Comparison of methods to account for autocorrelation in correlation analyses of fish data

Brian J. Pyper and Randall M. Peterman

Abstract: Autocorrelation in fish recruitment and environmental data can complicate statistical inference in correlation analyses. To address this problem, researchers often either adjust hypothesis testing procedures (e.g., adjust degrees of freedom) to account for autocorrelation or remove the autocorrelation using prewhitening or first-differencing before analysis. However, the effectiveness of methods that adjust hypothesis testing procedures has not yet been fully explored quantitatively. We therefore compared several adjustment methods via Monte Carlo simulation and found that a modified version of these methods kept Type I error rates near α. In contrast, methods that remove autocorrelation control Type I error rates well but may in some circumstances increase Type II error rates (probability of failing to detect some environmental effect) and hence reduce statistical power, in comparison with adjusting the test procedure. Specifically, our Monte Carlo simulations show that prewhitening and especially first-differencing decrease power in the common situations where low-frequency (slowly changing) processes are important sources of covariation in fish recruitment or in environmental variables. Conversely, removing autocorrelation can increase power when low-frequency processes account for only some of the covariation. We therefore recommend that researchers carefully consider the importance of different time scales of variability when analyzing autocorrelated data.

Résumé : L'autocorrélation du recrutement des poissons et des données écologiques peut compliquer l'inférence statistique dans les analyses de corrélation. Pour régler ce problème, les chercheurs ajustent souvent les méthodes de vérification des hypothèses (p. ex. ajustement des degrés de liberté) afin de tenir compte de l'autocorrélation, ou éliminent l'autocorrélation par l'introduction de bruit blanc ou le calcul des différences premières. Toutefois, l'efficacité des méthodes servant à l'ajustement des méthodes de vérification des hypothèses n'a pas encore été entièrement étudiée sur le plan quantitatif. Nous avons donc comparé plusieurs méthodes d'ajustement par la simulation de Monte Carlo et nous avons constaté qu'une version modifiée de ces méthodes permet de maintenir le taux d'erreur de type I près de α. Par contraste, les méthodes qui éliminent l'autocorrélation contrôlent bien le taux d'erreur de type I mais dans certains cas, elles augmentent le taux d'erreur de type II (probabilité de ne pas déceler certains effets écologiques) et donc, réduisent l'efficacité statistique comparativement à l'ajustement de la méthode de vérification. Plus particulièrement, nos simulations de Monte Carlo montrent que l'introduction de bruit blanc, et surtout le calcul des différences premières, diminuent l'efficacité dans des situations ordinaires où les processus à basse fréquence (à évolution lente) sont des sources importantes de covariation du recrutement des poissons ou des variables de l'environnement. Réciproquement, l'élimination de l'autocorrélation peut accroître l'efficacité statistique lorsque les processus à basse fréquence ne représentent qu'une certaine partie de la covariation. Nous recommandons donc aux chercheurs de considérer attentivement l'importance des différentes échelles de temps de la variabilité lorsqu'ils analysent des données autocorrélées.

[Traduit par la Rédaction]

Introduction

One of the greatest challenges in fish biology and management is to understand mechanisms associated with highly variable survival rates and recruitment in fish populations. This variability can occur over various time scales or frequencies. For instance, rapid year-to-year changes (variability at higher frequencies) in fish survival rates, zooplankton abundance, and oceanographic variables are commonly observed characteristics of marine ecosystems, where values of a given variable are largely independent from one year to the next. In recent years, however, fisheries scientists and oceanographers have found increasing evidence of slowly changing, long-term variations (variability at lower frequencies) in these biological and physical variables (e.g., Beamish 1995), such that values in a given year are closely related to values in previous years (i.e., the data contain positive autocorrelation). This longer term variability has been exhibited in the Northeast Pacific Ocean, for example, by extended periods of unusually high sea-surface temperatures, weak upwelling, and low biological productivity followed by periods of the opposite.

Correlation analysis has been a useful and widely applied tool for generating hypotheses about the effects of environ-

Received September 19, 1997. Accepted June 23, 1998.
J14211

B.J. Pyper[1] and R.M. Peterman. School of Resource and Environmental Management, Simon Fraser University, Burnaby, BC V5A 1S6, Canada.

[1] Author to whom all correspondence should be addressed. e-mail: pyper@sfu.ca

mental or other variables on recruitment at these various time scales (e.g., Myers et al. 1995*a*, and papers cited therein). However, a major statistical challenge exists when time series of recruitment and environmental data are strongly autocorrelated, i.e., dominated by low-frequency variability (e.g., Chelton 1984; Thompson and Page 1989). Such autocorrelation violates the assumption of serial independence required for most classical inference tests (Hurlbert 1984). In general, this means that a sample correlation between two autocorrelated time series has fewer degrees of freedom (or a larger variance) than that assumed under the classical significance test. Consequently, the test will have a Type I error rate greater than the specified α (i.e., there will be an increased chance of concluding that a correlation is statistically significant when in fact no correlation is present; Jenkins and Watts 1968, p. 338).

To address this problem, fisheries scientists and oceanographers have typically applied two qualitatively different approaches. One approach is to modify the hypothesis testing procedure by computing either a corrected degrees of freedom for the sample correlation (Garrett and Petrie 1981; Chelton 1984) or, equivalently, a corrected variance for the sample correlation (Kope and Botsford 1990). The usefulness of these methods has been questioned, however, because they depend on the autocorrelation function of each time series, which can be poorly estimated for short time series typical of fisheries data sets (Thompson and Page 1989). Despite the importance of this statistical problem, we are not aware of any publications examining the ability of these methods to control Type I error rates. This paper therefore attempts to fill this gap.

The second approach to dealing with autocorrelation attempts to remove it from each time series of data before computing and testing correlations. This has been done either by fitting time series models to the data and using residuals from them (often referred to as "prewhitening"; e.g., Milicich et al. 1992; Quinn and Niebauer 1995) or by "first-differencing" the data (subtracting each data point from the next; e.g., Thompson and Page 1989). The premise behind these methods is that if the new transformed data series are free of autocorrelation, then classical inference tests are appropriate for correlations computed between them. However, because removing autocorrelation is equivalent to removing low-frequency variability from data, a potential drawback of prewhitening and first-differencing seems clear: if the low-frequency components of variability in two time series of recruitment or environmental data are *common* (i.e., synchronous or asynchronous), as some researchers are finding (see Beamish 1995), then removing autocorrelation may also remove much of this covariance (Thompson and Page 1989). Such procedures may thus increase the probability that researchers will miss finding some important relationship between long-term, slowly changing environmental processes and fish population dynamics. In other words, they may face an increased Type II error rate (i.e., probability of failing to reject the null hypothesis of no correlation when it is in fact false). In many of the papers we reviewed, however, the researchers did not seem to recognize this potential drawback of removing autocorrelation.

This paper has two main objectives. First, we use Monte Carlo simulations to examine the effectiveness of several commonly employed methods for adjusting the test procedure of the null hypothesis of no correlation between autocorrelated data (Garrett and Petrie 1981; Chelton 1984; Kope and Botsford 1990). Specifically, we examine how closely the Type I error rates of these methods match the specified α. Second, we use a simple model and Monte Carlo simulations to illustrate the degree to which prewhitening and first-differencing can alter levels of real covariation between autocorrelated data and affect Type II error rates in comparison with the best method for adjusting the test procedure. For completeness, we also examine how levels of covariation and Type II error can be altered by "smoothing" data, a method that, in contrast with prewhitening and first-differencing, is frequently used to remove high-frequency components of variability from data. In addition, we provide an empirical example using recruitment data for several stocks of sockeye salmon (*Oncorhynchus nerka*) from Bristol Bay, Alaska. We limit our analyses of Type I and Type II error rates to the case where correlations are being computed at a single lag, which is typical of most fisheries applications. Furthermore, because our main intended audience is fisheries biologists, all methods and results are in the context of the time domain (i.e., correlations and autocorrelations), rather than the frequency domain often used by oceanographers (i.e., spectral analysis).

Adjusting the test procedure of a sample correlation

Garrett and Petrie (1981), Chelton (1984), and Kope and Botsford (1990) provide similar methods for adjusting the null hypothesis test of a sample correlation between two autocorrelated time series, say *X* and *Y*. These methods, which are based on formulas from either Bayley and Hammersley (1946) or Bartlett (1946), assume that the time series are stationary, such that their underlying means remain constant over time. Furthermore, when testing a sample correlation at a given lag *k*, denoted $r_{XY}(k)$, each of these methods assumes a null hypothesis of no correlation at *all* lags.

These methods can be summarized using the following theoretical approximation of the "effective" number of degrees of freedom, N^*, of $r_{XY}(k)$

$$(1) \qquad \frac{1}{N^*} \approx \frac{1}{N} + \frac{2}{N}\sum_{j=1}^{\infty}\frac{(N-j)}{N}\rho_{XX}(j)\rho_{YY}(j)$$

where N is the sample size and $\rho_{XX}(j)$ and $\rho_{YY}(j)$ are the autocorrelations of *X* and *Y* at lag *j*. For example, Garrett and Petrie (1981) used a form of eq. 1 where $\rho_{XX}(j)\rho_{YY}(j)$ was replaced by $\rho(j)$, the autocorrelation at lag *j* of the *cross-product* of *X* and *Y* (each standardized to have a mean of zero). Given N^*, Garrett and Petrie (1981) used the standard critical value for $r_{XY}(k)$ at the α significance level that can be read from statistical tables or derived using the *t* distribution for either one- or two-tailed tests (Zar 1984, p. 309):

$$(2) \qquad r_{\text{crit}} = \sqrt{t^2_{\alpha,N^*}(t^2_{\alpha,N^*} + N^*)}.$$

Chelton (1984) recommended an expression for N^* that is equivalent to eq. 1 but without the weighting function $(N - j)/N$. Chelton also recommended an alternative critical value based on the chi-square distribution with 1 degree of freedom, which is equivalent to the following, based on the standard normal distribution (Z):

$$(3) \qquad r_{crit} = Z_{\alpha/2}\sqrt{1/N^*}$$

where values of α and $\alpha/2$ are used for one- and two-tailed tests, respectively.

Kope and Botsford (1990) provided an approximation for the *variance* of a sample correlation that is equivalent to $1/N^*$ as defined by eq. 1. They assumed that the null distribution of $r_{XY}(k)$ is normal, which also implies the critical value defined by eq. 3. Thus, the only difference between their method and that of Chelton is that Kope and Botsford included the weighting function $(N - j)/N$ in the computation of N^* (eq. 1).

Each of the above methods for adjusting the test procedure attempts to maintain Type I error rates at the specified α by generating critical values that account for autocorrelation. When both time series contain positive autocorrelation, it is clear from eq. 1 that values of N^* will be less than N, resulting in larger, and hence more conservative, critical values (i.e., fewer rejections of the null hypothesis of no correlation than if no adjustment were made). However, these methods may perform poorly in practice for a number of reasons: (*i*) the expressions for N^* are based on asymptotic (large sample size) formulas that may not be accurate for short time series, (*ii*) values of N^* depend on estimates of autocorrelation functions that are known to be both imprecise and biased, especially for short time series (Jenkins and Watts 1968, chap. 5), and (*iii*) the distributions assumed by critical values (2) and (3) may be inappropriate, particularly for short time series with high levels of autocorrelation. In addition, it is unclear how many lags j should be used when computing N^*. Given these potential problems, we conducted the following simulation analysis to examine the performance of these methods.

Methods—adjusting the test procedure

We used simple time series models to generate hypothetical data sets with lengths (number of years) and levels of positive autocorrelation that were typical of recruitment and environmental time series. Recruitment series used in correlation analyses often vary from about 15 years to usually not more than 50 years (e.g., Myers et al. 1995*a*), and it is common to observe estimates of lag-1 autocorrelation ranging from moderate (e.g., 0.5) to high values (e.g., 0.8) (e.g., Koslow et al. 1987).

In our baseline analysis, we simulated two *independent* time series X and Y of length N using first-order autoregressive models (AR(1) models) (Box and Jenkins 1976, p. 56):

$$(4) \qquad X_t = \phi_X X_{t-1} + \varepsilon_t$$

and similarly for Y, where ε was randomly and normally distributed and ϕ_X is the autoregressive parameter. The variance of X as defined by eq. 4 is $\sigma_X^2 = \sigma_\varepsilon^2/(1-\phi_X^2)$, which was used to generate the initial value of X. Note that X has a theoretical autocorrelation function

$$(5) \qquad \rho_{XX}(j) = \phi_X^{|j|}$$

such that greater absolute values of ϕ_X give rise to more autocorrelated time series. Also note that the value of ϕ_X corresponds to the theoretical lag-1 autocorrelation in the time series. AR(1) models can produce a wide range of autocorrelated time series and often provide a reasonable description of fisheries and ecological data.

We generated X and Y with N ranging from 15 to 50 years (in increments of 5) and with the same autoregressive parameter ϕ (i.e., $\phi_X = \phi_Y$) ranging from 0.4 to 0.9 (in increments of 0.1). For each combination of N and ϕ, 5000 Monte Carlo trials were done in which we computed sample correlations between the two time series at lag zero (i.e., using the N pairs of simulated data) and tabulated Type I error rates for two-tailed tests of the null hypothesis of no correlation for three levels of α (0.02, 0.05, and 0.10) using the following testing methods: (A) the standard significance test using critical value (2) and $N - 2$ degrees of freedom (i.e., not making any adjustments for autocorrelation), (B) the Garrett–Petrie method, (C) the Chelton method, and (D) the Kope–Botsford method. For methods B–D, autocorrelations were estimated over the first $N/4$ lags (i.e., $j = 1$ to $N/4$ in eq. 1) using an estimator recommended by Box and Jenkins (1976, p. 32):

$$(6) \qquad r_{XX}(j) = \frac{\sum_{t=1}^{N-j}(X_t - \bar{X})(X_{t+j} - \bar{X})}{\sum_{t=1}^{N}(X_t - \bar{X})^2}.$$

On occasion, variability in autocorrelation estimates resulted in values of N^* that were greater than the sample size N. We followed Kope and Botsford (1990) and constrained N^* to a maximum of N.

We examined the sensitivity of the methods to the choice of critical value, autocorrelation estimator, the number of lags for which autocorrelations were estimated, and the type of time series model used in the simulations.

Results and discussion—adjusting the test procedure

A summary of Type I error rates is shown in Table 1 for each of the methods A–D described above. Results for the standard significance test provide a baseline for comparison (Table 1, Case A). In this case, error rates were consistently much greater than the specified α values and increased rapidly as ϕ increased, as illustrated in Fig. 1A for $\alpha = 0.05$. This bias emphasizes the importance of accounting for autocorrelation. However, the efficacy of the methods designed to control Type I error rates varied considerably. The Garrett–Petrie method performed poorly at the three α values (Case B), with error rates larger than α and increasing with ϕ (e.g., Fig. 1B). In contrast, the Chelton and Kope–Botsford methods gave very similar error rates (Cases C and D) that were somewhat conservative at $\alpha = 0.02$, reasonably accurate although variable at $\alpha = 0.05$, but greater and more variable than desired at $\alpha = 0.10$. Error rates of these two

Table 1. Summary of simulated Type I error rates for different methods of adjusting the hypothesis test procedure, different time series models, and three levels of α.

		Hypothesis test procedure		Type I error rates					
				α = 0.02		α = 0.05		α = 0.10	
Case	Time series model	Method	No. of lags *j*	Average	Range	Average	Range	Average	Range
A	AR(1)^a	Standard significance test		0.143	0.034–0.389	0.212	0.080–0.474	0.29	0.15–0.55
B	AR(1)^a	Garrett–Petrie	*N*/4	0.095	0.027–0.215	0.154	0.063–0.299	0.23	0.12–0.39
C	AR(1)^a	Chelton	*N*/4	0.013	0.007–0.017	0.056	0.041–0.078	0.14	0.10–0.23
D	AR(1)^a	Kope–Botsford	*N*/4	0.015	0.008–0.018	0.062	0.043–0.098	0.15	0.10–0.25
E	AR(1)^a	Chelton using theoretical *N**		0.006	0.000–0.016	0.024	0.000–0.047	0.07	0.00–0.10
F	AR(1)^a	Chelton; theoretical *N**; critical value (2)		0.013	0.000–0.023	0.037	0.001–0.058	0.08	0.01–0.12
G	AR(1)^a	Chelton using critical value (2)	*N*/4	0.031	0.019–0.066	0.081	0.051–0.168	0.16	0.11–0.28
H	AR(1)^a	Modified Chelton	*N*/4	0.018	0.013–0.022	0.053	0.043–0.069	0.12	0.09–0.16
J	AR(1)^a	Modified Chelton with eq. 7	*N*/4	0.014	0.011–0.017	0.044	0.037–0.047	0.10	0.09–0.12
K	AR(1)^a	Modified Chelton with eq. 7	*N*/3	0.012	0.009–0.015	0.038	0.032–0.044	0.09	0.08–0.10
L	AR(1)^a	Modified Chelton with eq. 7	*N*/5	0.016	0.011–0.018	0.047	0.043–0.051	0.11	0.09–0.12
M	AR(1)^a	Modified Chelton with eq. 7	*N*/6	0.017	0.012–0.021	0.049	0.044–0.057	0.11	0.09–0.14
N	AR(2)^b	Standard significance test		0.141	0.026–0.384	0.212	0.064–0.469	0.29	0.13–0.55
P	AR(2)^b	Modified Chelton with eq. 7	*N*/5	0.016	0.011–0.021	0.049	0.040–0.058	0.11	0.09–0.14
Q	AR(2)^c	Standard significance test		0.118	0.078–0.358	0.265	0.140–0.436	0.35	0.22–0.52
R	AR(2)^c	Modified Chelton with eq. 7	*N*/5	0.011	0.005–0.021	0.035	0.021–0.061	0.08	0.06–0.13
S	ARIMA(1,1,0)^d	Standard significance test		0.593	0.303–0.768	0.654	0.395–0.808	0.71	0.48–0.84
T	ARIMA(1,1,0)^d	Modified Chelton with eq. 7	*N*/5	0.016	0.011–0.021	0.055	0.050–0.062	0.14	0.13–0.15
U	ARIMA(1,1,0)^e	Modified Chelton with eq. 7	*N*/5	0.047	0.033–0.063	0.114	0.081–0.155	0.22	0.17–0.28

Note: The average and range of each set of simulations were computed across all combinations of *N* (ranging from 15 to 50 in increments of 5) and the parameter values of the time series model. See text for details.

[a] Parameter range for ϕ: 0.4 to 0.9 (increments of 0.1).

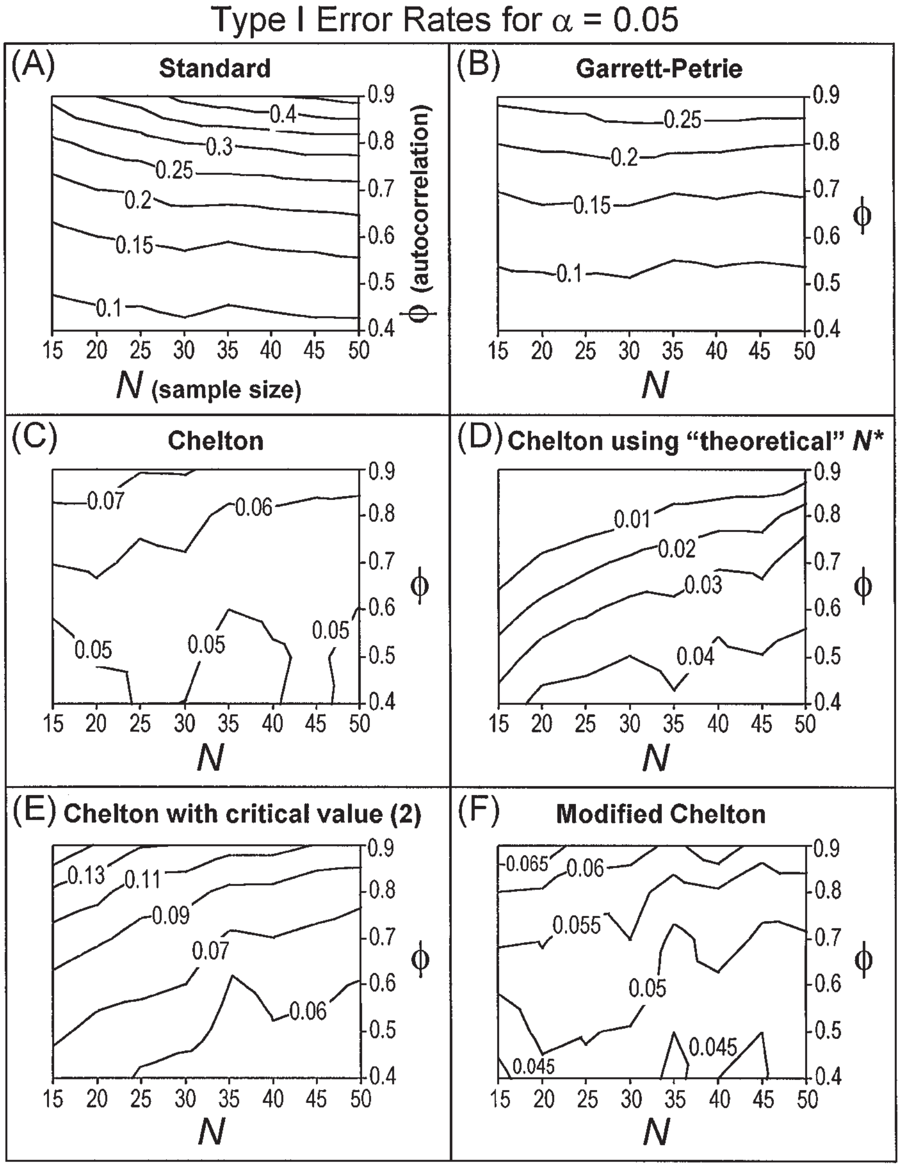
[b] Parameter range for ϕ_1: 0.2 to 0.7 (increments of 0.1); ϕ_2: 0.2 to $(0.9 - \phi_1)$ (increments of 0.1).

[c] Parameter range for ϕ_1: 0.8 to 1.8 (increments of 0.2); ϕ_2: $-\phi_1/2$.

[d] Parameter range for ϕ: 0.0 to 0.6 (increments of 0.3).

[e] Parameter value for ϕ: 0.9.

Fig. 1. Isopleths of simulated Type I error rates at $\alpha = 0.05$ for (A) the standard inference test, (B) the Garrett–Petrie method, (C) the Chelton method, (D) the Chelton method using theoretical values of N^*, (E) the Chelton method using critical value (2) instead of critical value (3), and (F) the Modified Chelton method (i.e., using critical value (2) with $N^* - 2$). See text for details.



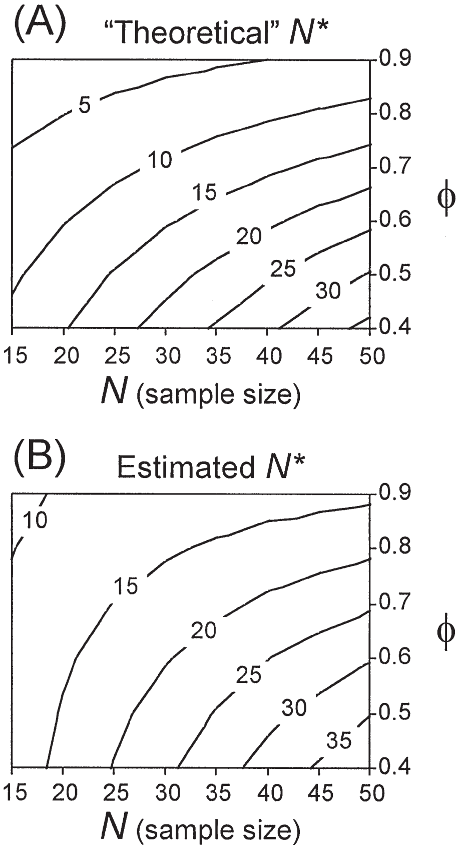
methods also increased as ϕ increased, especially for small N (e.g., Fig. 1C).

The Garrett–Petrie method was not as effective as the Chelton and Kope–Botsford methods primarily because estimates of $\rho(j)$, the autocorrelation function of the cross-product used in eq. 1, were highly imprecise and often negative for lags greater than 1 or 2, resulting in large overestimates of N^*. Changing the number of lags j used to estimate N^* did little to improve performance. For long time series (e.g., $N = 1000$), we found that estimates of $\rho(j)$ were very similar to those of $\rho_{XX}(j)\rho_{YY}(j)$ (the product used in the Chelton and Kope–Botsford methods), but for short time series, $\rho(j)$ was consistently smaller and much more variable. We therefore recommend that researchers not use the Garrett–Petrie approach.

For a given α value, Type I error rates of the Kope–Botsford method (Table 1, Case D) were slightly larger and more variable than those of the Chelton method (Case C). Recall that the only difference between these methods is that the Kope–Botsford method includes the weighting function $(N - j)/N$ in eq. 1. This added weighting function produced slightly larger estimates of N^*, which, over the various simulations we performed, consistently resulted in error rates that were slightly more variable than those of the Chelton method. We therefore limit further discussions of results to the Chelton method.

We found that the Type I error rates of the Chelton method were largely determined by an interaction among three main factors. First, it appears that eq. 1 (with or without the weighting function) does not provide accurate "theoretical" values of N^* for short, autocorrelated time series. For example, we used the theoretical autocorrelation func-

Fig. 2. Isopleths of (A) theoretical values of N^* computed using the theoretical autocorrelation functions of X and Y (eq. 5) and Chelton's form of eq. 1 and (B) average simulated estimates of N^* computed using Chelton's form of eq. 1.



tions of X and Y (eq. 5) and Chelton's form of eq. 1 to compute the theoretical N^* for each combination of N and ϕ (Fig. 2A). However, when these values of N^* were used with critical value (3), error rates were typically much too conservative (Table 1, Case E), falling well below the specified α values as ϕ increased (e.g., Fig. 1D). Results were similar for critical value (2) (Case F).

Second, estimates of autocorrelation were biased low in our simulations (i.e., toward less positive values), resulting in estimates of N^* that were considerably larger than the theoretical values of N^* (Fig. 2). Changing the number of lags j used to estimate N^* did little to reduce these differences. Standard estimators of autocorrelation can be seriously biased in this direction, with biases increasing for shorter time series or larger autocorrelations (e.g., Marriott and Pope 1954). Consequently, these biases compensated somewhat for the fact that theoretical values of N^* are overly conservative at high ϕ (e.g., Fig. 1D), resulting in error rates for the Chelton method that were reasonably accurate for $\alpha = 0.05$ (Fig. 1C).

Third, critical value (3) (based on the normal assumption) is inappropriate for autocorrelated data when N is small, just as it would be for the standard case where data are serially independent and critical value (2) is appropriate (Zar 1984, p. 309). When N is small (e.g., <100), the null distribution of a sample correlation has less weight in its tails (platykurtic) than assumed by critical value (3). This in part explains the tendency for error rates of the Chelton method to be lower than desired at $\alpha = 0.02$ and greater at $\alpha = 0.10$ (Table 1, Case C). Note, however, that using Chelton's form of eq. 1 with the standard critical value (2) gave larger and more variable Type I error rates for all combinations of N, ϕ, and α (Case G), with error rates increasing with ϕ (e.g., Fig. 1E). This result is not surprising given that critical value (3) will be larger than, and hence more conservative than, critical value (2) for a given estimate of N^*. Nevertheless, as we show below, critical value (2) *will* provide better coverage than critical value (3) when appropriate values of N^* are used. Thus, it would seem that the Chelton method (or the similar Kope–Botsford method) performed as well as it did because of a somewhat fortuitous and complex interaction between conservative values of theoretical N^*, biases in autocorrelation, and a conservative critical value (3).

Improvements to the Chelton method

Given these findings, we explored various adaptations of the Chelton method in an effort to improve its performance. The first and most useful adjustment we found was to use the standard critical value (2) instead of critical value (3), but with $N^* - 2$ rather than N^* degrees of freedom. We refer to this as the Modified Chelton method. Across all combinations of N and ϕ, this approach provided a better balance between estimates of N^* and Type I error rates at the three α levels (Table 1, Case H; Fig. 1F). In general, using $N^* - 1$ or $N^* - 3$ resulted in more variable error rates that were too large and too small, respectively. Despite the use of critical value (2), there was still a tendency for error rates to be lower than desired at $\alpha = 0.02$ and greater at $\alpha = 0.10$ (Case H). While this may indicate that critical value (2) is not entirely appropriate for autocorrelated data, it appeared to be largely caused by the fact that N^* is a random variable and hence, so is the critical value. This resulted in something analogous to the null distribution having less weight in its tails than assumed by critical value (2).

Performance was improved even further by using a different autocorrelation estimator (Chatfield 1989, p. 50):

$$(7) \qquad r_{XX}(j) = \frac{N}{N-j} \times \text{eq. 6.}$$

This estimator is often preferred to eq. 6 because it is less biased, although more variable (Chatfield 1989, p. 50). Using eq. 7 produced slightly smaller estimates of N^* (by roughly 0.5 on average) that were particularly influential at small N. As a result, error rates were reduced and, more importantly, were consistently less variable (Table 1, Case J). Unless stated otherwise, all further results are reported for the Modified Chelton method using eq. 7 to estimate autocorrelations.

A final factor influencing Type I error rates was the number of lags j for which autocorrelations were estimated in the computation of N^* (eq. 1). For example, increasing this number from $N/4$ to $N/3$ resulted in slightly smaller estimates of N^* (by roughly 0.3 on average) and overly conservative error rates for $\alpha = 0.02$ and 0.05 (Table 1, Case K). In contrast, using $N/5$ gave larger estimates of N^* (by only 0.2 on average) and provided perhaps the best combination of accuracy and precision in error rates for the three α values (Case L). Further reducing the number of lags to $N/6$ gave slightly larger and more variable error rates (Case M). Ideally, the optimal number of lags to use in a given situation would depend on both N and the theoretical autocorrelation functions of the time series. For instance, for a given N, error rates increased with ϕ (e.g., Fig. 1F), suggesting that more lags should be used at high ϕ than for low ϕ. However, because the theoretical autocorrelation functions are unknown and estimates of them can be highly variable, a reasonable approach is to use a simple rule, such as $N/5$, that yields reasonably accurate error rates with minimal variability across sample sizes and time series models.

Sensitivity analysis of the Modified Chelton method to various time series models

The results reported above were for the set of AR(1) models where both X and Y were simulated using the same value for the autoregressive parameter ϕ (i.e., $\phi_X = \phi_Y$) ranging from 0.4 to 0.9 (e.g., Table 1, Case L). For values of $\phi < 0.4$, error rates declined with the lowest error rates observed when $\phi = 0$ (i.e., no autocorrelation in either time series). In the latter case, the average error rates across N for $\alpha = 0.02$, 0.05, and 0.10 were 0.015, 0.040, and 0.086, respectively. In other words, the Modified Chelton method was slightly conservative at all α values when little or no autocorrelation was present. We also simulated cases where X and Y had different parameters (i.e., $\phi_X \neq \phi_Y$) ranging from 0 to 0.9; Type I error rates turned out to be very similar to the case where $\phi_X = \phi_Y$ (e.g., Case L). Thus, if AR(1) models with positive coefficients ϕ are a reasonable description of the time series being correlated (as they frequently are in fisheries research), then it appears that the Modified Chelton method using $N/5$ lags j provides a robust procedure for testing correlations (Case L).

However, it is possible that this method may not perform as well for other time series models that have different autocorrelation functions than AR(1) models. We therefore simulated additional models that often depict time series encountered in practice (see Box and Jenkins (1976) for details of the various models described below). For example, we simulated both X and Y using AR(2) models (i.e., $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t$) with values of ϕ_1 ranging from 0.2 to 0.7 (in increments of 0.1) and ϕ_2 ranging from 0.2 to $(0.9 - \phi_1)$ (in increments of 0.1). Again, Type I error rates of the standard significance test were typically much larger than the specified α (Table 1, Case N), while error rates for the Modified Chelton method were reasonably accurate and precise (Case P). For $\alpha = 0.05$ and 0.10, error rates for the Modified Chelton method were at or below 0.055 and 0.12, respectively, except in extreme cases when $\phi_1 + \phi_2 = 0.9$. Results were similar for simulations where values of ϕ_1 and ϕ_2 differed between X and Y.

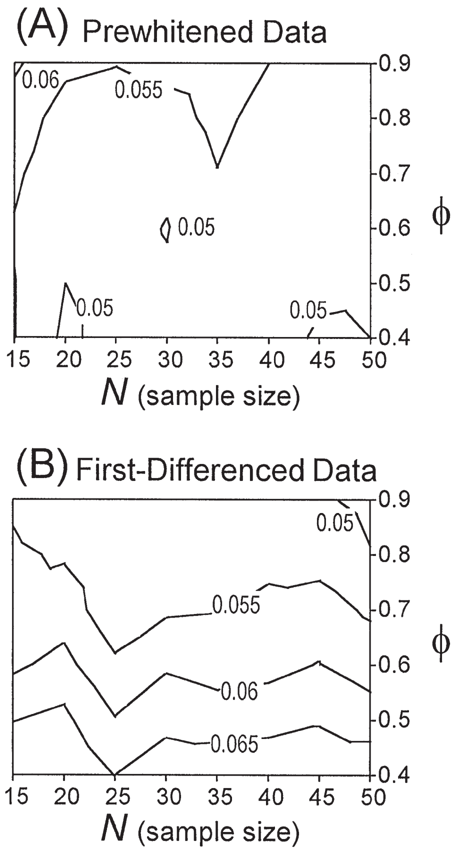
We examined an additional set of AR(2) models with ϕ_1 ranging from 0.8 to 1.8 (in increments of 0.2) and $\phi_2 = -\phi_1/2$. These models are of interest because they produce time series with apparent "cyclic" patterns (Box and Jenkins 1976, p. 62). A number of fish stocks in the Myers et al. (1995*b*) compendium of stock–recruitment data had autocorrelation functions of stock–recruitment residuals similar to the theoretical autocorrelation functions of these AR(2) models, namely, with large positive autocorrelations over the first several lags leading to negative autocorrelations at higher lags. For simulations with these models, the standard inference test once again gave large Type I error rates (Table 1, Case Q). In contrast, error rates for the Modified Chelton method were generally conservative (Case R). These conservative error rates were likely caused by the predominance of negative autocorrelations at higher lags. Because the autocorrelation estimator (eq. 7) is biased low, estimates of negative autocorrelations were slightly larger on average than their theoretical values. This resulted in lower values of N^*, and hence more conservative error rates, than were observed for the AR(1) and AR(2) models examined above.

We also examined a wide range of first- and second-order moving average models (MA(1) and MA(2) models) and mixed autoregressive moving average models (ARMA(1,1)) that produce time series with positive autocorrelations. However, in all cases, Type I error rates were roughly contained within the ranges observed for AR(1) models (i.e., Table 1, Case L). Furthermore, we simulated X and Y using various combinations of the AR(1) and AR(2) models described above, as well as models with moving average terms, and again found that averages and ranges of Type I error were consistent with those reported for Cases L, P, and R in Table 1.

Finally, we simulated a class of nonstationary time series to examine the performance of the Modified Chelton method under extreme conditions. As noted above, expressions for N^* are not valid for nonstationary data; however, for short time series, it may be difficult to distinguish between stationary and nonstationary processes. Specifically, we used ARIMA(1,1,0) models (i.e., $X_t = (1 + \phi)X_{t-1} - \phi X_{t-2} + \varepsilon_t$) with ϕ ranging from 0.0 (i.e., a random walk) to 0.9 (in increments of 0.3). The time series generated from such models will typically exhibit strong trends and high levels of autocorrelation. For values of ϕ ranging from 0.0 to 0.6, error rates of the standard inference test were extremely high (Table 1, Case S), yet the Modified Chelton method performed well, particularly at $\alpha = 0.02$ and 0.05 (Case T). However, when $\phi = 0.9$ (extreme trends, approaching an ARIMA(0,2,0) model), error rates for the standard test were greater than 0.7 for all combinations of N and α, while error rates of the Modified Chelton method were now well above the specified α levels (Case U).

In summary, results for the standard significance test of a sample correlation clearly illustrate that Type I error rates can be considerably greater than the specified α value when autocorrelation in the data is not taken into account. Unfortunately, there have been many published cases where correlation analyses were conducted but no adjustments were made for autocorrelation. Moreover, in cases where the testing procedure has been modified, the most commonly used method was that of Garrett and Petrie (1981), which we have shown to perform poorly at controlling Type I error rates. Nevertheless, our simulations showed that a modified version of the method of Chelton (1984) (or similarly, the method of Kope and Botsford (1990)) may often provide a useful procedure for testing sample correlations in the presence of positive autocorrelation. Specifically, we recommend that researchers use eq. 1 without the weighting function, with autocorrelations estimated over $N/5$ lags j using eq. 7, and with critical value (2) using $N^* - 2$ degrees of

Fig. 3. Isopleths of simulated Type I error rates at $\alpha = 0.05$ for prewhitening and first-differencing. See text for details.



freedom. Although we used a limited set of stationary models in our simulations, they will likely represent many of the time series encountered in practice. Using more highly parameterized models could give different results, depending on how different their autocorrelation functions are in comparison with those examined here, but in most cases, such models will have limited biological relevance.

Prewhitening, first-differencing, and smoothing

The previous section compared methods of adjusting the test procedure for sample correlations, but as an alternative, many researchers remove autocorrelation from their data. One common approach is to prewhiten the time series (e.g., Milicich et al. 1992; Quinn and Niebauer 1995), i.e., fit an appropriate time series model to each data set and use the residuals for computing correlations. An example is to fit an AR(1) model to time series X to give the prewhitened series X_{P}:

$$(8) \qquad X_{\mathrm{P}_t} = X_t - \hat{\phi}_X X_{t-1}.$$

Another very simple approach that is often effective at removing autocorrelation is to first-difference the time series (e.g., Thompson and Page 1989):

$$(9) \qquad \nabla X_t = X_t - X_{t-1}$$

where ∇X denotes the new, first-differenced time series. Note that first-differencing is equivalent to an extreme case of prewhitening where ϕ_X of the AR(1) model is equal to 1 (i.e., random walk), in which case, the process would be nonstationary (Box and Jenkins 1976, p. 56). Indeed, as discussed later, researchers have often first-differenced recruitment or environmental data prior to correlation analyses when these data exhibited strong time trends indicative of nonstationary time series.

To illustrate the potential effectiveness of prewhitening and first-differencing to control Type I error rates in tests of correlations, we used the baseline simulation procedure outlined above where AR(1) models were used to generate hypothetical data. Simulated time series were either prewhitened (by fitting AR(1) models to the data) or first-differenced, and then, correlations were computed between the new time series and tested using the standard significance test. Both procedures were very effective over the range of N and ϕ used in the simulations, as shown in Fig. 3 for $\alpha = 0.05$. Error rates for first-differencing did increase somewhat at low ϕ (Fig. 3B), where this procedure created moderate levels of negative lag-1 autocorrelation in the new time series, suggesting that first-differencing is most appropriate when high levels of autocorrelation are observed.

However, our concern in this section is not with Type I error rates but rather with how removing autocorrelation may affect Type II error rates in comparison with adjusting the test procedure. Specifically, removing autocorrelation (low-frequency variability) may limit a researcher's ability to detect the common effect of some slowly changing variable on fish population dynamics. This is quite important given the number of recent papers showing the presence of such slowly changing processes (e.g., Beamish 1995). Removing autocorrelation makes one or more implicit assumptions that low-frequency variability in time series (*i*) is unimportant, (*ii*) is not common between the two data series, (*iii*) potentially obscures the detection of common and more important high-frequency variability, or (*iv*) is caused by the same factors that cause high-frequency variability. However researchers rarely state such assumptions when removing autocorrelation and often do not seem to recognize the procedure's possible implications for their analyses. For example, if low-frequency variability is the dominant source of covariation between two time series, one would expect that removing autocorrelation from them, rather than adjusting the hypothesis testing procedure, may increase the Type II error rate (or equivalently, reduce statistical power (= 1 – Type II error rate)).

On the other hand, there may be cases where removing autocorrelation increases statistical power. This would be expected, for example, when high-frequency variability is the dominant source of covariation. Moreover, because autocorrelation can seriously reduce the degrees of freedom of a sample correlation, and hence limit statistical power, removing autocorrelation may increase power even when low-frequency variability is an important source of covariation by "restoring" degrees of freedom. In addition, it may be necessary to remove autocorrelation when there is a priori reason to believe that low-frequency patterns of variability may be spurious (e.g., Thompson and Page 1989).

In some cases, researchers who are particularly interested in patterns of low-frequency variability in fish data will smooth time series (e.g., Drinkwater and Myers 1987; Hollowed et al. 1987). Smoothing is an opposite approach to prewhitening or first-differencing because it removes high-frequency rather than low-frequency variation. In contrast with the vague assumptions often associated with prewhitening or first-differencing, smoothing is usually applied in correlation analyses with the clear assumption that high-frequency "noise" such as measurement error may obscure detection of common, important low-frequency variability. As an example of this approach, consider a simple two-point running mean:

$$(10) \qquad X_{\mathrm{S}_t} = (X_t + X_{t-1})/2$$

where X_{S} denotes the new, smoothed time series. Note that while smoothing can improve the detection of synchronous low-frequency variability (Davis 1977), it will increase autocorrelation in the time series and thus increase the probability of Type I error if the test procedure is not adjusted to account for the autocorrelation.

In the following section, we use a theoretical example to examine some of the possible implications of prewhitening, first-differencing, and, for comparison, smoothing time series in correlation analyses. In particular, we examine the extent to which these transformations can alter real correlations and affect the statistical power to detect them in comparison with using the original time series and the Modified Chelton test procedure.

Methods—prewhitening, first-differencing, and smoothing

Several researchers have attempted to generate hypotheses about sources of environmental variation on recruitment of fish stocks by testing whether separate stocks show covariation among themselves or with environmental variables (e.g., Koslow et al. 1987; Thompson and Page 1989). Consider the following simple model for how a correlation might arise between a pair of time series X and Y, where each is a linear combination of an "independent" source of variability, I_X and I_Y respectively, and a "common" source of variability, C:

$$(11) \qquad \begin{aligned} X_t &= I_X + aC_t \\ Y_t &= I_Y + bC_t \end{aligned}$$

where a and b are constants and I_X, I_Y, and C are independent of one another (i.e., uncorrelated). One could think of X and Y as time series of recruitment or environmental data that are both influenced by some unobservable, underlying environmental process C. We are interested in the case where each of the independent and common sources of variability is autocorrelated to some extent. Thus, we consider the simple case where I_X, I_Y, and C are each defined as an AR(1) process (eq. 4) with autoregressive parameters ϕ_{I_X}, ϕ_{I_Y}, and ϕ_C, respectively.

Intuitively, the greater the constants a and b (or variance of C), the greater the correlation will be between X and Y. Furthermore, larger values of ϕ_{I_X}, ϕ_{I_Y}, and ϕ_C will obviously result in greater autocorrelation in X and Y. However, when ϕ_C is large and ϕ_{I_X} and ϕ_{I_Y} are small, removing autocorrelation should also remove much of the common signal associated with C, thereby reducing the correlation between X and Y. The opposite should be true when ϕ_C is small and ϕ_{I_X} and ϕ_{I_Y} are large. We investigated the potential magnitude of these effects in the following ways.

Given eq. 11, we derived formulas for the theoretical asymptotic (large sample size) correlations between X and Y for the original time series and for series that were prewhitened via eq. 8, first-differenced (eq. 9), and smoothed (eq. 10) (see Appendix).

To examine the effects of prewhitening, first-differencing, and smoothing on the statistical power to detect real correlations, we used Monte Carlo simulations. Specifically, we simulated eq. 11 with the variances of I_X, I_Y, and C all held constant at 1. Simulations were done for $N = 20$ and $N = 50$. In each case, we generated time series of X and Y, with positive correlations between them, using values of a and b (i.e., $a = b$) from 0.4 to 2 (in increments of 0.2), values of ϕ_I (i.e., $\phi_{I_X} = \phi_{I_Y}$) from 0 to 0.9 (in increments of 0.1), and values of ϕ_C from 0 to 0.9 (in increments of 0.1). For each combination of (a,b), ϕ_I, and ϕ_C, 2000 Monte Carlo trials were done in which we computed sample correlations between X and Y at lag zero and tabulated statistical power (i.e., the proportion of cases where the null hypothesis of no correlation was rejected for a two-tailed test with $\alpha = 0.05$) for the following methods: (A) using the original (autocorrelated) time series and the Modified Chelton test procedure outlined above, (B) prewhitening each time series using AR(1) models and then using the standard inference test, (C) first-differencing each time series and the standard inference test, and (D) smoothing each time series using a two-point running mean and the Modified Chelton method.

Results and discussion—prewhitening, first-differencing, and smoothing

As demonstrated in formulas in the Appendix, prewhitening and, in particular, first-differencing can substantially reduce the asymptotic correlation (ρ_{XY}) between X and Y when the primary source of autocorrelation is also the source of covariation (i.e., $\phi_C > \phi_I$). For example, Table 2 presents asymptotic correlations for four combinations of ϕ_I and ϕ_C when $(a, b, \sigma_{I_X}, \sigma_{I_Y}, \sigma_C) = 1$. For the extreme case where $\phi_I = 0$ and $\phi_C = 0.9$, prewhitening and first-differencing will reduce ρ_{XY} from 0.50 to only 0.25 and 0.09, respectively. In contrast, smoothing will increase ρ_{XY} from 0.50 to 0.66. However, when independent sources of variability are the primary sources of autocorrelation ($\phi_I > \phi_C$), prewhitening and first-differencing will increase ρ_{XY}, while smoothing reduces it (right side of Table 2).

In the case where $\phi_C > \phi_I$, it is conceivable that reductions in correlations caused by prewhitening or first-differencing may be offset by the increase in degrees of freedom obtained by removing autocorrelation, which could increase or at least maintain statistical power. However, our simulations showed that in many cases, statistical power was reduced substantially. Figure 4 presents isopleths of changes in statistical power for three combinations of N and (a,b) when X and Y were either prewhitened, first-differenced, or smoothed. Positive values indicate increases in statistical power, while negative values (shaded areas in Fig. 4) indicate decreases. For example, a value of –0.2 means that out of 100 tests, there would be 20 *fewer* cases in which the null hypothesis was correctly rejected compared with using the

Table 2. Theoretical asymptotic correlations, computed from equations in Table A (see Appendix), between time series X and Y for the original time series, prewhitened series, first-differenced series, and smoothed series.

	Asymptotic correlation between X and Y (ρ_{XY})			
	$\phi_I = 0.0$		$\phi_C = 0.0$	
Transformation of data	$\phi_C = 0.6$	$\phi_C = 0.9$	$\phi_I = 0.6$	$\phi_I = 0.9$
(A) Original	0.50	0.50	0.50	0.50
(B) Prewhitened	0.40	0.25	0.60	0.75
(C) First-differenced	0.29	0.09	0.71	0.91
(D) Smoothed	0.62	0.66	0.38	0.34

Note: $(a, b, \sigma_{I_X}, \sigma_{I_Y}, \sigma_C) = 1$ and $\phi_{I_X} = \phi_{I_Y} = \phi_I$.

original time series and the Modified Chelton test procedure. (Note that eqs. A1 and A3 in the Appendix provide a rough indication of the levels of covariation and autocorrelation in the original time series X and Y for given values of (a,b), ϕ_I, and ϕ_C.)

For small values of (a,b) (i.e., small effect of the common factor C), prewhitening and first-differencing increased statistical power in most cases, especially for $\phi_I >> \phi_C$ (e.g., $N = 20$ and $(a,b) = 0.6$; top panels in Fig. 4). However, the situation changed appreciably when the common factor was a more important source of variability in X and Y (e.g., $N = 20$ and $(a,b) = 1.4$; middle panels in Fig. 4). Here, prewhitening and first-differencing resulted in reductions in power by as much as 0.2 or 0.3 for combinations where $\phi_C > \phi_I$ and resulted in increases in power that were greatest when both ϕ_I and ϕ_C were large. Larger changes in power were observed for given values of (a,b) when $N = 50$ (e.g., $N = 50$ and $(a,b) = 1.0$; bottom panels in Fig. 4).

The patterns shown in Fig. 4 for prewhitening and first-differencing were typical of the range of combinations of N and (a,b) that we explored. In general, these transformations increased power when $\phi_I > \phi_C$ by increasing degrees of freedom and levels of covariation. However, when $\phi_C > \phi_I$, removing autocorrelation typically reduced power despite increasing degrees of freedom because these procedures resulted in often large reductions in covariation.

Interestingly, smoothing using a two-point running mean increased power only slightly when $\phi_C >> \phi_I$ (Fig. 4). For most combinations where $\phi_C > \phi_I$, increases in covariation due to smoothing were offset by decreased degrees of freedom (through increased autocorrelation), resulting in reduced power. Using three-, four-, or five-point running means also resulted in only slight increases in power when $\phi_C >> \phi_I$.

Thus, as illustrated by this theoretical example, prewhitening and, in particular, first-differencing can remove much of the covariation between autocorrelated time series and consequently reduce the power to detect it *when* the source of that covariation is also the main source of the autocorrelation. Of course, the opposite is also true. That is, if autocorrelation is due in large part to variability that is unique to a given time series, then prewhitening or first-differencing can substantially increase power. It is not our purpose to present an exhaustive examination of the advantages or disadvantages of such methods as prewhitening, first-differencing, or smoothing, but rather to point out that such transformations may have unintended effects that alter the interpretation of analyses. By using prewhitening or first-differencing, past researchers may have missed detecting some environmental effect or some correlation among stocks that existed due to the common effect of some important but slowly changing variable.

A final frequently used method of treating time series in fisheries is to remove time trends. Recruitment and environmental time series often have distinct increasing or decreasing trends over time (e.g., Cohen et al. 1991) that could easily produce misleading correlations between the series when these trends are independent (Plosser and Schwert 1978) or could mask covariation at higher frequencies (Kope and Botsford 1990). Furthermore, strong trends indicate nonstationary time series, in which case, methods for adjusting the test procedure may not adequately control type I error rates. Thus, some researchers recommend removing time trends before computing correlations by either first-differencing (e.g., Cohen et al. 1991) or "detrending" the data (typically using residuals from a linear regression versus time) (e.g., Botsford and Brittnacher 1992). Once again, however, such practices may remove important information about a possible *common* mechanism influencing different data series. By removing time trends, we are assuming that they are unrelated, yet there are obvious mechanisms that could produce common time trends among recruitment data such as trends in environmental variables, habitat degradation, or trends in the abundances of competitor, prey, or predator species (e.g., Butler 1991). In addition, short time series can often appear to have deterministic or nonstationary trends, even though the series are stationary and only moderately autocorrelated. Thus, even though we may not be able to distinguish between time trends that are caused by a common mechanism and those that are not, researchers should consider both possibilities when choosing methods of analysis and they should interpret their results cautiously.

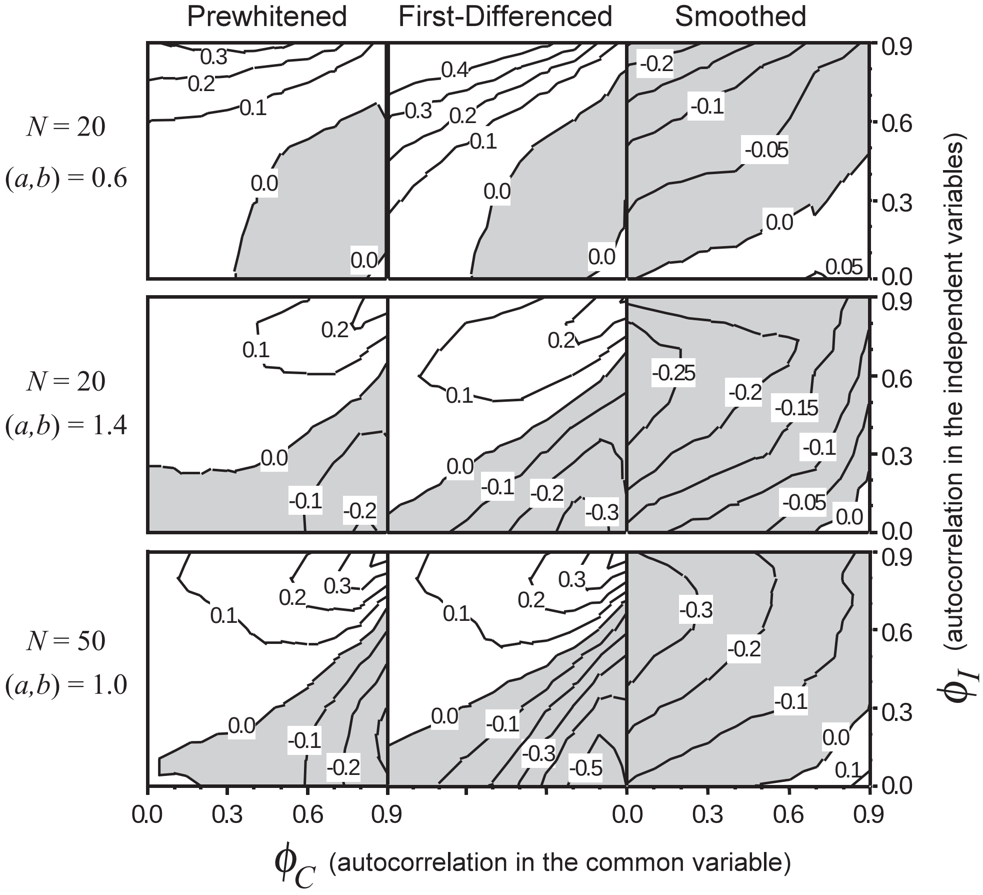
Empirical example: sockeye salmon of Bristol Bay, Alaska

The previous section quantified potential changes in correlations and statistical power that could be caused by prewhitening, first-differencing, and smoothing. Here, we briefly illustrate the possible effects of these transformations using recruitment data for four stocks of sockeye salmon from Bristol Bay, Alaska. There has been growing interest in the low-frequency variability, often referred to as interdecadal variability, of physical and biological processes in the North Pacific and their effects on the productivity of fish stocks such as Bristol Bay sockeye (e.g., Hare and Francis 1995; Adkison et al. 1996).

Methods—empirical example

We selected four stocks of Bristol Bay sockeye that had autocorrelated indices of survival rate: the Igushik, Kvichak, Ugashik, and Wood River stocks. For each stock, we used abundances of spawners and total recruits for 34 brood years from 1956 to 1989 (B. Cross, Alaska Department of Fish and Game, Anchorage, Alaska, personal communication). Using these data, we computed an index of survival rate (SR

Fig. 4. Isopleths of changes in statistical power for prewhitening, first-differencing, and smoothing compared with using the original time series and the Modified Chelton method for three sets of conditions of *N* and (*a*,*b*). Positive values indicate increases in statistical power over using the original time series, while negative values indicate decreases. See text for details.



index) for each stock to account for changes in spawner abundance, within-stock density dependence, and lognormal error structure. This index was the time series of residuals from the fit of the Ricker stock–recruitment relationship, i.e., linear regression of $\log_e$(recruits per spawner) versus spawners. The autocorrelation and partial autocorrelation functions of the four SR indices were consistent with AR(1) processes, with estimates of lag-1 autocorrelation ranging from 0.44 to 0.70. We therefore fit AR(1) models to these SR indices as an example of prewhitening.

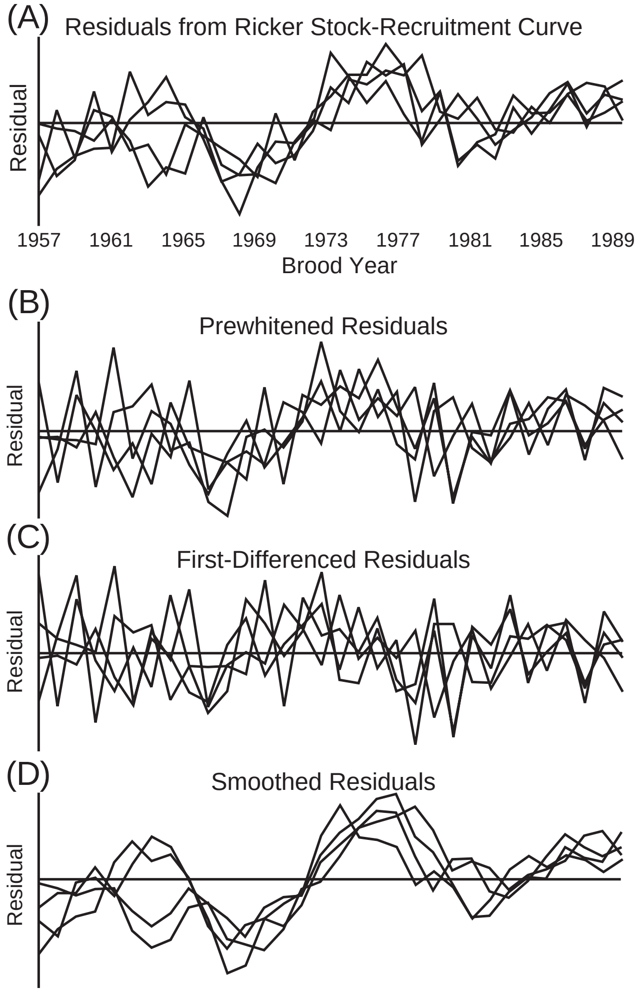
We computed pairwise correlations among the four SR data series using the original, prewhitened, first-differenced, and smoothed data (two-point running mean). Standard hypothesis tests (two-tailed, $\alpha = 0.05$) were used for correlations among prewhitened or first-differenced data, while the Modified Chelton method was used for correlations among the original or smoothed data. Statistical power was estimated for each pairwise comparison using the sample correlation and the equations given in Zar (1984, p. 312). Because these "analytical" equations are for serially independent data, they provide only a rough approximation of power for comparisons among the original and smoothed (i.e., autocorrelated) data. However, in our simulation analyses described above, we found that the analytical estimates of power consistently underestimated the actual, simulated power by roughly 0.05 on average for cases where levels of covariation and autocorrelation were similar to those observed here for the four Bristol Bay sockeye stocks. Thus, the estimates of power provided below for the original and smoothed data are likely slight underestimates of the actual statistical power.

Results and discussion—empirical example

The time series of SR indices and their transformations are shown in Fig. 5. Extended periods of negative residuals (1960s) and positive residuals (1970s) in each of the original SR indices provide evidence that *synchronous* low-frequency variability is an important feature of these data (Fig. 5A), suggesting a shared influence of some autocorrelated environmental factor. Prewhitening and, in particular, first-differencing appeared to remove much of the coherence between the SR indices (Figs. 5B and 5C). In contrast, using a two-point running mean to smooth the SR indices accentuated the low-frequency trends and the apparent coherence of the time series (Fig. 5D).

Correlations among the original and transformed SR indices (Table 3) reflect the patterns observed in Fig. 5. For example, prewhitening and first-differencing reduced the *average* correlation among the SR series from 0.50 to only 0.24 and 0.11, respectively (last row of Table 3), and reduced the number of significant correlations from 4 to 2 and 1, respectively. Moreover, prewhitening and, to a greater extent, first-differencing reduced the average statistical power from 0.5 to 0.34 and 0.18, respectively, despite having twice

Fig. 5. Time series for four Bristol Bay sockeye stocks of (A) residuals from the Ricker stock–recruitment curve (i.e., survival rate (SR) indices), (B) prewhitened SR indices, (C) first-differenced SR indices, and (D) smoothed SR indices.



the number of degrees of freedom as the average comparison among the original, autocorrelated SR indices (i.e., 31 versus an average of 14 degrees of freedom) (right half of Table 3). In contrast, smoothing increased the average correlation to 0.62, although statistical power changed little because of the corresponding decreases in degrees of freedom (Table 3).

Thus, this example represents a case where the researcher could easily miss detecting the shared influence of a slowly changing process on fish productivity by removing autocorrelation rather than adjusting the test procedure to account for autocorrelation. The tendency for correlations to be lower among prewhitened or first-differenced data and higher among smoothed data suggests that low-frequency components of variability were the dominant source of covariation in the original SR indices.

Conclusions

A growing body of literature documents the importance of slow changes, or low-frequency variability, in recruitment of fish species and in the physical and biological processes that

Table 3. Sample correlations and corresponding statistical power for comparisons among the survival rate (SR) indices of four Bristol Bay sockeye stocks: Igushik (IGU), Kvichak (KVI), Ugashik (UGA), and Wood (WOO).

	Effective degrees of freedom ($N^* - 2$)		Sample correlation coefficient				Statistical power			
Comparison	Original	Smoothed	Original	Prewhitened	First-differenced	Smoothed	Original	Prewhitened	First-differenced	Smoothed
(1) IGU vs. KVI	11	7	0.62*	0.43*	0.32	0.73*	0.65	0.70	0.44	0.64
(2) IGU vs. UGA	13	9	0.41	0.11	–0.02	0.52	0.31	0.08	0.02	0.35
(3) IGU vs. WOO	13	7	0.55*	0.28	0.14	0.68*	0.58	0.36	0.12	0.53
(4) KVI vs. UGA	21	13	0.33	0.05	–0.10	0.43	0.32	0.04	0.01	0.38
(5) KVI vs. WOO	16	8	0.51*	0.44*	0.35*	0.59	0.60	0.73	0.51	0.43
(6) UGA vs. WOO	12	7	0.56*	0.16	–0.06	0.76*	0.56	0.14	0.01	0.72
Average	**14**	**9**	**0.50**	**0.24**	**0.11**	**0.62**	**0.50**	**0.34**	**0.18**	**0.51**

Note: *Statistically significant ($P < 0.05$).

may affect recruitment (e.g., Hollowed and Wooster 1992). However, in an increasing number of applications in fisheries and ecological literature, autocorrelation (low-frequency variability) is removed from time series before correlation analyses are conducted in an effort to control Type I error rates (e.g., Milicich et al. 1992; Meekan et al. 1993; Robertson et al. 1993). We have shown that a specific method for adjusting the test procedure of a sample correlation may often provide a useful alternative to removing autocorrelation. Our theoretical and empirical examples showed that statistical power can be significantly greater for this adjustment method in comparison with removing autocorrelation when low-frequency variability is the major source of covariation, as many researchers are finding (e.g., Beamish 1995).

It is therefore important that fisheries researchers, when dealing with autocorrelated data, generate hypotheses about what time scales of variability in environmental processes are important. Given such hypotheses, researchers can then select appropriate tools for investigating patterns of covariation at these different time scales (e.g., Hollowed et al. 1987). For example, if high-frequency (i.e., rapid time scale) components of variability are the dominant source of covariation, then prewhitening (or first-differencing) may be useful because it should increase covariation and the statistical power to detect it by removing independent, low-frequency variability that obscures the detection of that covariation. Conversely, smoothing of some form may be useful for detecting low-frequency (slow time scale) sources of covariation, although correlations must be tested using the adjustment method outlined above to account for autocorrelation that will be present in the smoothed data. In any of these cases, researchers should be aware and cautious of the distorting effects that these various transformations can have on data.

While the concerns we have raised in this paper relate to simple correlation analysis, autocorrelation has similar implications for many other statistical methods such as regression (e.g., Bence 1995).

Acknowledgments

We thank Bev Cross for the Bristol Bay sockeye salmon data as well as the numerous biologists and technicians in the Alaska Department of Fish and Game who gathered the lengthy time series of data analyzed here. We thank Terrance Quinn II and Louis Botsford for useful suggestions on the draft manuscript and James R. Bence and an anonymous reviewer for their very useful comments. Funding was provided by grants to R.M. Peterman from the Ocean Production Enhancement Network (one of Canada's Networks of Centres of Excellence) and the Natural Sciences and Engineering Research Council of Canada.

References

Adkison, M.D., Peterman, R.M., Lapointe, M.F., Gillis, D.M., and Korman, J. 1996. Alternative models of climate effects on sockeye salmon, *Oncorhynchus nerka*, productivity in Bristol Bay, Alaska, and the Fraser River, British Columbia. Fish. Oceanogr. **5**: 137–152.

Bartlett, M.S. 1946. On the theoretical specification and sampling properties of autocorrelated time series. J. R. Stat. Soc. Ser. B (Methodol.), **8**: 27–41.

Bayley, G.V., and Hammersley, J.M. 1946. The "effective" number of independent observations in an autocorrelated time series. J. R. Stat. Soc. Ser. B (Methodol.), **8**: 184–197.

Beamish, R.J. (*Editor*). 1995. Climatic change and northern fish populations. Can. Spec. Publ. Fish. Aquat. Sci. No. 121.

Bence, J.R. 1995. Analysis of short time series: correcting for autocorrelation. Ecology, **76**: 628–639.

Botsford, L.W., and Brittnacher, J.G. 1992. Detection of environmental influence on wildlife: California quail as an example. *In* Wildlife 2001: populations. *Edited by* D.R. McCullough and R.H. Barrett. Elsevier Science, New York. pp. 158–169.

Box, G.E.P., and Jenkins, G.W. 1976. Time series analysis: forecasting and control. Revised ed. Holden-Day, San Francisco, Calif.

Butler, J.L. 1991. Mortality and recruitment of Pacific sardine, *Sardinops sagax caerulea*, larvae in the California Current. Can. J. Fish. Aquat. Sci. **48**: 1713–1723.

Chatfield, C. 1989. The analysis of time series: an introduction. 4th ed. Chapman and Hall, London, U.K.

Chelton, D.B. 1984. Commentary: short-term climatic variability in the Northeast Pacific Ocean. *In* The influence of ocean conditions on the production of salmonids in the North Pacific. *Edited by* W. Pearcy. Oregon State University Press, Corvallis, Oreg. pp. 87–99.

Cohen, E.B., Mountain, D.G., and O'Boyle, R. 1991. Local-scale versus large-scale factors affecting recruitment. Can. J. Fish. Aquat. Sci. **48**: 1003–1006.

Davis, R.E. 1977. Techniques for statistical analysis and prediction of geophysical fluid systems. Geophys. Astrophys. Fluid Dyn. **8**: 245–277.

Drinkwater, K.F., and Myers, R.A. 1987. Testing predictions of marine fish and shellfish landings from environmental variables. Can. J. Fish. Aquat. Sci. **44**: 1568–1573.

Garrett, C., and Petrie, B. 1981. Dynamical aspects of the flow through the Strait of Belle Isle. J. Phys. Oceanogr. **11**: 376–393.

Hare, S.R., and Francis, R.C. 1995. Climate change and salmon production in the northeast Pacific Ocean. *In* Climatic change and northern fish populations. *Edited by* R.J. Beamish. Can. Spec. Publ. Fish. Aquat. Sci. No. 121. pp. 357–372.

Hollowed, A.B., and Wooster, W.S. 1992. Variability of winter ocean conditions and strong year classes of Northeast Pacific groundfish. ICES Mar. Sci. Symp. **195**: 433–444.

Hollowed, A.B., Bailey, K.M., and Wooster, W.S. 1987. Patterns in recruitment of marine fishes in the northeast Pacific Ocean. Biol. Oceanogr. **5**: 99–131.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. **54**: 187–211.

Jenkins, G.M., and Watts, D.G. 1968. Spectral analysis and its applications. Holden-Day, San Francisco, Calif.

Kalbfleisch, J.G. 1985. Probability and statistical inference 1: probability. 2nd ed. Springer-Verlag, New York.

Kope, R.G., and Botsford, L.W. 1990. Determination of factors affecting recruitment of chinook salmon, *Oncorhynchus tshawytscha*, in central California. Fish. Bull. U.S. **88**: 257–269.

Koslow, J.A., Thompson, K.R., and Silvert, W. 1987. Recruitment to northwest Atlantic cod (*Gadus morhua*) and haddock (*Melanogrammus aeglefinus*) stocks: influence of stock size and environment. Can J. Fish. Aquat. Sci. **44**: 26–39.

Marriott, F.H.C., and Pope, J.A. 1954. Bias in the estimation of autocorrelations. Biometrika, **41**: 390–402.

Meekan, M.G., Milicich, M.J., and Doherty, P.J. 1993. Larval production drives temporal patterns of larval supply and recruitment of a coral reef damselfish. Mar. Ecol. Prog. Ser. **93**: 217–225.

Milicich, M.J., Meekan, M.G., and Doherty, P.J. 1992. Larval supply: a good predictor of recruitment of three species of reef fish (Pomacentridae). Mar. Ecol. Prog. Ser. **86**: 153–166.

Myers, R.A., Barrowman, N.J., and Thompson, K.R. 1995*a*. Synchrony of recruitment across the North Atlantic: an update. ICES J. Mar. Sci. **52**: 103–110.

Myers, R.A., Bridson, J., and Barrowman, M.J. 1995*b*. Summary of worldwide spawner and recruitment data. Can. Tech. Rep. Fish. Aquat. Sci. No. 2024.

Plosser, C.I., and Schwert, G.W. 1978. Money, income, and sunspots: measuring economic relationships and the effects of differencing. J. Monetary Econ. **4**: 637–660.

Quinn, T.J., II, and Niebauer, H.J. 1995. Relation of eastern Bering Sea walleye pollock (*Theragra chalcogramma*) recruitment to environmental and oceanographic variables. *In* Climatic change and northern fish populations. *Edited by* R.J. Beamish. Can. Spec. Publ. Fish. Aquat. Sci. No. 121. pp. 497–507.

Robertson, D.R., Schober, U.M., and Brawn, J.D. 1993. Comparative variation in spawning output and juvenile recruitment of some Caribbean reef fishes. Mar. Ecol. Prog. Ser. **94**: 105–113.

Thompson, K.R., and Page, F.H. 1989. Detecting synchrony of recruitment using short, autocorrelated time series. Can. J. Fish. Aquat. Sci. **46**: 1831–1838.

Zar, J.H. 1984. Biostatistical analysis. Prentice-Hall, Englewood Cliffs, N.J.

Appendix

Based on eq. 11, we derived theoretical expressions for the asymptotic correlation between the original time series of X and Y and between prewhitened, first-differenced, and smoothed versions of X and Y by evaluating the variances and covariances of these series (e.g., Kalbfleisch 1985, pp. 176–179). For example, it can be shown that the asymptotic Pearson product-moment correlation between the *original* time series of X and Y as defined by eq. 11 is

$$\text{(A1)} \quad \rho_{XY} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{ab\sigma_C^2}{\sigma_X \sigma_Y}$$

where

$$\text{(A2)} \quad \sigma_X^2 = \sigma_{I_X}^2 + a^2 \sigma_C^2$$

and similarly for σ_Y^2.

This asymptotic correlation is not affected by the presence of autocorrelation in either I_X, I_Y, or C. However, when I_X, I_Y, and C are defined as AR(1) processes as in eq. 4 (with first-order autoregressive parameters σ_{I_X}, σ_{I_Y}, and ϕ_C, respectively), it can be shown that X will have an asymptotic autocorrelation function

$$\text{(A3)} \quad \rho_{XX}(j) = \frac{\sigma_{I_X}^2 \phi_{I_X}^{|j|} + a^2 \sigma_C^2 \phi_C^{|j|}}{\sigma_X^2}$$

and similarly for Y. Note that the shape of the autocorrelation function defined by eq. A3 is similar but not identical to that of an AR(1) process (eq. 5) except when $\phi_{I_X} = \phi_C$.

To examine how prewhitening, first-differencing, and smoothing could affect the asymptotic correlation between X and Y, we used eq. 8 (prewhitening using an AR(1) model), eq. 9 (first-differencing), and eq. 10 (smoothing using a two-point running mean) to transform the series. When fitting AR(1) processes to X and Y, the first-order auroregressive parameters would, asymptotically, be equal to $\rho_{XX}(1)$ and $\rho_{YY}(1)$ as defined by eq. A3. For simplicity, we denote these parameters as ϕ_X and ϕ_Y.

The asymptotic correlations between the prewhitened, first-differenced, and smoothed times series of X and Y are shown in Table A. First, note that these correlations equal the original correlation (eq. A1) when there is no autocorrelation present (i.e., ϕ_{I_X}, ϕ_{I_Y}, and ϕ_C equal zero, and hence, ϕ_X and ϕ_Y also equal zero). However, if the common variable C is autocorrelated (i.e., $\phi_C > 0$), then the correlations between X and Y are *less* than the original correlation for both prewhitening and first-differencing and *greater* for smoothing. Larger values of ϕ_C result in larger differences in the correlations. The opposite is true if I_X and I_Y are autocorrelated but C is not. That is, prewhitening and first-differencing increase the correlation between X and Y, while smoothing reduces it.

Table A. Asymptotic correlations between prewhitened, first-differenced, and smoothed time series of X and Y (eq. 11) when I_X, I_Y, and C are defined as AR(1) processes.

Method of transformation	Asymptotic correlation between X and Y	Variances of X and Y
Prewhitening (eq. 8)	$\rho_{X_\text{P}Y_\text{P}} = \frac{ab\sigma_C^2(1 - \phi_C\phi_X - \phi_C\phi_Y + \phi_X\phi_Y)}{\sigma_{X_\text{P}}\sigma_{Y_\text{P}}}$	$\sigma_{X_\text{P}}^2 = \sigma_{I_X}^2(1 - 2\phi_{I_X}\phi_X + \phi_X^2) + a^2\sigma_C^2(1 - 2\phi_C\phi_X + \phi_X^2)$ and similarly for $\sigma_{Y_\text{P}}^2$
First-differencing (eq. 9)	$\rho_{\nabla X \nabla Y} = \frac{2ab\sigma_C^2(1 - \phi_C)}{\sigma_{\nabla X}\sigma_{\nabla Y}}$	$\sigma_{\nabla X}^2 = 2\sigma_{I\ X}^2(1 - \phi_{I_X}) + 2a^2\sigma_C^2(1 - \phi_C)$ and similarly for $\sigma_{\nabla Y}^2$
Smoothing (eq. 10)	$\rho_{X_S Y_S} = \frac{ab\sigma_C^2(1 + \phi_C)}{2\sigma_{X_S}\sigma_{Y_S}}$	$\sigma_{X_S}^2 = [\sigma_{I_X}^2(1 + \phi_{I_X}) + a^2\sigma_C^2(1 + \phi_C)]/2$ and similarly for $\sigma_{Y_S}^2$