Preservation of the Rescaled Adjusted Range
2. Simulation Studies Using Box-Jenkins Models

KEITH WILLIAM HIPEL

Department of Systems Design, University of Waterloo, Waterloo, Ontario, Canada

ANGUS IAN MCLEOD

Department of Mathematics, University of Western Ontario, London, Ontario, Canada

It is demonstrated that autoregressive moving average (Arma) models do preserve the rescaled adjusted range (RAR) or equivalently the Hurst coefficient K. Arma models are fit to 23 geophysical time series, and by using Monte Carlo techniques and a specified statistical test it is shown that the observed RAR or K is retained by the models. The empirical cumulative distribution function (ECDF) for these statistics can be calculated as closely as is required to the theoretical distribution. Furthermore, the distribution of the RAR is a function of the time series length N and the parameter values of the particular Arma process being considered. Various estimates for the Hurst coefficient are compared for the 23 geophysical data sets.

INTRODUCTION

A major challenge in stochastic hydrology is to determine models that preserve important historical statistics such as the rescaled adjusted range (RAR), or equivalently the Hurst coefficient K. In an accompanying paper labeled part 1 [*McLeod and Hipel,* 1978*a*] the development of fractional Gaussian noise (FGN) processes as one means of possibly retaining the RAR and K is appraised by *McLeod and Hipel* [1978*a*]. Furthermore, computer algorithms for exactly simulating FGN and obtaining maximum likelihood estimates (MLE) for the model parameters are given by *Hipel and McLeod* [1978] in part 3. The purpose of this paper is to demonstrate that the autoregressive moving average (Arma) models do preserve the RAR and K when these processes are fit to a wide range of geophysical time series. Therefore in many practical situations it may be unnecessary to employ FGN processes in order to retain the Hurst statistics.

The theory, model construction stages, and simulation of Box-Jenkins models are briefly surveyed. Following the identification, estimation, and diagnostic check stages of model development, Arma models are determined for 23 geophysical time series. Simulation studies are then performed to determine the small sample empirical cumulative distribution function (ECDF) of the RAR or K for various Arma models. The ECDF for these statistics is shown to be a function of the time series length N and the parameter values of the specific Arma process being considered. Furthermore, it is possible to determine as accurately as desired the distribution of the RAR or K. A theorem is given to obtain confidence intervals for the ECDF in order to guarantee a prescribed precision. Then it is shown by utilizing simulation results and a given statistical test that Arma models do preserve the observed RAR or K of the 23 geophysical time series. Finally, various estimates for the Hurst coefficient are estimated and compared for the 23 given time series.

BOX-JENKINS MODELING

Theory

In 1970, Box and Jenkins published a text that describes a family of linear stochastic models. These models are often collectively referred to as the Box-Jenkins models. However, if the process is stationary, the label autoregressive moving average (Arma) is employed, while if differencing is required to eliminate nonstationarity, the process is called an autoregressive integrated moving average (Arima) model.

Research related to the Hurst phenomenon, RAR, and K has involved the stochastic analysis of nonseasonal data such as average annual river flows, tree ring indicies, and mud varve data. Therefore only the mathematical theory of nonseasonal Box-Jenkins models is now summarized.

Let $z_1, z_2, \cdots, z_{t-1}, z_t, z_{t+1}, \cdots, z_N$ be a discrete time series measured at equal time intervals. Suppose z_t can be modeled by a nonseasonal multiplicative Box-Jenkins model of the form

$$\phi(B)\{[(1 - B)^d z_t^{(\lambda)}] - \mu\} = \theta(B)a_t \quad (1)$$

or

$$\phi(B)(w_t - \mu) = \theta(B)a_t \quad (2)$$

where

$z_t^{(\lambda)}$ appropriate transformation of z_t such as a Box-Cox transformation [*McLeod,* 1974; *Box and Cox,* 1964] (no transformation is a possible option);

t discrete time;

B backward shift operator defined by $B^k z_t^{(\lambda)} = z_{t-k}^{(\lambda)}$ for $k = 1, 2, \cdots$;

μ mean level of the process, usually taken as the average of the w_t series;

a_t identically independently distributed white noise residual with mean 0 and variance σ_a^2 (written as IID (0, σ_a^2)), often the residuals are assumed to be normally independently distributed (denoted by NID (0, σ_a^2));

$\phi(B) = 1 - \phi_1 B - \phi_2 B - \cdots - \phi_p B^p$ nonseasonal autoregressive (AR) operator or polynomial of order p such that the roots of the characteristic equation $\phi(B) = 0$ lie outside the unit circle for nonseasonal stationarity and the ϕ_i, $i = 1, 2, \cdots, p$ are the nonseasonal AR parameters;

$(1 - B)^d = \nabla^d$ nonseasonal differencing operator of order d to produce nonseasonal stationarity of the dth differences, usually $d = 0, 1,$ or 2;

Paper number 7W0520.
0043-1397/78/037W-0520$01.00

$w_t = \nabla^d z_t^{(\lambda)}$ stationary series formed by differencing the $z_t^{(\lambda)}$ series, $n = N - d$ is the number of terms in the series;

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ nonseasonal moving average (MA) operator or polynomial of order q such that the roots of $\theta(B) = 0$ lie outside the unit circle for invertibility and θ_i, $i = 1, 2, \cdots, q$ are the nonseasonal MA parameters.

The notation (p, d, q) is used to list the orders of the nonseasonal operators of the Arima model given in (1) and (2). When a nonseasonal process is stationary and requires no differencing, the Arma model is denoted by (p, q) or equivalently $(p, 0, q)$. A pure nonseasonal AR process of order p without differencing is often written as AR(p). Likewise, a nonseasonal MA process of order q is sometimes labeled as MA(q). Of course, an AR(p) model can be equivalently represented by the notation $(p, 0)$ or $(p, 0, 0)$, while a MA(q) process can also be denoted by $(0, q)$ or $(0, 0, q)$.

Model Construction

When determining a Box-Jenkins model for a particular time series, it is recommended to adhere to the identification, estimation, and diagnostic check stages of model development [*Box and Jenkins*, 1970; *Box and Tiao*, 1973]. Recently, *Hipel et al.* [1977*a*] have presented some new procedures to simplify and also to substantiate the three stages of model construction. For example, in addition to the autocorrelation function (ACF) and partial autocorrelation function (PACF) it is recommended to employ the inverse autocorrelation function (IACF) and the inverse partial autocorrelation function (IPACF) for model identification. At the estimation stage, parameters can be estimated more efficiently by using the modified sum of squares technique [*McLeod*, 1977]. *Hipel et al.* [1977*a*] also describe sensitive diagnostic tests for checking the assumptions of independence, homoscedasticity (i.e., constant variance), and normality of the model residuals. For instance, because the distribution of the residual autocorrelation function is now known [*McLeod*, 1978], valuable diagnostic checks are available for testing the key assumption of independence of the model residuals.

In practice, heteroscedasticity and nonnormality of the residuals can be removed by a Box-Cox transformation that is defined by [*McLeod*, 1974, p. 14; *Box and Cox*, 1964]

$$z_t^{(\lambda)} = \lambda^{-1} [(z_t + \text{const})^\lambda - 1] \qquad \lambda \neq 0$$
$$z_t^{(\lambda)} = \ln (z_t + \text{const}) \qquad \lambda = 0 \qquad (3)$$

where const is a constant. The parameter const is usually assigned a magnitude that will cause all the values in the time series to be greater than zero. In many situations, all of the observations are positive, and therefore the constant is set equal to zero. If it is suspected what type of power transformation is required to remove anomalies in the residuals, then λ can be given a fixed value. For example, a square root transformation (i.e., $\lambda = 0.5$) or perhaps natural logarithms (i.e., $\lambda = 0$) may be the appropriate transformation to invoke. Alternatively, a MLE for λ along with a corresponding standard error can be calculated at the estimation stage.

McLeod et al. [1977] have demonstrated the utility of the contemporary modeling techniques given by *Hipel et al.* [1977*a*] by applying these procedures to both nonseasonal and seasonal time series. A constrained nonseasonal Arma model is determined for average annual river flows, a nonmultiplicative Arma process is fit to a yearly sunspot series, and a multiplicative Arima model is found for modeling monthly international airline passengers data. In all three cases a better model is obtained than was previously cited in the literature. For a comprehensive presentation of the development and use of the Box-Jenkins models in water resources the reader should refer to the textbook by *Hipel and McLeod* [1977].

Simulation Techniques

Recently, *McLeod and Hipel* [1978*b*] have developed improved simulation procedures for generating synthetic traces from Box-Jenkins models. These techniques have been labeled WASIM (Waterloo simulation procedure 1) and WASIM2 (Waterloo simulation procedure 2) and are used for the Monte Carlo studies in the ensuing sections of this paper. The exact theoretical development of WASIM1 and WASIM2 and the types of situations where it is advantageous to use one simulation technique in preference to the other are described in detail by *McLeod and Hipel* [1978*b*]. In addition, the computer programs along with other supporting subroutines and documentation for these contemporary simulation procedures have been listed in the microfiche edition of the paper.

McLeod and Hipel [1978*b*] cite many distinct advantages for using WASIM1 and WASIM2. Of prime importance is the fact that random realizations of the underlying stochastic process are employed as starting values. Because fixed initial values are not utilized, systematic bias is not introduced into the generated data.

In this paper it is assumed that the Box-Jenkins model residuals are approximately normally distributed. When it is necessary a Box-Cox transformation is invoked to insure that the normality assumption is satisfied. WASIM1 is employed

TABLE 1. Annual River Flows and Miscellaneous Geophysical Data

Code Name	Type	Location	Period	N
Mstouis	Mississippi River	St. Louis, Missouri	1861–1957	96
Neumunas	Neumunas River	Smalininkai, USSR	1811–1943	132
Danube	Danube River	Orshava, Romania	1837–1957	120
Rhine	Rhine River	Basle, Switzerland	1807–1957	150
Ogden	St. Lawrence River	Ogdensburg, New York	1860–1957	97
Gota	Gota River	Sjotorp-Vanersburg, Sweden	1807–1957	150
Espanola	mud varves	Espanola, Ontario	−471 to −820 (Swedish time)	350
Temp	temperature data			
Precip	precipitation	English Midlands	1698–1952	255
Sunyr	yearly sunspots	London, England	1813–1912	100
Minimum	minimum flows of Nile River	sun	1798–1960	163
		Rhoda, Egypt	622–1469	848

TABLE 2. Tree Ring Indicies Data

Code Name	Type of Tree	Location	Period	N
Snake	Douglas fir	Snake River Basin	1282–1950	669
Exshaw	Douglas fir	Exshaw, Alberta, Canada	1460–1965	506
Naramata	Ponderosa pine	Naramata, B.C., Canada	1451–1965	515
Dell	Limber pine	Dell, Montana	1311–1965	655
Lakeview	Ponderosa pine	Lakeview, Oregon	1421–1964	544
Ninemile	Douglas fir	Nine Mile Canyon, Utah	1194–1964	771
Eaglecol	Douglas fir	Eagle, Colorado	1107–1964	858
Navajo	Douglas fir	Navajo National Monument (Belatakin), Arizona	1263–1962	700
Bryce	Ponderosa pine	Bryce Water Canyon, Utah	1340–1964	625
Tioga	Jeffrey pine	Tioga Pass, California	1304–1964	661
Bigcone	Big cone spruce	Southern California	1458–1966	509
Whitemtn	Bristlecone pine	White Mountains, California	800–1963	1164

for the pure MA models, while WASIM2 is utilized for pure AR and mixed Arma models.

ARMA MODELING OF GEOPHYSICAL PHENOMENA

In this section, Arma models are determined for 23 geophysical time series. Table 1 lists the average annual river flows and miscellaneous geophysical phenomena that are modeled. The river flows are the longer records that are available in a paper by *Yevjevich* [1963]. Although the flows were converted to cubic meters per second, it is irrelevant which units of measurement are used, since the parameter estimates for the Arma process fit to the data are independent of the measuring system used. The mud varve, temperature, rainfall, sunspot numbers, and minimum flows of the Nile River are obtained from articles by *De Geer* [1940], *Manley* [1953, pp. 255–260], *Kendall and Stuart* [1963, p. 343], *Waldmeier* [1961], and *Toussoun* [1925], respectively.

Table 2 lists 12 sets of tree ring indices comprising six different species of trees from western North America. The indices labeled Snake are from *Schulman*'s [1956, p. 77] book, and the rest were selected from a report by *Stokes et al.* [1973].

By employing the three stages of model construction that are suggested by *Hipel et al.* [1977*a*] and *McLeod et al.* [1977] the proper Box-Jenkins model is determined for each of the 23 time series. An inspection of the ACF for each data set at the identification stage reveals that differencing of the data is not required. Therefore stationary Arma (p, q) models, or equivalently Arima $(p, 0, q)$ processes, successfully model all the time series considered.

When obtaining MLE estimates for the Arma model parameters, *Hipel et al.* [1977*a*] recommend employing the modified sum of squares method which was recently developed by *McLeod* [1977]. *McLeod et al.* [1977] use the modified sum of squares technique when determining MLE of the parameters of the models examined in their paper. Because the modified sum of squares technique was evolved subsequent to the development of the models in this article, the unconditional sum of squares method [*Box and Jenkins,* 1970, chapter 7] was used to obtain MLE for the parameters of the models for the 23 geophysical series. However, for the data sets considered, the two estimation techniques produce almost the same results. Nevertheless, as was demonstrated by *McLeod et al.* [1977], in certain situations this may not be the case.

Table 3 catalogs the type of Arma model, Box-Cox transformation, and parameter estimates and standard errors for each data set. The standard errors are given in parentheses. For all the Box-Cox transformations, the constant is set equal to zero. When $\lambda = 1$ there is no transformation, while $\lambda = 0$ means that natural logarithms are taken of the data. Whenever a MLE of λ is calculated, the standard error is included in parentheses.

By utilizing (1) and (3) and the parameter estimates in Table 3 the finite difference equation can be written in operator form for each model. For instance, the Arma model for Ogden River is

$$(1 - 0.626B - 0.184B^3)(z_t - 6818.69) = a_t \qquad (4)$$

where 6818.69 is the mean level of the Ogden data. As was explained by *McLeod et al.* [1977], ϕ_2 is constrained to zero in order to have a parsimonious model. If a (3, 0) process is estimated with ϕ_2 not set equal to zero, the ϕ_2 parameter is not significantly different from zero and therefore should not be included in the model. It is interesting to note that an examination of the ACF, PACF, IACF, and IPACF reveals at the identification stage that probably a (3, 0) model with ϕ_2 confined to zero is the best model to estimate. Notice also for Sunyr that the selected model is (9, 0) with ϕ_3–ϕ_8 constrained to zero.

DISTRIBUTION OF THE RAR OR K

Suppose the determination of the exact distribution of the RAR (i.e., $\bar{R}_N^*$) or K is required. The expected value of $\bar{R}_N^*$ is now known theoretically for both an independent and a symmetrically correlated Gaussian process [*Anis and Lloyd,* 1976]. At present the cumulative distribution function (CDF) of $\bar{R}_N^*$ for a white noise process and in general any Arma model is analytically intractable. However, by simulation it is possible to determine as accurately as is desired for practical purposes the CDF for $\bar{R}_N^*$. Because both $\bar{R}_N^*$ and K are functions of N, their CDF are defined for a particular length of series N. The CDF for $\bar{R}_N^*$ is

$$F = F(r; N, \phi, \theta) = Pr(\bar{R}_N^* \leq r) \qquad (5)$$

where

- N length of each individual time series;
- ϕ set of known AR parameters;
- θ set of known MA parameters;
- r any possible value of $\bar{R}_N^*$.

As was mentioned previously, when simulating a time series of length N it is recommended to employ the improved simulation techniques of *McLeod and Hipel* [1978*b*]. In this section, WASIM1 is utilized for the $(0, q)$ models, while WASIM2 is used for the $(p, 0)$ and (p, q) processes. Because the RAR of K

TABLE 3. Arma Models Fit to the Geophysical Data

Code Name	Model	λ*	Parameter	Value*	Parameter	Value*	Parameter	Value*
Mstouis	(0, 1)	1.0	θ_1	−0.309 (0.094)				
Neumunas	(0, 1)	0.0	θ_1	−0.222 (0.086)				
Danube	(0, 0)	1.0						
Rhine	(0, 0)	1.0						
Ogden	(3, 0)	1.0	ϕ_1	0.626 (0.083)	ϕ_2	0.0	ϕ_3	0.184 (0.086)
Gota	(2, 0)	1.0	ϕ_1	0.591 (0.079)	ϕ_2	−0.274 (0.080)		
Espanola	(1, 1)	0.0	ϕ_1	0.963 (0.016)	θ_1	0.537 (0.051)		
Temp	(0, 2)	1.0	θ_1	−0.115 (0.063)	θ_2	−0.202 (0.057)		
Precip	(0, 0)	0.0						
Sunyr	(9, 0)	1.0	ϕ_1	1.219 (0.060)	ϕ_2	−0.508 (0.056)	ϕ_9	0.232 (0.029)
Minimum	(2, 1)	−0.778 (0.316)	ϕ_1	1.254 (0.060)	ϕ_2	−0.279 (0.051)	θ_1	0.842 (0.049)
Snake	(3, 0)	1.0	ϕ_1	0.352 (0.039)	ϕ_2	0.093 (0.041)	ϕ_3	0.100 (0.039)
Exshaw	(1, 1)	1.0	ϕ_1	0.725 (0.067)	θ_1	0.395 (0.090)		
Naramata	(2, 0)	1.0	ϕ_1	0.196 (0.044)	ϕ_2	0.131 (0.044)		
Dell	(2, 0)	1.0	ϕ_1	0.367 (0.039)	ϕ_2	0.185 (0.039)		
Lakeview	(3, 0)	0.717 (0.130)	ϕ_1	0.525 (0.038)	ϕ_2	0.0	ϕ_3	0.143 (0.039)
Ninemile	(2, 1)	0.684 (0.060)	ϕ_1	1.225 (0.063)	ϕ_2	−0.274 (0.047)	θ_1	0.850 (0.049)
Eaglecol	(2, 1)	0.624 (0.054)	ϕ_1	1.156 (0.114)	ϕ_2	−0.237 (0.082)	θ_1	0.693 (0.103)
Navajo	(1, 1)	1.0	ϕ_1	0.683 (0.082)	θ_1	0.424 (0.103)		
Bryce	(1, 0)	1.366 (0.107)	ϕ_1	0.598 (0.033)				
Tioga	(1, 0)	1.458 (0.098)	ϕ_1	0.556 (0.033)				
Bigcone	(2, 0)	1.0	ϕ_1	0.375 (0.044)	ϕ_2	0.159 (0.044)		
Whitemtn	(1, 1)	1.414 (0.061)	ϕ_1	0.641 (0.086)	θ_1	0.408 (0.104)		

* The parenthetical values are standard errors.

is independent of the variance of the innovations, any value of σ_a^2 may be used. Consequently, it is simplest to set $\sigma_a^2 = 1$ and hence to assume that the residuals are NID (0, 1).

Suppose that $\bar{N}$ simulations of length N are generated for a specific Arma model and the $\bar{N}$ RARs $\bar{R}_{N1}^*, \bar{R}_{N2}^*, \cdots, \bar{R}_{N\bar{N}}^*$ are calculated for each of the simulated series. If the sample of RAR is reordered such that $\bar{R}_{N(1)}^* \leq \bar{R}_{N(2)}^* \leq \cdots \leq \bar{R}_{N(\bar{N})}^*$, it is known that the MLE of F is given by the ECDF [*Gnedenko*, 1968, pp. 444–451]:

$$
\begin{aligned}
F_{\bar{N}} &= F_{\bar{N}}(r; N, \boldsymbol{\phi}, \boldsymbol{\theta}) = 0 \qquad r \leq \bar{R}_{N(1)}^* \\
F_{\bar{N}} &= F_{\bar{N}}(r; N, \boldsymbol{\phi}, \boldsymbol{\theta}) = k/N \qquad \bar{R}_{N(k)}^* < r \leq \bar{R}_{N(k+1)}^* \qquad (6) \\
F_{\bar{N}} &= F_{\bar{N}}(r; N, \boldsymbol{\phi}, \boldsymbol{\theta}) = 1 \qquad r > \bar{R}_{N(\bar{N})}^*
\end{aligned}
$$

The Kolmogoroff theorem [*Gnedenko*, 1968, p. 450] can be used to obtain confidence intervals for $F_{\bar{N}}$ and to indicate the number of samples $\bar{N}$ necessary to guarantee a prescribed accuracy. This theorem states that if $\bar{N}$ is moderately large (it has been shown that $\bar{N} > 100$ is adequate), then

$$
Pr(\max_r |F_{\bar{N}} - F| < \epsilon/N^{1/2}) \simeq K(\epsilon) \qquad (7)
$$

where

$$
K(\epsilon) = 0 \qquad \epsilon \leq 0
$$

$$
K(\epsilon) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\epsilon^2} \qquad \epsilon > 0
$$

For example, when $\epsilon = 1.63$, then $K(\epsilon) = 0.99$. If $\bar{N} = 10^4$ simulations are done for a series of length N, then by Kolmogoroff's theorem, all the values of $F_{\bar{N}}$ are accurate to at least within 0.0163 with probability 0.99.

In actual simulation studies it is useful to examine the convergence of $F_{\bar{N}}$ by printing out a summary of the ECDF for increasing values of $\bar{N}$ (such as $\bar{N}$ = 100, 200, 500, 1000, 2000, $\cdots$) until sufficient accuracy has been obtained. To curtail the computer time required in simulations, there are efficient algorithms available called 'quicksorts' [*Knuth*, 1973] for ordering the sample values for the RAR.

If simulation studies are done for $\bar{R}_N^*$, the ECDF for K can be obtained from the transformation

$$
K = \log \bar{R}_N^*/\log (N/2) \qquad (8)
$$

Alternatively, when the ECDF for K is known then the ECDF for $\bar{R}_N{}^*$ can be calculated by substituting each value of K into

$$\bar{R}_N{}^* = (N/2)^K \quad (9)$$

The tables of various ECDF for different types of Arma models are listed in the appendix on microfiche.[1] In Table 1 of the appendix (on microfiche) the ECDF of K is shown for various values of N for white noise that is NID $(0, \sigma_a{}^2)$. For each value of N (i.e., each row) in that table an ECDF is determined using $\bar{N} = 10^4$ samples of length N. By substituting all values for K in this table into (9) the ECDF for the RAR can be found for each value of N.

When a particular time series is modeled by an Arma process other than white noise, the ECDF for either $\bar{R}_N{}^*$ or K can be calculated by simulation for each desired value of N. For example, Tables 2–10 of the appendix show the ECDF for the RAR for various values of N for Markov processes with $\phi_1 = 0.1, 0.2, \cdots, 0.9$, respectively. In all the tables for a particular value of N the number of samples $\bar{N}$ simulated is 10^4.

The ECDF tables (in the appendix) can be used to make statistical inferences about $\bar{R}_N{}^*$, or equivalently K. For instance, the 95% confidence interval for $\bar{R}_N{}^*$ with $N = 100$ for a Markov (1, 0) process with $\phi_1 = 0.4$ can be determined by utilizing Table 5 of the appendix. Opposite $N = 100$ select the values of $\bar{R}_N{}^*$ below the 0.025 and 0.975 quantiles. The 95% confidence interval for the RAR is then 9.85–24.02. By substituting these interval limits into (8) the 95% confidence interval for K is 0.585–0.813.

The ECDF tables illustrate certain properties of the RAR or K. In Table 1 of the appendix an examination of the median for K for white noise definitely shows that K slowly decreases asymptotically toward 0.5 with increasing N and is consequently a function of N. Because of this a separate ECDF must be developed for each value of N for a specified process. Note that the median values for K in Table 1 of the appendix are almost identical with the values of K tabulated in Table 6 in part 1 [*McLeod and Hipel*, 1978*a*]. These latter values of K are calculated by using (8) when the exact theoretical expected values of $\bar{R}_N{}^*$ are found from a formula given by *Anis and Lloyd* [1976] and also by employing simulation techniques to estimate $E(K)$. As can be seen from a perusal of Table 6 of *McLeod and Hipel* [1978*a*], the expected value of K is obviously a function of N and decreases in magnitude with increasing N.

It can be proven theoretically that for any Arma process the RAR or K is a function of the time series length N and the AR and MA parameters [*Hipel*, 1975, Appendix B]. This fact is confirmed by the ECDF for the RAR for various Markov processes in Tables 2–10 of the appendix. It can be seen that the median and all other values of the RAR at any quantile for all of the models increase in value for increasing N. Furthermore, the distribution of the RAR is also a function of the value of the AR parameter ϕ_1.

PRESERVATION OF THE RAR AND K BY ARMA MODELS

By employing the ECDF of the RAR or K in conjunction with a specified statistical test it is now shown that Arma models do preserve the historically observed Hurst statistics. Because the Hurst coefficient K is widely cited in the literature, the research results for this statistic are described. However, K and $\bar{R}_N{}^*$ are connected by the simple transformation given in (9), and therefore preservation of either statistic automatically implies retention of the other by an Arma model.

The Arma models fit to 23 geophysical time series ranging in length from $N = 96$ to $N = 1164$ are listed in Table 3. For exactly the same N as the historical data, 10^4 simulations are done for each model to determine the ECDF of K, or equivalently $\bar{R}_N{}^*$. The probability p_i of having K for the ith model greater than the K calculated for the ith historical series is determined from the ith ECDF as

$$Pr(K > K_i{}^{obs} | \text{model}) = p_i \quad (10)$$

where $K_i{}^{obs}$ is the K value calculated for the ith observed historical time series. If the chosen Arma model is correct, then by definition, p_i would be uniformly distributed on (0, 1). For k time series it can be shown [*Fisher*, 1970, p. 99] that

$$-2 \sum_{i=1}^{k} \ln p_i \simeq \chi_{2k}{}^2 \quad (11)$$

Significance testing can be done by using (11) to determine whether the observed Hurst coefficient or the RAR is preserved by Arma models. The test could fail if the incorrect model were fit to the data (for example, if the Ogden data were incorrectly modeled by a (1, 0) process with $\phi_1 = 0.4$) or if Arma models do not retain the Hurst K. Careful model selection was done, thereby largely eliminating the former reason for test failure. If it is thought (as was suggested by *Mandelbrot and Wallis* [1968]) that the observed K is larger than that implied by an appropriate Brownian domain model, then a one-tailed rather than a two-tailed test may be performed.

The results of the χ^2 test in (11) for the 23 geophysical phenomena confirm that there is no evidence that the observed K, or equivalently the RAR, is not adequately preserved by the ARMA models. Table 4 summarizes the information used in the test. The observed Hurst coefficient, $E(K)$, from the simulations and the p_i value are listed for each of the time series. In Table 5 it can be seen that the calculated χ^2 value from (11) is not significant at the 5% level of significance for the 23 time

TABLE 4. Geophysical Time Series Calculations

Code Name	N	Observed K	Arma Model $E(K)$	p_i
Mstouis	96	0.648	0.667	0.624
Neumunas	132	0.660	0.649	0.420
Danube	120	0.633	0.613	0.534
Rhine	150	0.614	0.609	0.468
Ogden	97	0.894	0.832	0.149
Gota	150	0.689	0.659	0.283
Espanola	350	0.855	0.877	0.674
Temperature	255	0.694	0.646	0.157
Precip	100	0.618	0.610	0.434
Sunspot numbers	163	0.723	0.768	0.728
Minimum	848	0.815	0.786	0.264
Snake	669	0.687	0.693	0.559
Exshaw	506	0.637	0.702	0.938
Naramata	515	0.595	0.649	0.905
Dell	655	0.687	0.694	0.569
Lakeview	544	0.706	0.729	0.709
Ninemile	771	0.740	0.726	0.378
Eaglecol	858	0.645	0.747	0.995
Navajo	700	0.653	0.670	0.660
Bryce	625	0.732	0.698	0.203
Tioga	661	0.701	0.687	0.362
Bigcone	509	0.611	0.695	0.981
Whitemtn	1164	0.695	0.648	0.095

[1] The appendix is available with entire article on microfiche. Order from American Geophysical Union, 1909 K Street, N.W., Washington, D. C. 20006. Document W78-003; $1.00. Payment must accompany order.

TABLE 5. Results of the χ^2 Test for Geophysical Time Series

Data Set	Degrees of Freedom	$-2 \ln \sum p_i$
River flows	12	11.78
Miscellaneous	10	9.46
Tree rings	24	16.08
Total	46	37.32

series for either a one-sided or a two-sided test. Therefore on the basis of the given information, Arma models do preserve K or the RAR when considering all the time series. Furthermore, when the set of annual river flows, miscellaneous data, and tree ring indicies are inspected individually, it can be seen from Table 5 that Arma models preserve the historical Hurst statistics for all three cases.

In Table 4 the average of the observed K is calculated to be 0.693 with a standard deviation of 0.076. The $E(K)$ from the simulations has an average of 0.698 with a standard deviation of 0.068. The average of the observed K is therefore slightly less than that for the simulated case, but this difference is not statistically different.

If the results of the RAR had been given rather than K, only columns 3 and 4 of Table 4 would be different, due to the transformation in (9). The p_i values and the results of the χ^2 test in Table 6 would be identical. Therefore preservation of either K or $\bar{R}_N^*$ infers retention of the other statistic by Arma models.

ESTIMATES OF THE HURST COEFFICIENT

From empirical studies of approximately 690 geophysical time series, *Hurst* [1951, 1956] found the RAR to vary as

$$\bar{R}_N^* \propto N^h \quad (12)$$

where h is a constant often referred to as the generalized Hurst coefficient. The above equation can be written in the general form

$$\bar{R}_N^* = aN^h \quad (13)$$

where a is a coefficient. Hurst assumed the coefficient a to have a value of $(1/2)^h$ and then estimated h by K in (9) and (8).

Recently, *Siddiqui* [1976] has employed the functional central limit theorem and the theory of Brownian motion to derive many statistical formulae that may be of interest to hydrologists. Of particular importance is the asymptotic result for calculating $E(\bar{R}_N^*)$ for Arma processes. This formula is given as

$$E(\bar{R}_N^*) \simeq a'N^{1/2} \quad (14)$$

where

$$a' = 1.2533\gamma_0^{-1/2}\left(1 - \sum_{i=1}^{q} \theta_i\right) \Big/ \left(1 - \sum_{i=1}^{p} \phi_i\right)$$

and γ_0 is the theoretical autocovariance function at lag 0 that is evaluated by using the algorithm of *McLeod* [1975] with $\sigma_a^2 = 1$. If the random variables are IID, a special case of (14) that was previously derived by *Feller* [1951] is

$$E(R_N^*) \simeq 1.2533N^{1/2} \quad (15)$$

By comparing (14) and (13) a possible alternative method of evaluating h may be to employ the equation

$$\bar{R}_N^* = a'N^{SH} \quad (16)$$

where SH is Siddiqui's estimate of the generalized Hurst coefficient h. When logarithms are taken of (15), Siddiqui's estimate for h is [*Siddiqui*, 1976]

$$SH = (\log \bar{R}_N^* - \log a')(\log N)^{-1} \quad (17)$$

It should be noted that due to the way *Hurst* [1951, 1956] and *Siddiqui* [1976] calculate the coefficient a in (13), the Hurst coefficient K and the Siddiqui coefficient SH are in fact two different statistics. Nevertheless, as was suggested by *Siddiqui* [1976], it may be of interest to determine whether h exhibits the Hurst phenomenon if the estimate SH is employed. Accordingly, for the 23 geophysical time series given in Tables 1 and 2 the K and SH statistics are compared.

Table 3 lists the Arma models fit to the 23 time series. If a Box-Cox transformation is included in a model, then K and SH are calculated for the transformed series to which the model is fit. This is because the formula for calculating SH in (17) does not have the capability of incorporating a Box-Cox transformation in order to get an estimate of SH for the untransformed data. Table 6 displays the values of K and SH that are calculated for each time series by using (8) and (17), respectively. Notice that the entries for K in Table 6 differ from the K values in Table 4 wherever the data used in Table 6 have been transformed by a Box-Cox transformation.

An examination of Table 6 reveals that in all cases except three the value of SH is less than K for the corresponding time series. The K statistic has an arithmetic mean of 0.701 with a standard deviation of 0.084. However, the mean of the SH statistic is 0.660 and possesses a standard deviation of 0.131. The mean value of SH is therefore well within 2 standard deviations of 0.500.

Another technique to estimate h can be found by comparing (15) and (13). Accordingly, *Gomide* [1975] suggests the following equation to evaluate h:

$$\bar{R}_N^* = 1.2533N^{YH} \quad (18)$$

TABLE 6. Estimates of the Hurst Coefficient

Code Name	K	SH	YH
Mstouis	0.648	0.591	0.500
Neumunas	0.677	0.591	0.535
Danube	0.633	0.495	0.495
Rhine	0.614	0.484	0.484
Ogden	0.894	0.929	0.709
Gota	0.689	0.636	0.549
Espanola	0.928	0.927	0.779
Temp	0.694	0.640	0.567
Precip	0.615	0.473	0.473
Sunyr	0.723	0.949	0.580
Minimum	0.817	0.746	0.699
Snake	0.687	0.663	0.579
Exshaw	0.637	0.580	0.530
Naramata	0.595	0.543	0.492
Dell	0.687	0.667	0.579
Lakeview	0.703	0.706	0.590
Ninemile	0.727	0.642	0.617
Eaglecol	0.761	0.701	0.650
Navajo	0.653	0.584	0.550
Bryce	0.734	0.727	0.620
Tioga	0.704	0.691	0.594
Bigcone	0.611	0.595	0.507
Whitment	0.695	0.623	0.595

where YH is Gomide's estimate of the generalized Hurst coefficient h. By taking logarithms of (18), Gomide's estimate of h is

$$YH = (\log \bar{R}_N^* - \log 1.2533)(\log N)^{-1} \quad (19)$$

When (19) is utilized to estimate the Hurst coefficient, *Gomide* [1975] obtains an average value for YH of 0.57 for the 690 series considered by *Hurst* [1951, 1956]. On the other hand, *Hurst* [1951, 1956] calculated K to have an average of 0.73 for the 690 series. Therefore lower values are obtained for the Hurst coefficient h if YH is employed rather than K.

Table 6 lists the values of YH for the same 23 geophysical time series that are considered for SH. Therefore if a Box-Cox transformation is included with an Arma model in Table 3, then YH is determined for the transformed series to which the model is fit. Obviously, because YH, as calculated in (19), is not a function of the Arma model parameters, it is not in general necessary to consider the transformed series. However, the aforementioned procedure is adopted so that appropriate comparisons can be formulated for the three estimates given in Table 6.

A perusal of Table 6 shows that for each time series the value of YH is consistently less than the magnitude of K. For the series to which white noise models are fit in Table 3 (i.e., Danube, Rhine and Precip) the values of YH and SH in Table 6 are equivalent. However, for all the other data sets the magnitudes of YH are less than SH. The mean of the 23 YH values is 0.577 with a standard deviation of 0.078. The YH statistic is within 1 standard deviation of 0.500. Therefore it can perhaps be argued that for the data considered, the Hurst phenomenon is not significant for the YH statistic. A similar argument can be made for the SH estimate of h.

CONCLUSIONS

Arma models preserve the observed RAR and K when fit to a variety of geophysical time series. Because important stochastic characteristics of hydrologic time series are retained by Box-Jenkins models, this should give engineers confidence in water resource projects that are designed with the aid of simulation techniques. In particular, the RAR statistic is directly related to storage problems, and this makes Arma models desirable for reservoir design, operation, and evaluation.

Following the identification, estimation, and diagnostic check stages of model development it is a straightforward procedure to develop a model for a particular time series [*Hipel et al.*, 1977*a*; *McLeod et al.*, 1977]. If the phenomenon being modeled has been influenced significantly by external interventions, these effects can be incorporated into the model [*Hipel et al.*, 1975, 1977*b*]. By employing Monte Carlo techniques the ECDF of statistics such as the RAR or K can be developed to any desired accuracy. The ECDF are used in conjunction with a specified statistical test to check for the preservation of historical statistics. This testing procedure can be used to check for the retention of any observed statistics by Arma or by other types of stochastic models.

Besides considering Hurst's estimate K of the coefficient h it is possible to consider other types of estimates. For the 23 time series examined in this paper the Siddiqui coefficient SH [*Siddiqui*, 1976] and Gomide's statistic YH [*Gomide*, 1975] possess a mean value less than K. By examining the standard deviations of the SH and YH statistics the Hurst phenomenon is seen to be less pronounced for these estimates (especially for YH) than it is for K.

NOTATION

a coefficient used in (13).
a' coefficient used in (14).
a_t white noise time series value at time t.
AR(p) autoregressive process of order p.
B backward shift operator.
const additative constant for a Box-Cox transformation.
d order of the nonseasonal differencing operator.
$E(K)$ expected value of K.
F theoretical CDF for $\bar{R}_N^*$.
FGN fractional Gaussian noise.
$F_{\bar{N}}$ ECDF for F.
h generalized Hurst coefficient.
IID identically independently distributed variable.
$K(\epsilon)$ limit approached for Kolmogoroff's theorem.
K_i^{obs} K value calculated for the ith observed time series.
K Hurst's estimate of the generalized Hurst coefficient.
MLE maximum likelihood estimates.
MA(q) moving average process of order q.
N number of data in a time series.
n length of the w_t time series.
$\bar{N}$ Number of simulated series of length N.
NID normally independently distributed variable.
p order of the nonseasonal AR operator.
p_i probability that $K > K_i^{obs}$ for a particular model.
(p, d, q) orders of the operators of a nonseasonal Arima model.
(p, q) orders of the operators of a nonseasonal Arma model.
q order of the nonseasonal MA operator.
r any possible value of $\bar{R}_N^*$.
$\bar{R}_N^*$ rescaled adjusted range.
$\bar{R}_{Ni}^*$ value of the RAR for the ith simulated series of length N.
$\bar{R}_{N(i)}^*$ ith largest value of the RAR when the RAR for $\bar{N}$ simulated series of length N are put in ascending order.
SH Siddiqui's estimate of the generalized Hurst coefficient h.
t discrete time.
w_t stationary series formed by differencing the $z_t^{(\lambda)}$ series.
YH Gomide estimate of the generalized Hurst coefficient h.
z_t discrete time series value at time t.
$z_t^{(\lambda)}$ transformation of the z_t series.
γ_k autocovariance function at lag k.
ϵ increment.
$\boldsymbol{\theta}$ set of unkonwn MA parameters.
$\theta(B)$ nonseasonal MA operator of order q.
θ_i ith nonseasonal MA parameter.
λ exponent for a Box-Cox transformation.
μ mean level of the w_t series.
σ_a^2 variance of a_t.
$\boldsymbol{\phi}$ set of unknown AR parameters.
ϕ_i ith nonseasonal AR parameter.
$\phi(B)$ nonseasonal AR operator of order p.
χ_ν^2 chi-squared random variable with ν degrees of freedom.
∇^d nonseasonal differencing operator of order d.

Acknowledgments. The authors appreciate the advice and encouragement given by T. E. Unny and W. C. Lennox of the University of Waterloo and also N. C. Matalas of the U.S. Geological Survey. V. C. LaMarche of the University of Arizona generously provided the authors with a copy of the reference by Stokes et al. (1973). Bruce Hinton's assistance with part of the computational work is appreciated. The mathematics faculty at the University of Waterloo kindly provided free use of the Mathematics Faculty Computer Facility for computational purposes. A portion of the research was funded by the National Research Council of Canada. Finally, the authors wish to thank Sheila Hipel for doing a professional job of typing and editing this paper.

REFERENCES

Anis, A. A., and E. H. Lloyd, The expected value of the adjusted rescaled Hurst range of independent normal summands, *Biometrika, 63,* 111-116, 1976.

Box, G. E. P., and D. R. Cox, An analysis of transformations, *J. Roy. Statist. Soc., Ser. B, 26,* 211-252, 1964.

Box, G. E. P., and G. M. Jenkins, *Time Series Analysis: Forecasting and Control,* Holden-Day, San Francisco, Calif., 1970.

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

De Geer, G., *Geochronologia Suecica Principles,* Almqvist and Wiksells, Stockholm, 1940.

Feller, W., The asymptotic distribution of the range of sums of independent random variables, *Ann. Math. Statist., 22,* 427-432, 1951.

Fisher, R. A., *Statistical Methods for Research Workers,* Oliver and Boyd, Edinburgh, England, 1970.

Gnedenko, B. V., *The Theory of Probability,* Chelsea, N.Y., 1968.

Gomide, F. L. S., Range and deficit analysis using Markov chains, *Hydrol. Pap. 79,* Colo. State Univ., Fort Collins, 1975.

Hipel, K. W., Contemporary Box-Jenkins modelling in hydrology, Ph.D. thesis, Univ. of Waterloo, Waterloo, Ont., Canada, 1975.

Hipel, K. W., and A. I. McLeod, *Applied Box-Jenkins Modeling for Water Resources Engine rs,* Elsevier, New York, in press, 1977.

Hipel, K. W., and A. I. McLeod, Preservation of the rescaled adjusted range, 3, Fractional Gaussian noise algorithms, *Water Resour. Res., 14,* this issue, 1978.

Hipel, K. W., W. C. Lennox, T. E. Unny, and A. I. McLeod, Intervention analysis in water resources, *Water Resour. Res., 11*(6), 855-861, 1975.

Hipel, K. W., A. I. McLeod, and W. C. Lennox, Advances in Box-Jenkins modeling, 1, Model construction, *Water Resour. Res., 13*(3), 567-576, 1977*a*.

Hipel, K. W., A. I. McLeod, and E. A. McBean, Stochastic modelling of the effects of reservoir operation, *J. Hydrol., 32,* 97-113, 1977*b*.

Hurst, H. E., Long-term storage of reservoirs, *Trans. Amer. Soc. Civil Eng., 116,* 770-808, 1951.

Hurst, H. E., Methods of using long-term storage in reservoirs, *Proc. Inst. Civil Eng., 1,* 519-543, 1956.

Kendall, M. G., and A. Stuart, *The Advanced Theory of Statistics,* vol. 1, *Distribution Theory,* Hafner, New York, 1963.

Knuth, D. F., *The Art of Programming,* vol. 3, Addison-Wesley, Reading, Mass., 1973.

Mandelbrot, B. B., and J. R. Wallis, Noah, Joseph, and operational hydrology, *Water Resour. Res., 4*(5), 909-918, 1968.

Manley, G., The mean temperatures of central England (1698-1952), *Quart. J. Roy. Meteorol. Soc., 79,* 242-261, 1953.

McLeod, A. I., Contributions to applied time series, masters thesis, Univ. of Waterloo, Waterloo, Ont., Canada, April 1974.

McLeod, A. I., Derivation of the theoretical autocovariance function of autoregressive-moving average time series, *J. Roy. Statist. Soc., Ser. C, 24*(2), 255-256, 1975.

McLeod, A. I., Improved Box-Jenkins estimators, *Biometrika, 64*(3), 531-534, 1977.

McLeod, A. I., On the distribution of residual autocorrelations in Box-Jenkins models, *J. Roy. Stat. Soc., Ser. B,* in press, 1978.

McLeod, A. I., and K. W. Hipel, Preservation of the rescaled adjusted range, 1, A reassessment of the Hurst phenomenon, *Water Resour. Res., 14,* this issue, 1978*a*.

McLeod, A. I., and K. W. Hipel, Simulation procedures for Box-Jenkins models, *Water Resour. Res., 14,* in press, 1978*b*.

McLeod, A. I., K. W. Hipel, and W. C. Lennox, Advances in Box-Jenkins modeling, 2, Applications, *Water Resour. Res., 13*(3), 577-586, 1977.

Schulman, E., *Dendroclimatic Change in Semi-Arid America,* University of Arizona Press, Tucson, Ariz., 1956.

Siddiqui, M. M., The asymptotic distribution of the range and other functions of partial sums of stationary processes, *Water Resour. Res., 12*(6), 1271-1276, 1976.

Stokes, M. A., L. G. Drew, and C. W. Stockton, Tree ring chronologies of western America, *Chron. Ser. 1,* Lab. of Tree-Ring Res., Univ. of Ariz., Tucson, 1973.

Toussoun, O., *Memoires sur l'Histoire de Nil, Memoires de l'Institut d'Egypte,* Imprimerie de L'institut Francais d'Archéologie Orientale, Cairo, 8-10, 1925.

Waldmeier, M., *The Sunspot Activity in the Years 1610-1960,* Schulthas and Co., Zurich, Switzerland, 1961.

Yevjevich, V. M., Fluctuation of wet and dry years, 2, Research data assembly and mathematical models, *Hydrol. Pap. 1,* Colo. State Univ., Fort Collins, 1963.

(Received January 20, 1976;
revised May 15, 1977;
accepted June 6, 1977.)