reno, NV 89557, USA. (yonas@unr.edu)