Effect of persistence on trend detection via regression

Nicholas C. Matalas
Vienna, Virginia, USA

A. Sankarasubramanian
International Research Institute for Climate Prediction, Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York, USA

Received 28 April 2003; revised 13 August 2003; accepted 28 August 2003; published 5 December 2003.

[1] Trends in hydrologic sequences may be assessed in various ways. The coefficient of regression of flow on time may be used, particularly if the sequences are very long. Under the assumption of stationarity the variance of the regression coefficient is expressed as a function of sequence length and the autocorrelation coefficients of relevant order. Thus the variance inflation factor for assessing the statistical significance of estimated regression coefficients may be readily determined for any given stationary process. The variance inflation factor is determined for four stationary processes: independent, Markov, autoregressive-moving average of order (1, 1), and fractional Gaussian noise. The effectiveness of prewhitening observed sequences with a Markov process is nearly the same whether the first order autocorrelation coefficient is known per se or through estimation. *INDEX TERMS:* 1869 Hydrology: Stochastic processes; 1860 Hydrology: Runoff and streamflow; 1833 Hydrology: Hydroclimatology; *KEYWORDS:* persistence, stationarity, trends

Citation: Matalas, N. C., and A. Sankarasubramanian, Effect of persistence on trend detection via regression, *Water Resour. Res.*, *39*(12), 1342, doi:10.1029/2003WR002292, 2003.

1. Introduction

[2] It is generally recognized that persistence affects the testing of statistical hypotheses. In the case of positive persistence marked by the tendency for high values to follow high values and for low values to follow low values, persistence compromises tests of significance: tests are likely to yield a lesser frequency of acceptance of a null hypothesis when in fact the null hypothesis is true. To fully account for persistence, the coefficients measuring persistence would need to be known as well as the process yielding the realization from which the statistic to be tested for significance is extracted.

[3] Several investigators have explored the interaction between deterministic trend and autoregressive process [*Wilkes*, 1997; *Yue and Pilon*, 2003; *Yue et al.*, 2002]. *Wilkes* [1997] dealt with the effect of persistence, as well as spatial correlation, on testing the null hypothesis that the difference in mean values is a specific value. The most common case is testing the hypothesis that the difference in mean values is equal to zero. Wilkes obtained approximate variance inflation factors for the test statistics in the case where observed sequences are realizations of specific stochastic processes, in particular, a Markov process, an ARMA(1, 1) process and an autoregressive process of order 2, i.e., an ARMA(2, 0) process. For these processes, persistence has the effect of rejecting the null hypothesis more frequently than would be the case if persistence was totally absent, i.e., if the observed sequences were realizations of independent processes.

0043-1397/03/2003WR002292$09.00

[4] *Zheng et al.* [1997] explored climate trends for the New Zealand region and found that a linear trend was adequate to explain annual time series of air temperature and sea surface temperature. They showed that ignoring serial correlation exhibited by sequences would result in smaller confidence intervals for trends measured as the regression coefficient of air temperature on time. *Von Storch* [1999] noted that prewhitening observed sequences lessens the adverse effects of persistence on tests of null hypotheses. He illustrated the effectiveness of prewhitening under the assumption that observed sequences are realizations of Markov processes. Storch notes that prewhitening with an assumed Markov process may improve the performance of test statistics, but the tests may yet reject the null hypothesis with a frequency that is unsatisfactorily high. *Douglas et al.* [2000] took into account a regional partition of the sequences in their assessment of trends, based on the Mann-Kendall test, in sequences of low flows and in sequences of high flows. They found that prewhitening, based on Markov processes with sample estimates of first order autocorrelation coefficients, yielded fewer regions being interpreted as having statistically significant trends. The interaction between trend and persistence was examined by *Matalas and Olsen* [2001] in terms of changes in the frequency of sequences showing significant trends following detrending and removal of Markovian persistence. *Yeu and Wang* [2002] suggest that prewhitening is not effective in eliminating the effect of serial correlation on trend detection through the Mann-Kendall test when sequences do in fact have trending observations. The effectiveness was assessed through simulation via the estimation of the first order autocorrelation coefficient in observed sequences generated by a Markov process coupled with trend.

[5] In the following discussions the effect of persistence on the variance of the regression coefficient is assessed - the regression coefficient is taken as the measure of trend in an observed sequence. An observed hydrologic sequence is assumed to be a realization of one of four stochastic processes, an independent process, a Markov process, an autoregressive-moving average process of order (1,1), or a process of fractional Gaussian noise. For each of these processes, the variance inflation factor of the regression coefficient is determined. The effectiveness of prewhitening a generated sequence under the assumption that the sequence is a realization of a Markov process when in "fact" the process is Markov or one of the other three processes is assessed via Monte Carlo experiments. The effectiveness of prewhitening is assessed using the population first order autocorrelation coefficient as well as using the sample estimate of first order autocorrelation coefficient.

2. Hydrologic Process

[6] Assume that a hydrologic sequence, $\{x_t: t = 1,2, \ldots, n\}$, e.g., a streamflow sequence, is a realization of a stationary process, P, and also assume that $x_t \sim N(\mu, \sigma^2)\ \forall t$. Persistence may be a reflection of either short-term (finite) memory exhibited, e.g., by an autoregressive-moving average process, $ARMA(p, q)$ of nonnegative orders $p < \infty$ and $q < \infty$, or a reflection of long-term (infinite) memory exhibited, e.g., by a process of fractional Gaussian noise (FGN). In the following discussions, four processes are considered, namely, $ARMA(0, 0)$, $ARMA(1, 0)$, $ARMA(1, 1)$, and FGN, where $ARMA(0, 0)$ is an independent process and is designated as I, $ARMA(1, 0)$ is a Markov process and is designated as M, and $ARMA(1, 1)$ is an autoregressive-moving average process of order (1, 1) and is designated as AM. FGN is a fractional Gaussian noise process, an $ARMA(0, \infty)$ [see *Mandelbrot and Taqqu*, 1979]. Each of these processes has been used to describe hydrologic processes, particularly streamflow.

[7] The correlograms for $P \equiv M$, $P \equiv AM$ and $P \equiv FGN$ are alike in the sense that for each of the processes, the autocorrelation of lag k diminishes monotonically as k increases. The correlogram for an independent process ($P \equiv I$) [*Kendall*, 1975] is defined as

$$\rho_k = \begin{cases} 1; & k = 0 \\ 0; & |k| > 0 \end{cases} \tag{1}$$

The correlogram for ARMA(1, 0) ($P \equiv M$) process [*Kendall*, 1975] is defined as

$$\rho_k = \begin{cases} 1; & k = 0 \\ \rho^{|k|}; & |k| > 0 \end{cases} \tag{2}$$

where $\rho \equiv \rho|l|$. If $\rho = 0$, $P \equiv M$ reduces to $P \equiv I$.

[8] The correlogram for ARMA(1, 1) process ($P \equiv AM$) [*Wilkes*, 1997] is defined as

$$\rho_k = \begin{cases} 1; & k = 0 \\ \rho_1; & |k| = 1 \\ \rho_2; & |k| = 2 \\ \phi^{k-2}\rho_2; & |k| \geq 3 \end{cases} \tag{3}$$

where $\phi = \rho_2/\rho_1$. If $\rho_2 = \rho_1^2$, $P \equiv AM$ reduces to $P \equiv M$. If $\rho_1 \to 0$ and $\rho_2 \to 0$, such that $\phi \to 1$, $P \equiv AM$ reduces to $P \equiv I$.

[9] The correlogram for the fractional Gaussian noise ($P \equiv FGN$) process [*Mandelbrot and Taqqu*, 1979] is defined as

$$\rho_k = \begin{cases} 1; & k = 0 \\ (1/2)\left[|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right]; & |k| > 0 \end{cases} \tag{4}$$

where H is referred to as the Hurst coefficient. Given ρ_1,

$$H = (1/2)\left[1 + \frac{\ln(1+\rho_1)}{\ln(2)}\right] \tag{5}$$

If $H = 1/2$, the FGN process reduces to an independent process. An $ARMA(p, q)$ process is characterized by $H = 1/2 \forall\ p$ and $q < \infty$.

[10] In the following discussions, unless otherwise noted, it is assumed that the value of ρ_1 is the same for each of the three processes, $P \equiv M$, $P \equiv AM$ and $P \equiv FGN$, and ρ_2 is the same for $P \equiv AM$ and $P \equiv FGN$. For a specific value of ρ_1, H is determined from equation (5), whereupon ρ_2 is determined from equation (4). From equations (2)–(5), it is readily seen that $\rho_2\ (M) < \rho_2\ (AM) = \rho_2\ (FGN)$, and that $\rho_k\ (M) < \rho_k\ (AM) < \rho_k\ (FGN)\ \forall k \geq 3$.

[11] The cumulative sums of the autocorrelation coefficients for the processes are given as

$$S(m) = \sum_{k=0}^{m} \rho_k = \begin{cases} 1; & P \equiv I \\ 1 + \dfrac{\rho_1(1-\rho_1^m)}{(1-\rho_1)}; & P \equiv M \\ 1 + \dfrac{\rho_1^2}{(\rho_1-\rho_2)}\left[1 - \left(\dfrac{\rho_2}{\rho_1}\right)^m\right]; & P \equiv AM \\ \textit{Determined numerically}; & P \equiv FGN \end{cases} \tag{6}$$

where for $m \geq 3$, $S(m|I) < S(m|M) < S(m|AM) < S(m|FGN)$.

[12] As $m \to \infty$, $S(m)$ tends to a finite value for each of the processes with the exception of $P \equiv FGN$:

$$\lim_{m\to\infty} S(m) = S = \begin{cases} 1; & P \equiv I \\ 1 + \dfrac{\rho_1}{(1-\rho_1)}; & P \equiv M \\ 1 + \dfrac{\rho_1^2}{(\rho_1-\rho_2)}; & P \equiv AM \\ \infty; & P \equiv FGN \end{cases} \tag{7}$$

3. Measure of Trend Via Regression

[13] If a hydrologic process is nonstationary in the mean, then x_t may be expressed as

$$x_t = \beta t + (1 - R^2)^{1/2}\delta_t \tag{8}$$

where β denotes the coefficient of regression of x_t on t, and R, the coefficient of correlation between x_t and t and δ_t denotes the deviation about the trend line βt. Both β and R are measures of linear trend: $\beta = R/V[t]$, where $V[t]$ denotes the variance of t. Under the assumption that a hydrologic process is at least a second order stationary process, then $\beta = 0$, whereby $R = 0$, so that $x_t = \delta_t$. Given the observed sequence $\{x_t: t = 1, 2, \ldots, n\}$, the regression of x_t on t yields

$$\hat{x}_t = \bar{x} + b(t - \bar{t}) \tag{9}$$

where

$$\bar{x} = \sum_{t=1}^{n} x_t/n \tag{10}$$

$$\bar{t} = \sum_{t=1}^{n} t/n = (n+1)/2 \tag{11}$$

$$b = \left[\left(\sum_{t=1}^{n} t x_t/n\right) - \bar{x}\bar{t}\right] \Big/ J \tag{12}$$

where

$$J \equiv V[t] = \left[\left(\sum_{t=1}^{n} t^2/n\right) - \bar{t}^2\right] = (n^2-1)/12 \tag{13}$$

Assume $\mu_x = E[x_t] = 0\ \forall t$, whereby $\sigma_x^2 = E[\{x_t^2\}]\ \forall t$. Given that the null hypothesis, H_0: $\beta = 0$, is true, and that the hydrologic process is at least second order stationary, then $E[tx_t] = tE[x_t] = 0\ \forall t$. It follows that $E[b] = 0$ and therefore b is an unbiased estimator of $\beta = 0$. The variance of b is conditioned on the generating process of $\{x_t\colon t = 1, 2, \ldots, n\}$ and may be expressed as

$$V[b|P] = E[b^2|P] \tag{14}$$

where P denotes the generating process, M, AM or FGN. From equation (12),

$$J^2 b^2 = \left[\sum_{t=1}^{n} t^2 x_t^2 + 2\sum_{t=1}^{n-1}\sum_{\tau=t+1}^{n} t\tau x_t x_\tau\right] \Big/ n^2 - 2\left[\sum_{t=1}^{n} x_t \sum_{t=1}^{n} t \sum_{t=1}^{n} t x_t\right] \Big/ n^3 + \left\{\left[\sum_{t=1}^{n} x_t^2 + 2\sum_{t=1}^{n-1}\sum_{\tau=t+1}^{n} x_t x_\tau\right]\left[\sum_{t=1}^{n} t^2 + 2\sum_{t=1}^{n-1}\sum_{\tau=t+1}^{n} t\tau\right]\right\} \Big/ n^4 \tag{15}$$

By taking the expectation of both sides of equation (15) and then dividing through by J^2, the variance of b may be expressed as

$$V[b|P] = \sigma_x^2\{[A(P) + B(P)]n^{-2} - C(P)n^{-3} + D(P)n^{-4}\} \Big/ J^2 = \sigma_x^2 K(P) \tag{16}$$

where J is defined by equation (13) and where the terms A(P), B(P), C(P) and D(P) are defined as follows

$$A(P) = E\left[\sum_{t=1}^{n} t^2 x_t^2\right] = \sum_{t=1}^{n} t^2 E[x_t^2] = \sum_{t=1}^{n} t^2 = n(n+1)(2n+1)/6 \tag{17}$$

$$\begin{aligned} B(P) &= 2E\left\{\sum_{t=1}^{n-1}\sum_{\tau=t+1}^{n} [t\tau x_t x_\tau]\right\} = 2\sum_{t=1}^{n-1}\sum_{\tau=t+1}^{n} [t\tau\rho_{\tau-t}] \\ &= 2\left\{\sum_{\tau=2}^{n} \tau\rho_{\tau-1} + 2\sum_{\tau=3}^{n} \tau\rho_{\tau-2} + 3\sum_{\tau=4}^{n} \tau\rho_{\tau-3} + \cdots + (n-1)\sum_{\tau=n}^{n} \tau\rho_{\tau-(n-1)}\right\} \\ &= 2\left\{\sum_{k=1}^{n-1} (k/\rho_{k-1})\left[\sum_{i=1}^{n} i\rho_{i-1} - \sum_{j=1}^{k} j\rho_{j-1}\right]\right\} \\ &= \frac{1}{3}\sum_{t=1}^{n} \rho_t (n-t)(n-t+1)(2n+t+1) \end{aligned} \tag{18}$$

$$\begin{aligned} C(P) &= 2E\left\{\sum_{t=1}^{n} x_t \sum_{t=1}^{n} t \sum_{t=1}^{n} t x_t\right\} = 2\sum_{t=1}^{n} tE\left\{\sum_{t=1}^{n} x_t \sum_{t=1}^{n} t x_t\right\} \\ &= n(n+1)\left\{\begin{array}{l}\left[\sum_{\tau=1}^{n} \rho_{\tau-1} + 2\sum_{\tau=2}^{n} \rho_{\tau-2} + \cdots + n\sum_{\tau=n}^{n} \rho_{\tau-n}\right] \\ + \left[2\sum_{t=1}^{1} \rho_{2-t} + 3\sum_{t=1}^{2} \rho_{3-t} + \cdots + n\sum_{t=1}^{n-1} \rho_{n-t}\right]\end{array}\right\} \\ &= n(n+1)\left\{\left[\sum_{j=1}^{n} j\sum_{\tau=j}^{n} \rho_{\tau-j}\right] + \left[\sum_{j=2}^{n} j\sum_{t=1}^{j-1} \rho_{j-t}\right]\right\} \\ &= \frac{n(n+1)^2}{2}\left[n + 2\sum_{t=1}^{n-1} \rho_t (n-t)\right] \end{aligned} \tag{19}$$

$$\begin{aligned} D(P) &= E\left\{\left[\sum_{t=1}^{n} x_t^2 + 2\sum_{t=1}^{n-1}\sum_{\tau=t+1}^{n} x_t x_\tau\right]\left[\sum_{t=1}^{n} t^2 + 2\sum_{t=1}^{n-1}\sum_{\tau=t+1}^{n} t\tau\right]\right\} \\ &= \frac{n^2(n+1)^2}{4}\left[n + 2E\sum_{t=1}^{n-1}\sum_{\tau=t+1}^{n} x_t x_\tau\right] \\ &= \frac{n^2(n+1)^2}{4}\left\{n + 2\left[\sum_{\tau=2}^{n} \rho_{\tau-1} + \sum_{\tau=3}^{n} \rho_{\tau-2} + \cdots + \sum_{\tau=n}^{n} \rho_{\tau-(n-1)}\right]\right\} \\ &= \frac{n^2(n+1)^2}{4}\left[n + 2\sum_{t=1}^{n-1}\sum_{\tau=t+1}^{n} \rho_{\tau-t}\right] \\ &= \frac{n^2(n+1)^2}{4}\left[n + 2\sum_{t=1}^{n-1} \rho_t (n-t)\right] = \frac{n}{2}C(P) \end{aligned} \tag{20}$$

The term $A(P)$ is a function only of the sequence length, n, and therefore holds for any stationary stochastic process. Given the correlogram of an arbitrary stationary stochastic process, the variance of the regression coefficient, measure of trend for a realization of the process, may be readily determined via equations (17)–(20). In the case of $P \equiv M$, only the value of ρ_1 need be given to determine the variance of the regression coefficient:

$$B(P \equiv M) = 2\left\{\begin{array}{l}\dfrac{[1+n(1-\rho)][\rho^3(1-\rho^{n-1}) - (n-1)(1-\rho)\rho^2]}{(1-\rho)^4} \\ + \dfrac{n(n-1)[3\rho + (2n-1)\rho(1-\rho)]}{6(1-\rho)^2}\end{array}\right\} \tag{21}$$

Table 1. Variance Inflation Factor U(b|P) for Regression Coefficient Conditioned on P Given n and ρ_1

	ρ_1							
n	0	0.01	0.1	0.3	0.5	0.7	0.9	0.99
				$P \equiv M$				
5	1.000	1.008	1.077	1.194	1.200	0.996	0.450	0.051
25	1.000	1.018	1.193	1.711	2.526	3.893	5.051	1.135
50	1.000	1.019	1.207	1.784	2.761	4.745	9.650	4.099
75	1.000	1.019	1.212	1.808	2.840	5.048	12.257	8.356
100	1.000	1.020	1.215	1.820	2.880	5.201	13.794	13.496
300	1.000	1.020	1.220	1.845	2.960	5.511	17.207	61.256
500	1.000	1.020	1.221	1.850	2.976	5.573	17.922	97.492
1000	1.000	1.020	1.221	1.853	2.988	5.620	18.460	141.949
∞	1.000	1.020	1.222	1.857	3.000	5.667	19.000	199.000
				$P \equiv AM$				
5	1.000	1.006	1.048	1.079	0.991	0.747	0.306	0.033
25	1.000	1.026	1.284	2.026	3.081	4.373	4.027	0.672
50	1.000	1.029	1.322	2.221	3.706	6.390	9.783	2.523
75	1.000	1.030	1.335	2.288	3.929	7.229	14.416	5.354
100	1.000	1.031	1.342	2.321	4.042	7.670	17.797	8.994
300	1.000	1.032	1.355	2.389	4.270	8.579	27.084	53.006
500	1.000	1.032	1.357	2.402	4.316	8.764	29.239	101.077
1000	1.000	1.032	1.359	2.412	4.350	8.902	30.883	187.196
				$P \equiv FGN$				
5	1.000	1.004	1.027	1.010	0.884	0.634	0.248	0.027
25	1.000	1.027	1.284	1.866	2.282	2.195	1.113	0.133
50	1.000	1.038	1.412	2.426	3.424	3.732	2.116	0.264
75	1.000	1.044	1.493	2.828	4.341	5.091	3.081	0.395
100	1.000	1.048	1.553	3.154	5.136	6.345	4.022	0.526
300	1.000	1.065	1.807	4.780	9.767	14.713	11.123	1.564
500	1.000	1.072	1.938	5.800	13.168	21.753	17.850	2.597
1000	1.000	1.083	2.132	7.540	19.752	36.981	33.916	5.169

$$C(P \equiv M) = \frac{n(n+1)^2[n(1-\rho^2) - 2\rho(1-\rho^n)]}{2(1-\rho)^2} \tag{22}$$

$$D(P \equiv M) = \frac{[n^2(n+1)^2][n(1-\rho^2) - 2\rho(1-\rho^n)]}{4(1-\rho)^2} = \frac{n}{2}C(P \equiv M) \tag{23}$$

where $\rho \equiv \rho_1$. For $\rho = 0$, equations (21), (22), and (23) reduce to the expressions for $P \equiv I$.

4. Variance Inflation Factor for the Regression Coefficient

[14] The extent to which the variance of the regression coefficient is inflated by persistence may be measured by

$$U(b|P) = V[b|P]/V[b|I] = \sigma_x^2 K(P)/\sigma_x^2 K(P \equiv I) = \{[n(n^2-1)]/12\}K(P) \tag{24}$$

where $K(P \equiv I)$ is given by equations (16)–(20) with $\rho_t = 0$ $\forall t > 0$.

[15] If the inflation factor $U[b|P] > 1$, then the effect of persistence on detection of linear trend is to increase the variance of b, the coefficient of regression of x_t on t. In this case, if account is not taken of persistence and sequences are dealt with as though they are realizations of independent processes, then it is likely that the frequency of sequences apparently exhibiting trend will be greater than if the sequences were indeed realizations of independent processes. In such a case, the null hypothesis of stationarity would be rejected when in "fact" the null hypothesis is true. On the other hand, if $U[b|P] < 1$, then it is likely that the frequency of sequences apparently exhibiting trend will be less than if the sequences were indeed realizations of independent processes. In such a case, it is highly unlikely that the null hypothesis of stationary would be rejected.

[16] Values of $U(b|P)$ for specific values of $0.01 \leq \rho_1 \leq 0.99$ and $5 \leq n \leq 1{,}000$ are given in Table 1 for $P \equiv M$, $P \equiv AM$ and $P \equiv FGN$ with ρ_2 being computed using equations (4) and (5) conditioned on ρ_1 for P ≡ AM and P ≡ FGN. As a point of reference, the value $U(b|P \equiv I) = 1 \; \forall \; n$ is given in Table 1. The limiting values of $U(b|P \equiv M)$ as $n \to \infty$,

$$\lim_{n\to\infty} U(b|P \equiv M) = U^*(b|P \equiv M) = (1+\rho)/(1-\rho) \tag{25}$$

where $\rho \equiv \rho_1$, and are shown in Table 1. For a given value of ρ_1, the inflation factor increases monotonically with n, at least up to n = 1,000. For a given value of n, the inflation factor increases with ρ_1, reaching a peak and then decreasing. The peak value of the inflation factor is reached with higher values of ρ_1 as n becomes larger. Except for very high levels of persistence coupled with small sequence lengths, the effect of persistence is to inflate the variance of the regression coefficient. Unless the level of persistence is very high, $U(b|M) < U(b|AM) < U(b|FGN)$. The inflation factor is not sensitive to P for very low levels of persistence. However, as the level of persistence increases, P becomes a more important decision variable in evaluating trend via regression. In the case of streamflow, the level of persistence on an annual scale is low and the lengths of observed sequences are generally less than 100, whereby, the degree to which variance of the regression coefficient is inflated is not overly sensitive to the choice of P for describing the streamflow process. For surrogate hydrologic processes, e.g., variations in the widths of tree rings and mud varves, higher levels of persistence and much longer sequences yield greater sensitivity to the choice of P in determining the degree to which the variance of regression is inflated.

5. Effect of Prewhitening on Trend Detection

[17] Prewhitening is an operation seeking to transform a process P into a process I via an assumed process P'. For prewhitening to be fully effective, P and its parameter values would need to be known. If P is not known, then prewhitening with an arbitrary P' is not likely to yield a process P^* that is identically I. In the following discussions, the effect of prewhitening is assessed in the case where $P' \equiv M$, i.e., the prewhitening process is a first order Markov process characterized by parameters $\omega = \rho_1$ and σ_z.

5.1. Markov Parameter $\omega = \rho_1$ Known

[18] Prewhitening P by $P' \equiv M$ yields

$$y_t = x_t - \nu x_{t-1} \tag{26}$$

where ν denotes the presumed value of ω, and where x_t derives from the process P, whereby y_t derives from process P^*. From the assumption that $\mu_x = E[x_t] = 0 \; \forall t$, it follows

Table 2. Variance Inflation Factor Following Prewhitening P, U(b*|P*), With P′ = M Characterized by Parameter $\nu = \rho_1$

	ρ_1							
n	0	0.01	0.1	0.3	0.5	0.7	0.9	0.99
				$P \equiv AM$				
5	1.000	0.996	0.965	0.901	0.857	0.842	0.864	0.888
25	1.000	1.008	1.073	1.164	1.171	1.043	0.724	0.619
50	1.000	1.010	1.093	1.234	1.312	1.281	0.906	0.598
75	1.000	1.011	1.100	1.258	1.361	1.381	1.067	0.598
100	1.000	1.011	1.104	1.269	1.386	1.434	1.187	0.606
300	1.000	1.012	1.110	1.293	1.437	1.541	1.518	0.748
500	1.000	1.012	1.112	1.297	1.447	1.563	1.595	0.910
1000	1.000	1.012	1.113	1.301	1.454	1.579	1.654	1.204
				$P \equiv FGN$				
5	1.000	0.996	0.957	0.888	0.857	0.875	0.950	1.005
25	1.000	1.009	1.069	1.066	0.878	0.579	0.342	0.327
50	1.000	1.018	1.165	1.335	1.198	0.759	0.272	0.192
75	1.000	1.024	1.228	1.541	1.483	0.966	0.276	0.140
100	1.000	1.028	1.275	1.711	1.738	1.170	0.301	0.112
300	1.000	1.044	1.480	2.577	3.262	2.611	0.619	0.052
500	1.000	1.051	1.587	3.125	4.392	3.846	0.960	0.041
1000	1.000	1.062	1.745	4.060	6.584	6.528	1.795	0.042

that the process P^* is characterized by $\mu_y = E[y_t] = 0\ \forall t$ and $\sigma_y^2 = \sigma_x^2\ (1 - 2\nu\rho_1 + \nu^2)\ \forall t$.

[19] The correlogram of P^* is given by

$$R_k = \frac{E[y_{t+k}y_t]}{\sigma_y^2} = \begin{cases} 1;\ k = 0 \\ \dfrac{\rho_k(1+\nu^2) - \nu(\rho_{k+1} + \rho_{k-1})}{(1 - 2\nu\rho_1 + \nu^2)};\ k > 0 \end{cases} \quad (27)$$

where ρ_k denotes the kth order autocorrelation of P. Refer to equations (1)–(5). If $P \equiv M$ with parameter $\nu = \rho_1$, then $R_k = 0\ \forall k > 0$, whereby $P^* \equiv I$. If $P \equiv I$, then $R_1 = -\nu/(1 + \nu^2)$ and $R_k = 0\ \forall\ k > 1$.

[20] Assume that P is prewhitened by $P' \equiv M$ with parameter ν resulting in P^*. Let $u = t - 1$, where $t = 2, \ldots, n$, whereby $u = 1, \ldots, n - 1$. Given the observed sequence $\{y_u: u = 1, \ldots, n - 1\}$, a realization of P^*, the regression of y_u on u yields

$$\hat{y}_u = \bar{y} + b^*(u - \bar{u}) \quad (28)$$

where $\bar{y}$, u, $\bar{u}$ and b^* are given by equations (10)–(12) upon corresponding change in notation. Note equation (13) yields the term J^* upon corresponding change in notation and with n replaced by $n - 1$.

[21] The process $P' \equiv M$ is second order stationary. It is assumed that the processes P is at least second order stationary, whereby the process P^* is at least second order stationary. It is further assumed that the null hypothesis H_0: $\beta^* = 0$ is true, and therefore $E[uy_t] = uE[y_t] = 0\ \forall\ u$. Thus $E[b^*] = 0$, whereby b^* is an unbiased estimator of $\beta^* = 0$. The variance of b^* may be expressed as

$$V[b^*|P^*] = E\left[(b^*)^2|P^*\right] \quad (29)$$

where P^* derives from prewhitening P by $P' \equiv M$ with parameter ν. The variance of b^* is given by the right hand side of equation (16) upon replacing P by P^*, n by $n - 1$ and J by J^*. The terms $A(P^*)$, $B(P^*)$, $C(P^*)$ and $D(P^*)$ are given by equations (17)–(20) upon replacing P by P^*, n by $n - 1$, t by u and ρ_t by $R_{k\equiv u}$. See equation (29).

[22] Following prewhitening, the extent to which the variance of the regression coefficient is inflated is given by

$$U(b^*|P^*) = V[b^*|P^*]/V[[b^*|I]] \quad (30)$$

where $V[b^*|P^*]$ is given as described above, and where $V[b^*|I]$ is given by equation(16) upon replacing b, n and J by b^*, $n - 1$ and J^*, respectively. For selected values of n and ρ_1, values of $U(b^*|P^*)$ are given in Table 2 with ρ_2 being computed using equations (4) and (5) conditioned on ρ_1 for P ≡ AM and P ≡ FGN. From Tables 1 and 2 it is seen that prewhitening P with $P' \equiv M$ in the case where $\omega = \rho_1$, the characterizing parameter of M is known, greatly reduces the inflation factor. For the case $P \equiv M$, prewhitening with $P' \equiv M$ yields $P^* \equiv I$, whereby $U(b^*|I) = 1\ \forall\ n$ and ρ_1.

[23] The effectiveness of prewhitening is defined as the relative change in the variance of the estimate of the regression slope prior to and following prewhitening:

$$\begin{aligned} \eta &= [U(b^*|P) - U(b^*|P^*)]/U(b^*|P) \\ &= \{V[b^*|P] - V[b^*|P^*]\}/V[b^*|P] \\ &= 1 - V[b^*|P^*]/V[b^*|P] \end{aligned} \quad (31)$$

If $\eta \leq 0$, then prewhitening P by P' is counter-productive. For selected values of n and ρ_1, the effectiveness,η, of prewhitening $P \equiv AM$ and $P \equiv FGN$ with $P' \equiv M$ are given in Table 3 with ρ_2 being computed using equations (4) and (5) conditioned on ρ_1 for P ≡ AM and P ≡ FGN. From Table 3 it is seen that prewhitening with $P' \equiv M$ is as effective in reducing persistence characterizing $P \equiv AM$ as in reducing persistence characterizing $P \equiv FGN$. For a given value of ρ_1, the effectiveness of $P' \equiv M$ is very nearly the same for $P \equiv AM$ and $P \equiv FGN$ for any sequence length, n. Except for ρ_1 very large, say $\rho_1 > 0.8$, the rate of increase in the effectiveness prewhitening as ρ_1 increases does not depend upon the length of sequence, n.

5.2. Markov Parameter $\omega = \rho_1$ Not Known

[24] In practice ρ_1 is not known and therefore it must be estimated from an observed sequence. Let $\hat{\rho}_1$ denote the estimate of ρ_1 [see, e.g., *Kendall*, 1975]. The degrees of

Table 3. Effectiveness in Prewhitening P, η, With P′ ≡ M Characterized by Parameter $\nu = \rho_1$

	ρ_1							
n	0	0.01	0.1	0.3	0.5	0.7	0.9	0.99
				$P \equiv AM$				
5	0.000	0.010	0.079	0.165	0.135	0.127	−1.824	−25.909
25	0.000	0.018	0.164	0.425	0.620	0.761	0.820	0.079
50	0.000	0.018	0.173	0.444	0.646	0.800	0.907	0.763
75	0.000	0.018	0.176	0.450	0.654	0.809	0.926	0.888
100	0.000	0.019	0.177	0.453	0.657	0.813	0.933	0.933
300	0.000	0.019	0.181	0.459	0.663	0.820	0.944	0.986
500	0.000	0.019	0.181	0.460	0.665	0.822	0.945	0.991
1000	0.000	0.019	0.181	0.461	0.666	0.823	0.946	0.994
				$P \equiv FGN$				
5	0.000	0.008	0.068	0.121	0.031	−0.380	−2.833	−36.222
25	0.000	0.018	0.167	0.429	0.615	0.736	0.693	−1.462
50	0.000	0.019	0.175	0.450	0.650	0.797	0.872	0.273
75	0.000	0.020	0.178	0.455	0.658	0.810	0.910	0.646
100	0.000	0.019	0.179	0.457	0.662	0.816	0.925	0.787
300	0.000	0.020	0.181	0.461	0.666	0.823	0.944	0.967
500	0.000	0.019	0.181	0.461	0.666	0.823	0.946	0.984
1000	0.000	0.020	0.182	0.461	0.667	0.823	0.947	0.992

effectiveness in using $P' \equiv M$ to prewhiten sequences that are realizations of $P \equiv M$ and of $P \equiv AM$ are assessed via simulation. The generating process and the parameters of $P \equiv M$ are given by equation (32) through (34).

$$x_t = \omega x_{t-1} + z_t \tag{32}$$

where

$$\omega = \rho_1 \tag{33}$$

denotes the lag 1 autocorrelation coefficient, and where

$$\sigma_x^2 = \sigma_z^2/(1-\omega^2) \tag{34}$$

denotes the standard deviation of the white noise term, z_t, distributed as $N(0, \sigma_z)$. Without loss of generality, assume that $E[x_t] = \mu_x = 0$ and $\sigma_z = 1\ \forall t$, whereby the variance of x_t can be written as $\sigma_x^2 = 1/(1-\rho_1^2)\ \forall t$. For $P \equiv AM$

$$x_t = \phi x_{t-1} + z_t - \lambda z_{t-1} \tag{35}$$

where

$$\phi = \rho_2/\rho_1 \tag{36}$$

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4(\rho_1 - \phi)^2}}{2(\rho_1 - \phi)} \tag{37}$$

$$\sigma_x^2 = \left[\left(1 - 2\phi\lambda + \lambda^2\right)/\left(1-\phi^2\right)\right]\sigma_z^2 \tag{38}$$

where

$$b = 1 - 2\phi\rho_1 + \phi^2 \tag{39}$$

[25] The generation of a realization of $P \equiv AM$ is always constrained to satisfy causality, $(-1 < \phi < 1)$, and invertibility, $(-1 < \lambda < 1)$ [see, e.g., *Box and Jenkins*, 1976]. An ensemble of $N = 10{,}000$ sequences, where each sequences was of length n generated with $P \equiv M$ with a specified value of ρ_1. For each value of n = 10, 30, 60, 100, 500, the specified values of ρ_1 were 0, 0.1, 0.3, 0.6, 0.9, 0.99. In all, 30 ensembles were generated. For $P \equiv AM$, 25 ensembles, each consisting of $N = 10{,}000$ sequences, were generated: n = 10, 30, 60, 100, 500 and ρ_1 = 0.1, 0.3, 0.6, 0.9, 0.99. Ensembles for $\rho_1 = 0$ were not generated since the results, estimated degrees of effectiveness of prewhitening $P \equiv I$ with $P' \equiv M$ having parameter $\nu = \hat{\rho}_1$, for those ensembles would have yielded the same results, apart from sampling errors, as the ensemble for $P \equiv M$ with $\rho_1 = 0$, equivalently for $P \equiv I$. For $P \equiv AM$, the values of ρ_2 corresponding to the values of ρ_1 (0.1, 0.3, 0.6, 0.9, 0.99) were 0.05, 0.05, 0.3, 0.75 and 0.98 respectively.

[26] Prewhitening a realization of length n of the process P with $P' \equiv M$ characterized by $\nu = \hat{\rho}_1$ yields a realization of length $n - 1$ of P^*:

$$y_u = x_u - \hat{\rho}_1 x_{u-1} \tag{40}$$

where $u = t - 1$: $t = 2,3, \ldots, n$, whereby $u = 1, \ldots, n-1$. Given $\{x_t: t = 1, \ldots, n\}$, a realization of P, and $\{y_u: u = 1, \ldots, n-1\}$, a realization of P^*, the regression of x_t on t yields

$$\hat{x}_t = \bar{x} + b_x(t - \bar{t}) \tag{41}$$

and the regression of y_u on u yields

$$\hat{y}_u = \bar{y} + b_y(u - \bar{u}) \tag{42}$$

where b_x is determined directly from equation (12) and b_y is determined upon appropriate changes in notation from equations (10)–(13).

[27] For given values of n and ρ_1, the variances of b_x and b_y are given by

$$\hat{V}[b_x|P^*] = \frac{1}{N}\sum_{i=1}^{N} b_x^2 - \left(\frac{1}{N}\sum_{i=1}^{N} b_x\right)^2 \tag{43}$$

$$\hat{V}\left[b_y|P^*\right] = \frac{1}{N}\sum_{i=1}^{N} b_y^2 - \left(\frac{1}{N}\sum_{i=1}^{N} b_y\right)^2 \tag{44}$$

where $N = 10{,}000$ denotes the number of generated sequences forming the ensemble defined by n, ρ_1 and P^*, the process derived from prewhitening P with $P' \equiv M$, where for each realization, $P' \equiv M$ is characterized by $\nu = \hat{\rho}_1$. The estimated degree of effectiveness in prewhitening P with $P' \equiv M$ characterized by $\nu = \hat{\rho}_1$, is given by

$$\hat{\eta} = 1 - \hat{V}\left[b_y|P^*\right]/\hat{V}[b_x|P^*] \tag{45}$$

Prewhitening is counterproductive if $\eta \leq 0$. The effectiveness of prewhitening would be counterproductive if realizations of $P \equiv I$ were to be prewhitened with $P' \equiv M$ characterized by $\nu = \hat{\rho}_1$. For specific values of ρ_1 and n, the effectiveness of prewhitening $P \equiv M$ and $P \equiv AM$ with $P' \equiv M$ characterized by $\nu = \hat{\rho}_1$ are given in Table 4. From Table 4 it is seen that effectiveness of prewhitening with $P \equiv M$ or $P \equiv AM$ with $P' \equiv M$ characterized by $\nu = \hat{\rho}_1$ increases with ρ_1 given n and with n given ρ_1. For specific values of ρ_1 and n, the effectiveness of prewhitening with $P' \equiv M$ characterized by $\nu = \hat{\rho}_1$ is very nearly the same for $P \equiv M$ with $P' \equiv M$. Moreover, the effectiveness of prewhitening $P \equiv AM$ with $P' \equiv M$ is nearly the same whether ρ_1 is known or not. Further, Tables 3 and 4 show that prewhitening $P \equiv AM$ and $P \equiv FGN$ with $P' \equiv M$ is effective even if one does not apply the underlying model, since the effectiveness in prewhitening is greater than zero for almost all cases (except for samples of very small size with high correlation coefficient). It is an open question whether $P' \equiv M$ characterized by $\nu = \hat{\rho}_1$ would be equally as effective in prewhitening realizations of $P \equiv FGN$ as in prewhitening realizations of $P \equiv M$ and $P \equiv AM$.

6. Conclusions

[28] The variance of the sample regression coefficient is expressed as a function of sample size and the autocorrelation coefficients of relevant order. Thus the variance of the

Table 4. Effectiveness in Prewhitening P, $\hat{\eta}$, With $P' \equiv M$ Characterized by $\nu = \hat{\rho}_1$

	ρ_1					
n	0	0.1	0.3	0.6	0.9	0.99
			$P \equiv M$			
10	−0.30	−0.14	0.17	0.53	0.78	0.83
30	−0.08	0.10	0.42	0.77	0.95	0.98
60	−0.03	0.16	0.47	0.81	0.98	0.99
100	−0.03	0.17	0.49	0.82	0.98	0.99
500	−0.00	0.19	0.51	0.84	0.99	1.00
			$P \equiv AM$			
10	–[a]	−0.13	0.16	0.53	0.80	0.83
30	–[a]	0.11	0.42	0.72	0.95	0.98
60	–[a]	0.16	0.47	0.81	0.98	0.99
100	–[a]	0.17	0.49	0.82	0.98	1.00
500	–[a]	0.19	0.51	0.84	0.99	1.00

[a]Situation is the same as prewhitening $P \equiv M$ with $\rho_1 = 0$, i.e., $P \equiv I$, with $P' \equiv M$ and therefore was not assessed.

regression coefficient may be accurately determined for any process whose correlogram may be expressed in analytical form. Given the variance of the regression coefficient, the variance inflation factor for a particular stochastic process may be determined in a straightforward manner. The variance inflation factor increases monotonically with sequence length, n, conditioned on a specific value of the first order autocorrelation coefficient, ρ_1. Given n, the variance inflation factor increases monotonically with ρ_1 until ρ_1 becomes very high. For values of ρ_1 in the hydrologic range for annual flow sequences, say $\rho_1 < 0.5$, the variance inflation factor for a FGN process is greater than that for an AM process, which in turn is greater than that for a M process, which in turn is greater than that for an P $\equiv$ I process.

[29] Prewhitening with $P' \equiv M$ in the case where ρ_1 is known yields smaller inflation factors for $P \equiv FGN$ than for $P \equiv AM$ unless ρ_1 is exceedingly large. Relative to not prewhitening, the effectiveness of prewhitening with $P' \equiv M$ in the case where ρ_1 is known is very nearly the same for $P \equiv AM$ and $P \equiv FGN$. Prewhitening $P \equiv M$ or $P \equiv AM$ with $P' \equiv M$ is of equal effectiveness. The effectiveness of prewhitening $P \equiv AM$ with $P' \equiv M$ is nearly the same whether ρ_1 is known or not.

References

Box, G. E. P., and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 575 pp., Holden-Day, Boca Raton, Fla., 1976.

Douglas, E. M., R. M. Vogel, and C. N. Kroll, Trends in floods and low flows in the United States: Impact of spatial correlation, *J. Hydrol.*, *240*(1–2), 90–105, 2000.

Kendall, M., *Time-Series*, 2nd ed., 40 pp., Hefner, New York, 1975.

Mandelbrot, B. B., and M. S. Taqqu, Robust R/S analysis of long run serial correlation, paper presented at the 42nd Session of the International Statistical Institute, Int. Stat. Inst., Manila, 4–14 Dec. 1979.

Matalas, N. C., and J. R. Olsen, Analysis of trends and persistence in hydrologic records, *Risk-Based Decision Making In Water Resources IX*, *Proceedings of the Ninth Conference*, edited by Y. Y. Haimes, D. A. Moser, and E. Z. Stakhiv, pp. 61–76, Am. Soc. of Civ. Eng., Reston, Va., 2001.

von Storch, H., Misuses of statistical analysis in climate research, in *Analysis of Climate Variability*, edited by H. von Storch, and A. Navarra, pp. 11–26, Springer-Verlag, New York, 1999.

Wilkes, D. S., Resampling hypothesis tests for autocorrelated fields, *J. Clim.*, *10*, 65–82, 1997.

Yue, S., and P. Pilon, Interaction between deterministic trend and autoregressive process, *Water Resour. Res.*, *39*(4), 1077, doi:10.1029/2001WR001210, 2003.

Yue, S., and C. Y. Wang, Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test, *Water Resour. Res.*, *38*(6), 1068, doi:10.1029/2001WR000861, 2002.

Yue, S., P. Pilon, B. Phinney, and G. Cavadias, The influence of autocorrelation on the ability to detect trend in hydrological series, *Hydrol. Processes*, *16*, 1807–1829, 2002.

Zheng, X., R. E. Basher, and C. S. Thompson, Trend detection in regional-mean temperature series: Maximum, minimum, diurnal range and SST, *J. Clim.*, *11*, 317–326, 1997.

N. C. Matalas, 709 Glyndon Street, S.E., Vienna, VA 22180, USA. (nmatalas@aol.com)

A. Sankarasubramanian, International Research Institute for Climate Prediction, Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA. (sankar@iri.columbia.edu)