Interpretation of North Pacific Variability as a Short and Long Memory Process

Donald B. Percival[1], James E. Overland[2,3] and Harold O. Mofjeld[2]

[1]Applied Physics Laboratory
Seattle, WA 98195–5640

[2]Pacific Marine Environmental Laboratory/NOAA
Seattle, WA 98115–6349

[3]corresponding author
PMEL/NOAA
7600 Sand Point Way NE
Seattle, WA 98115–6349
overland@pmel.noaa.gov
206.526.6795
Seattle, WA 98115–6349

April 2, 2001

Contribution No. 2249 from NOAA/Pacific Marine Environmental Laboratory

Abstract

A major difficulty in investigating the nature of interdecadal variability of climatic time series is their shortness. An approach to this problem is through comparison of models. In this paper we contrast a first order autoregressive (AR(1)) model with a fractionally differenced (FD) model as applied to the winter averaged sea level pressure time series for the Aleutian low (the North Pacific (NP) index), and the Sitka winter air temperature record. Both models fit the same number of parameters. The AR(1) model is a 'short memory' model in that it has a rapidly decaying autocovariance sequence, whereas an FD model exhibits 'long memory' because its autocovariance sequence decays more slowly.

Statistical tests cannot distinguish the superiority of one model over the other when fit with 100 NP or 146 Sitka data points. The FD model does equally well for short term prediction and has potentially important implications for long term behavior. In particular, the zero crossings of the FD model tend to be further apart, so they have more of a 'regime'-like character; a quarter century interval between zero crossings is four times more likely with the FD than the AR(1) model. The long memory parameter δ for the FD model can be used as a characterization of regime-like behavior. The estimated δs for the NP index (spanning 100 years) and the Sitka time series (168 years) are virtually identical, and their size implies moderate long memory behavior. Although the NP index and the Sitka series have broadband low frequency variability and modest long memory behavior, temporal irregularities in their zero crossings are still prevalent. Comparison of the FD and AR(1) models indicates that regime-like behavior cannot be ruled out for North Pacific processes.

1. Introduction

Difficulties in interpretation of climatic time series center on two issues. The first is recognizing multidecadal variability from relatively short time series of a century or perhaps a little longer. The second is isolating a low frequency signal from a time series with large interannual or several year variability. For example, only 37% of the winter interannual variance of the Aleutian low sea level pressure time series is on time scales greater than five years (Overland *et al.* 1999). There is a current interest in decadal and multidecadal variability of Northern Hemisphere time series. One facet of this interest is that possible climate change will occur as an increase in amplitude or a persistent phase of ongoing large scale atmospheric variability (Palmer 1999). An additional interest is the impact of low frequency variability on ecosystems. Many species such as salmon are adapted to interannual variability, but are strongly modulated by interdecadal changes (Mantua *et al.* 1997). Because of the age structure of marine organisms, Hare and Mantua (2000) in fact conclude that monitoring North Pacific ecosystems might allow for an earlier identification in sign changes than is possible from monitoring climate data alone. It is often difficult to establish a statistically significant difference when comparing different models of climate variability due to the large confidence intervals associated with the relatively short time series. This lack of data puts us on the less certain ground of deciding whether a given model is useful based on additional criteria (von Storch and Zwiers, 1999).

Because of these limitations, it is likely that the true nature of North Pacific time series of climate processes will remain unknown for a long time to come. In such a situation it is important to understand the potential consequences of model choices for interpreting the underlying process. For example, Minobe (1999) suggests bidecadal and pentadecadal oscillations for the Aleutian low sea level pressure time series. A second concept is regimes where objective change point techniques suggest significantly different sections in the time series, for example, before and after 1977 (Overland *et al.* 1999; Hare and Mantua 2000). A chaotic model suggests rapid transitions, but eventual return to the vicinity of previous locations in state space (Overland *et al.* 2000). A

purely stochastic white and red noise view is suggested by Pierce (2001) for the North Pacific and Wunsch (1999) for the North Atlantic. The assumption of a white noise model is that there are instabilities in the atmosphere and that energy is cascading to all frequencies in equal amounts. As Wunsch and Pierce point out, there is by definition considerable low frequency energy in a white noise process. What is of interest in interpreting climate time series is whether an enhanced low frequency variance exists. This enhancement might suggest a physical feedback process, such as air-ocean interaction, which would permit some measure, however small, of enhanced predictability.

This paper investigates the influence of model choice in representing North Pacific atmospheric processes. We use the wintertime average (Nov–Mar) Aleutian low sea level pressure field as the basic time series (Trenberth and Paolino 1980). This time series is referred to as the North Pacific (NP) index. The NP index is a surface index associated with the Pacific North American (PNA) pattern of hemispheric variability in the troposphere. The NP index also serves as a measure of atmospheric forcing of North Pacific Ocean variability. We investigate whether the NP index has characteristics of a long memory process, which is a broadband process that exhibits a persistent dependence between distant observations (Beran 1994). A default stochastic model suggested by several authors (von Storch and Zwiers 1999) is the autoregressive moving average (ARMA) process. We contrast a simple ARMA model, namely, an first order autoregressive (AR(1)) model with a simple long memory model, namely, a fractionally differenced (FD) process. Both are fit to the NP index. By inspecting various diagnostic statistics, we conclude that the FD model for the North Pacific is an equally as viable model of North Pacific variability as the AR(1) model and that, given the amount of available data, we cannot expect to be able to distinguish between the two models. While the two models have similar behavior for short term prediction, the FD model suggests different zero crossing behavior. We then compare our analysis of the NP index (presented in §2 below) with an analysis of a somewhat longer North Pacific time series, namely, a temperature record from Sitka, Alaska (§3). We discuss the implications of the two stochastic models applied to the North Pacific in §4 in terms of run lengths, which could serve as one definition for regime-like behavior. We state our conclusions in §5.

2. Statistical Models for the NP index

In this section we consider two stationary models for the NP index. This time series consists of $N = 100$ annual values for the years 1900 to 1999 (Figure 1). The two models are a first order autoregressive (AR(1)) process and a fractionally differenced (FD) process, both of which are completely determined by three parameters and hence can be regarded as being equally simple. For each process, one of the parameters is its expected value, and another effectively controls the shape of the spectral density function (SDF) and the corresponding autocovariance sequence (ACVS). The final parameter merely adjusts the levels (heights) of the SDF and ACVS. The essential difference between these two models is that an AR(1) process postulates a rapidly decaying ACVS whereas the ACVS for an FD process decays much more slowly. This qualitative difference is what is meant in the literature when an AR(1) process is said to have 'short memory' whereas a comparable FD model is said to have 'long memory.'

In what follows, we first define each model and outline a procedure for estimating the model parameters from the NP index, from which we learn that the autocorrelation in this series is fairly weak overall. We then consider some goodness of fit tests (model diagnostics) for both short and long memory models, from which we conclude that both models are quite reasonable for the NP index and that, from a statistical point of view, each model fits equally well. Next we explore how well we can expect to discriminate between short and long memory models, given a time series with the length and characteristics of the NP index. We conclude that, since the overall autocorrelation in the NP index is weak, we would need much more than $N = 100$ years of data to be able to reject a short memory model if in fact the NP index were a realization of an FD process (and *vice versa*).

a. A Short Memory Model for the NP index

Let us consider a short memory model first by regarding the NP index as a realization of a portion $X_0, X_1, \ldots, X_{N-1}$ of a stationary Gaussian AR(1) process; i.e., we assume

$$X_t - \mu_X = \phi(X_{t-1} - \mu_X) + \epsilon_t, \tag{1}$$

where $\mu_X \equiv E\{X_t\}$ and $|\phi| < 1$, while $\{\epsilon_t\}$ is a Gaussian white noise process with mean zero and variance σ_ϵ^2 (this parameter is sometimes called the 'innovations' or 'prediction error' variance and can be interpreted as the mean squared difference between X_t and the best linear predictor of X_t based upon prior values $X_{t-1}, X_{t-2}, \ldots$). This process has three parameters, namely, μ_X, ϕ and σ_ϵ^2. The last two parameters determine the correlation properties of the AR(1) process because these fully specify both its SDF $S_X(\cdot)$ and its ACVS $\{s_{X,\tau}\}$:

$$S_X(f) \equiv \frac{\sigma_\epsilon^2}{\left|1 - \phi e^{-i2\pi f}\right|^2}, \quad |f| \le \tfrac{1}{2}, \text{ and } s_{X,\tau} = \frac{\sigma_\epsilon^2 \phi^{|\tau|}}{1-\phi^2}, \quad \tau = \ldots, -1, 0, 1, \ldots . \tag{2}$$

When $\phi = 0$, the AR(1) process becomes a white noise process. We can thus assess the null hypothesis that the NP index is realization of a Gaussian white noise process by estimating ϕ and then ascertaining whether or not the estimated value is significantly different from zero.

In order to fit an AR(1) model to the NP index, we must estimate the three unknown parameters. We adopt the following standard strategy. First we estimate μ_X using the sample mean $\overline{X} \equiv \frac{1}{N}\sum X_t$, after which we recenter the series by forming $\widetilde{X}_t \equiv X_t - \overline{X}$. We then use the maximum likelihood (ML) method to estimate the parameters ϕ and σ_ϵ^2 in the AR(1) model $\widetilde{X}_t = \phi \widetilde{X}_{t-1} + \epsilon_t$, i.e., Equation (1) with $\widetilde{X}_t$ replacing X_t and with μ_X set to zero. (A possible refinement is to use the ML approach to estimate μ_X along with ϕ and σ_ϵ^2, but this more complicated procedure leads to very little gain in the quality of the estimator for μ_X – see §8.2 of Beran 1994.) For completeness, details on the formulation of the ML estimators $\hat{\phi}$ and $\hat{\sigma}_\epsilon^2$ for the parameters ϕ and σ_ϵ^2 are given in Appendix A. The theory behind these estimators says that asymptotically they are (i) independent of each other, (ii) unbiased and (iii) normally distributed with $\mathrm{var}\,\{\hat{\phi}\} = (1-\phi^2)/N$ and $\mathrm{var}\,\{\hat{\sigma}_\epsilon^2\} = 2\sigma_\epsilon^4/N$. Approximate 95% confidence intervals (CIs) for ϕ and for σ_ϵ can thus be constructed using, respectively,

$$\left[\hat{\phi} - 1.96\frac{(1-\hat{\phi}^2)^{1/2}}{N^{1/2}}, \hat{\phi} + 1.96\frac{(1-\hat{\phi}^2)^{1/2}}{N^{1/2}}\right] \text{ and } \left[\left(\hat{\sigma}_\epsilon^2 - 1.96\frac{\hat{\sigma}_\epsilon^2\sqrt{2}}{\sqrt{N}}\right)^{1/2}, \left(\hat{\sigma}_\epsilon^2 + 1.96\frac{\hat{\sigma}_\epsilon^2\sqrt{2}}{\sqrt{N}}\right)^{1/2}\right]. \tag{3}$$

A set of residuals (sometimes called observed innovations or observed prediction errors) that can

be examined to evaluate the adequacy of the model is given by

$$\hat{\epsilon}_0 \equiv \widetilde{X}_0(1-\hat{\phi}^2)^{1/2} \text{ and } \hat{\epsilon}_t \equiv \widetilde{X}_t - \hat{\phi}\widetilde{X}_{t-1}, \quad t = 1, \ldots, N-1. \tag{4}$$

Application of the ML procedure to the NP index yields the estimates of ϕ and σ_ϵ and associated 95% CIs given in the upper third of Table 1. Note that the interval for ϕ just barely misses zero, so there is evidence that the true ϕ differs from zero at an observed level of significance of approximately 0.05. Since this result depends on the large sample approximation for the variance of $\hat{\phi}$, we checked its applicability by generating 10,000 simulated series of length $N = 100$ from an AR process with parameter $\phi = 0.21$ (i.e., the observed $\hat{\phi}$) and fitting an AR model to each series (details on how the simulated series were created are given in Kay 1981). This Monte Carlo experiment yielded an estimate for var $\{\hat{\phi}\}$ that, when used in place of the large sample approximation, yielded a 95% CI for ϕ in perfect agreement with the one stated in Table 1. The upper left-hand plot of Figure 2 shows the theoretical autocorrelation sequence (ACS) $\rho_{X,\tau} \equiv s_{X,\tau}/s_{X,0}$ for an AR(1) process with parameter $\hat{\phi}$ as a solid curve, along with the first part of the sample ACS $\hat{\rho}_\tau$ for the NP index (plotted as deviations from zero), where

$$\hat{\rho}_\tau \equiv \frac{\sum_{t=0}^{N-\tau-1} \widetilde{X}_t \widetilde{X}_{t+\tau}}{\sum_{t=0}^{N-1} \widetilde{X}_t^2}, \quad \tau = 0, 1, \ldots . \tag{5}$$

Also shown are upper and lower 95% CIs (thin curves) for the ACS under the assumption that the NP index is a realization of a white noise process (see Corollary 6.3.6.2 of Fuller 1996 for details). The lower left-hand plot shows an SDF estimate (thick smooth curve) obtained by substituting $\hat{\phi}$ and $\hat{\sigma}_\epsilon^2$ for the corresponding quantities in Equation (2), along with the periodogram $\hat{S}(f_k)$ for the NP index (thin jagged curve); i.e.,

$$\hat{S}(f_k) \equiv \frac{1}{N}\left|\sum_{t=0}^{N-1} \widetilde{X}_t e^{-i2\pi f_k t}\right|^2 \tag{6}$$

where $f_k \equiv k/N$, $1 \le k < N/2$. In the left-hand part of this plot is a confidence interval about a circle. If we move this interval such that the center of the circle is positioned at a particular $\hat{S}(f_k)$, then we have a 95% confidence interval for the true SDF at frequency f_k (this interval is based upon the standard assumption that $\hat{S}(f_k)$ is proportional to a chi-square random variable with two

degrees of freedom). Such an interval in fact traps the theoretical AR(1) SDF at all but two of the fifty Fourier frequencies, indicating that there are no serious discrepancies between the NP index and the fitted AR(1) model.

b. A Long Memory Model for NP

Suppose now that the NP data is a realization of a portion $Y_0, Y_1, \ldots, Y_{N-1}$ of a stationary Gaussian fractionally differenced (FD) process. By definition such a process has a first moment given by $\mu_Y \equiv E\{Y_t\}$ and an SDF and ACVF given by

$$S_Y(f) = \frac{\sigma_\varepsilon^2}{|2\sin(\pi f)|^{2\delta}}, \quad |f| \le \tfrac{1}{2}, \text{ and } s_{Y,\tau} = \frac{\sigma_\varepsilon^2 \sin(\pi\delta)\Gamma(1-2\delta)\Gamma(\tau+\delta)}{\pi\Gamma(\tau+1-\delta)}, \tag{7}$$

where $\sigma_\varepsilon^2 > 0$ and $-\frac{1}{2} \le \delta < \frac{1}{2}$ (Granger and Joyeux 1980; Hosking 1981). The parameter σ_ε^2 can be interpreted as an innovations variance, while δ causes an FD process to exhibit long memory when $0 < \delta < \frac{1}{2}$. At low frequencies, we have $S_Y(f) \approx \sigma_\varepsilon^2/|2\pi f|^{2\delta}$, so an FD process is approximately proportional to a power law $|f|^\alpha$ with exponent $\alpha = -2\delta$. It follows from Equation (7) and standard relationships for the Γ function (see, e.g., Abramowitz and Stegun 1964) that the variance of this process is given by

$$s_{Y,0} = \frac{\sigma_\varepsilon^2 \Gamma(1-2\delta)}{\Gamma^2(1-\delta)} \tag{8}$$

and that, for $\tau \ge 1$, the ACVF can be computed recursively using the formula

$$s_{Y,\tau} = s_{Y,\tau-1}\frac{\tau+\delta-1}{\tau-\delta}.$$

When $\delta = 0$, the FD process becomes a white noise process, so we can assess the null hypothesis of white noise for the NP index based upon an estimate of δ.

The unknown parameters in an FD model are μ_Y, δ and σ_ε^2. We adopt a strategy similar to what we used in the AR(1) case by first recentering the series to form $\widetilde{Y}_t \equiv Y_t - \overline{Y}$, after which we use the ML method to estimate δ and σ_ε^2 via $\hat{\delta}$ and $\hat{\sigma}_\varepsilon^2$ as described in Appendix A. The theory behind $\hat{\delta}$ and $\hat{\sigma}_\varepsilon^2$ says that these estimators are asymptotically independent of each other, unbiased

and normally distributed with $\mathrm{var}\{\hat{\delta}\} = 6/(\pi^2 N)$ and $\mathrm{var}\{\hat{\sigma}_\varepsilon^2\} = 2\sigma_\varepsilon^4/N$. Approximate 95% CIs for δ and σ_ε are given by

$$\left[\hat{\delta} - 1.96\frac{\sqrt{6}}{\pi\sqrt{N}}, \hat{\delta} + 1.96\frac{\sqrt{6}}{\pi\sqrt{N}}\right] \quad \text{and} \quad \left[\left(\hat{\sigma}_\varepsilon^2 - 1.96\frac{\hat{\sigma}_\varepsilon^2\sqrt{2}}{\sqrt{N}}\right)^{1/2}, \left(\hat{\sigma}_\varepsilon^2 + 1.96\frac{\hat{\sigma}_\varepsilon^2\sqrt{2}}{\sqrt{N}}\right)^{1/2}\right]. \qquad (9)$$

As is true when fitting an AR(1) model, the ML procedure leads to a set of residuals $\hat{\varepsilon}_t$ that we can use to evaluate the adequacy of the fitted FD model (see Appendix A for details).

Application of the ML procedure to the NP index yields the estimates and 95% CIs shown in Table 1. As was true for the AR(1) parameter ϕ, the interval for δ just barely misses zero, so evidently the true δ differs from zero at an observed level of significance of approximately 0.05. To ascertain the validity of this CI, we carried out a Monte Carlo experiment analogous to the one for the AR(1) case and obtained a CI of $[0.01, 0.33]$ for δ, which is very close to the interval $[0.02, 0.32]$ reported in Table 1 (the simulated FD series were created using an 'exact' method described in Davies and Harte 1987 and Wood and Chan 1994). The upper right-hand plot of Figure 2 shows the theoretical ACS for an FD process with parameter $\hat{\delta}$ (thick curve). The lower right-hand plot shows an SDF estimate (thick smooth curve) obtained by plugging $\hat{\delta}$ and $\hat{\sigma}_\varepsilon^2$ into the corresponding quantities in Equation (7). If we compare this estimate to 95% confidence intervals for the true SDF based upon the periodogram, we find the confidence intervals trap the FD-based estimates at all but one of the fifty Fourier frequencies, indicating that there are no serious discrepancies between the NP index and the fitted FD model.

c. Goodness of Fit Tests for Short and Long Memory Models

Figure 2 indicates that, when we take their sampling variability into consideration, the sample ACS and the periodogram for the NP index are visually in reasonably good agreement with the corresponding theoretical quantities derived from the fitted AR(1) and FD models. A more quantitative approach for assessing the adequacy of these models is to consider four well-known goodness of fit test statistics. The results of one or more of these tests could in principle lead us to favor one model over the other. The first test statistic T_1 is an SDF test that compares the periodogram for the NP

index to the SDF corresponding to the fitted model. The remaining three test statistics make use of the residuals $\hat{\epsilon}_t$ and $\hat{\varepsilon}_t$ obtained in the process of fitting the AR(1) and FD models. These tests are built using the concept that, if the proposed model is in fact correct, the residuals should be approximately a realization of a white noise process. The first such statistic T_2 is the cumulative periodogram test, while the remaining test statistics T_3 and T_4 are variations on the portmanteau test, which looks at the squares of a small number of sample autocorrelations of the residuals. Details about T_j, $j = 1, \ldots, 4$, are given in Appendix B, but the manner in which we use each test statistic is quite similar. Thus, based upon a computed T_j and a predetermined significance level α, we can reject the null hypothesis that the NP index is a realization from one of the fitted models if T_j exceeds the $(1-\alpha) \times 100\%$ percentage point $Q_j(1-\alpha)$ for the statistic under the null hypothesis. If, for example, we let $\alpha = 0.05$, we would incorrectly reject the null hypothesis about 5% of the occasions when in fact it is true. Alternatively, if we do not want to use a prespecified significance level, we can compute the observed critical level $\hat{\alpha}$, which is the smallest significance level for which we would end up rejecting the null hypothesis. Thus, if $\hat{\alpha}$ is quite small, we have good reason to doubt the validity of the null hypothesis; on the other hand, if $\hat{\alpha}$ is large compared to typical preselected values for α (e.g., 0.05 or 0.01), we have no real reason to reject the null hypothesis.

Table 2 gives the results of applying the four goodness of fit tests to the AR(1) and FD models for the NP index (in keeping with recommendations in the literature, we set $K = N/20 = 5$ when computing and evaluating T_3 and T_4, but we obtained comparable results with $K = 10$). For the sake of comparison, we also used each test statistic on the NP index itself; i.e., we entertained the null hypothesis that the NP index is a realization of a white noise (WN) process. None of the four tests rejects the null hypothesis at the 0.05 level of significance for either the AR(1) or FD models, and all reject the null hypothesis of white noise for the NP index itself. The observed critical levels $\hat{\alpha}$ are larger for the FD model for all four T_j, which might suggest that the FD model is slightly better than the AR(1) model; however, at best this is quite weak evidence. The main conclusion we can draw from these tests is that the short and long memory models are quite comparable and

that both models are to be prefered over a simple white noise model.

d. Discriminating Between Short and Long Memory Models

The goodness of fit tests indicate that both models seem to fit equally well and that there is no compelling statistical evidence that would favor either the short or long memory model for the NP index. Here we explore the question of how well we can expect to discriminant between a short and long memory process given a realization as short as the NP index. To do so, let us first assume that the NP index is in fact a realization of an FD process with model parameters given by the estimates shown in Table 1. Given this assumption, we could use the machinery outlined in Appendix A to convert the NP index into a set of residuals that would in fact be a realization of a white noise process. Suppose, however, that we incorrectly entertain an AR(1) model with parameter $\hat{\phi}$ as given in Table 1 to obtain a set of residuals via Equation (4). These residuals would be approximately a realization of a stationary process, say V_t, whose SDF is given by

$$S_V(f) = \hat{\sigma}_\varepsilon^2 \frac{|1 - \hat{\phi} e^{-i2\pi f}|^2}{|2\sin(\pi f)|^{2\hat{\delta}}}. \tag{10}$$

The creation of V_t amounts to subjecting an FD process to a prewhitening filter in which the filter is in fact appropriate for an AR(1) process. If the fitted AR(1) and FD models are in fact comparable over a range of frequencies whose lower limit is approximately equal to the inverse of the total time span of the available data, we might expect a goodness of fit statistics to be unable to distinguish between V_t and a white noise process, but, by making N sufficiently, we can expect to make the distinction. Conversely, we can swap the roles of the AR(1) and FD processes in this exercise, leading us to a set of residuals that are a realization of a process, say W_t, with SDF given by

$$S_W(f) = \hat{\sigma}_\epsilon^2 \frac{|2\sin(\pi f)|^{2\hat{\delta}}}{|1 - \hat{\phi} e^{-i2\pi f}|^2}. \tag{11}$$

Andersson (1998) has previously studied the implications of missmatching short and long memory models in the context of forecasting economic time series.

To determine how large N must be before we can reasonably expect to reject the null hypothesis that a fitted model is adequate when in fact the time series is a realization of a different process,

we conducted a series of Monte Carlo experiments that yielded an estimate of the probablilty of rejecting the null hypothesis for sample sizes ranging from $N = 400$ up to $N = 6000$ (see Appendix C for details). Figure 3 shows plots of the probability of rejection as a function of sample size under the two scenarios. The left-hand plot in this figure is for the case when we fit an AR(1) model to a realization of an FD process with parameters $\hat{\delta}$ and $\hat{\sigma}_{\varepsilon}^2$. This plot indicates that, in order to have a 50% chance of rejecting the null hypothesis , we would require sample sizes of about $N = 2500, 1700, 500$ and 500 when using, respectively, the T_1, T_2, T_3 and T_4 test statistics. The right-hand plot shows that, when the roles of the AR(1) and FD processes are swapped, we would now need $N = 750$ when using the T_2, T_3 and T_4 test statistics and an N in excess of 4000 for T_1. All of these sample sizes are considerably larger than the $N = 100$ values that make up the NP index, thus reinforcing the notion that, given the weak overall correlation that is exhibited by the NP index and given the amount of data that is available to us, we cannot hope to distinguish between short and long memory models.

3. Statistical Models for Sitka Air Temperature

Let us now consider the same two statistical models for the Sitka winter air temperature time series (Figure 4). This time series consists of 146 data values collected over a 168 period (1829 to 1996), so there are 22 years for which there are no recorded values. Sitka lies in the eastern Gulf of Alaska. Winter temperature anomalies relate to changes in the wind field with more southerly winds producing warm anomalies. These winds would respond in part to both the intensity and east/west location of the Aleutian Low. For comparison, we fit AR(1) and FD models both to the original unequally sampled series and to an equally sampled version of the Sitka series formed by linearly interpolating values for the missing years. While the interpolated series can be handled using exactly the same ML estimation procedures as in the case of the NP index, the unequally sampled Sitka series requires an adaptation of these procedures that can deal with missing values (see Appendix A for details). The resulting estimates and corresponding 95% confidence intervals for the uninterpolated and interpolated series are displayed in, respectively, the middle and bottom

thirds of Table 1 (for the uninterpolated series, we used Monte Carlo experiments to verify that the CIs based upon Equations (3) and (9) with $N = 146$ are indeed accurate). The estimated ϕ and δ parameters for the uninterpolated series are quite comparable to the estimates for the NP index. The corresponding parameter estimates for the interpolated series are somewhat higher, suggesting that a slightly stronger degree of autocorrelation has been artificially introduced by the interpolation procedure. As was done in Figure 2 for the NP index, Figure 5 shows the sample ACS and periodogram for the interpolated series, along with the theoretical ACSs and SDFs corresponding to the AR and FD processes that were fit to the uninterpolated series. Based upon these plots and the goodness of fit tests, we can conclude, as for the NP index, that the AR(1) and FD models are quite comparable for the Sitka series and that there is no statistical evidence to favor one model over the other.

4. Discussion

a. Implications of Short versus Long Memory Models

Based upon the previous sections, there is no statistical reason to prefer an AR(1) process over an FD process as a model for the NP and Sitka series (or *vice versa*). Both processes depend upon three parameters, so both have the same degree of simplicity. We cannot thus appeal to the principle of Occram's razor here to make a case for one process over the other. Nonetheless, the fact that the two processes appear to describe both series equally well does not mean that there are not potentially important implications if we arbitrarily select one of these processes to model certain statistical properties of these series. As an illustration of this fact, here we consider the extent to which the two processes lend support to the notion of 'regimes' in the NP index.

Loosely speaking, a regime is an interval of time during which a time series remains predominantly either above or below its long term average value. To clarify this idea, let us consider Figure 1, which shows the NP index (thin curve) along with a five year running average of the index (thick curve) and a horizontal line indicating the sample mean of the entire series (1009.8).

In the NP index itself there appear to be intervals over which the index is predominantly above its sample mean. For example, from 1901 to 1923, all of the NP values were above the sample mean with the exception of the ones for 1905 and 1919. This stretch of 23 years would constitute a positive regime and is clearly identified in the five year running average (see also Minobe 1999). The idea behind the running average is to quantify the notion of 'predominantly above,' but, while the choice of five years is admittedly subjective, it is in keeping with smoothing procedures typically applied to climatological time series in the literature to reduce the influence of interannual variability. After 1923, we can see that the running averages are predominantly (but not strictly) below the sample mean up to 1946. Based upon this visual inspection, we might be tempted to deem this 23 year interval to be a negative regime (and to formulate a hypothesis that a 'typical' regime lasts 23 years), but this is obviously a subjective judgement that is open to valid criticism. If we take the definition of a regime to be a contiguous stretch over which a five year running average is strictly above or below the sample mean, then the period from 1924 to 1946 breaks up into seven regimes (four of length one, and one each of lengths three, seven and nine). If in fact climatological series such as the NP index were to exhibit regimes with 'typical' sizes, we could presumably use this information to help predict when a switch from, say, a positive to a negative regime is about to occur.

While the predictability of regime shifts for the NP index is open to question when we view the series as a realization of either a short or long memory process, it is nonetheless of interest to see how the fitted FD and AR(1) models impact what we would deduce about the distribution of regime sizes. Knowledge of this distribution gives us some idea as to how compatible these two processes are with the idea of regimes, or at least tell us which process is more likely to generate realizations that supporters of the regime idea would deem to be realistic. While it is difficult to determine the distribution of regime sizes analytically, it is easy to do so via Monte Carlo experiments. To do so, we generated 1000 realizations of size 1024 from a zero mean FD process whose parameter δ is dictated by our fitted FD model for the NP index. In order to account for the uncertainity in the parameter estimate $\hat{\delta}$, we used its large sample distribution to generate the FD parameter

δ_k for the kth such realization; i.e., we selected δ_k from a Gaussian distribution with mean $\hat{\delta}$ and variance $6/(100\pi^2)$. We need not concern ourselves with uncertainty in the estimate for σ_ε^2 because this parameter has no effect on how much time a realization spends above or below the process mean. For each realization so generated, we then determined how many regimes there were of sizes unity and up. We used the following two definitions for a regime. Let $x_t, t = 0, \ldots, 1023$, denote one of the realizations. In the first definition, the starting index t_s for a positive regime is one for which $x_{t_s} \geq 0$ and $x_{t_s-1} < 0$, while the ending index t_e is the one satisfying $t_e \geq t_s$, $x_{t_e} \geq 0$ and $x_{t_e+1} < 0$ with $x_{t_s+k} \geq 0$ for $k = 0, \ldots, t_e - t_s$ (a negative regime is defined analogously by $x_{t_s} < 0$ and $x_{t_s-1} \geq 0$ along with $x_{t_e} < 0$ and $x_{t_e+1} \geq 0$). In the second definition, we define $y_t = (x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2})/5$ and then identify regimes by applying the first definition to this five point running average. Here we wish to suppress the influence of the large interannual variability. For either definition, the regime size is taken to be $t_e - t_s + 1$. We also only tabulated "fully expressed" regimes; i.e., regimes that might have started prior to index $t = 0$ or after index $t = 1023$ were not used. This procedure, in fact, biases the distribution somewhat toward smaller regime sizes, but this bias should be relatively small and can be lessened by increasing the size of each realization beyond 1024. An analogous procedure was followed for the AR(1) model.

Figure 6 summarizes the results of the Monte Carlo experiments. Here we plot the empirically determined probability of a regime size being greater than or equal to a specified length for the AR(1) model (thin curves) or FD model (thick) when used with the first or second definition for a regime (left- and right-hand plots, respectively). We see that the FD model tends to yield regime sizes that are longer than those for the AR(1) model. For example, if we consider the 'typical' regime size of 23 years suggested by our visual inspection of the first half of the NP index, we see that a run of this length or longer is four times more likely to occur in the five year running averages with the FD model than with the AR(1) model. As a second example, a run of 35 years is ten times more likely to occur with the FD model. Thus, even though the statistics for short runs are quite comparable in both models, the FD model suggests a greater likelihood of observing long runs, which is in keeping with visual analyses that inspired the notion of regimes.

In the previous sections we have noted that the FD model is as viable a model for the NP and Sitka series as the AR(1) model. In this section one notion for regime-like behavior is presented in which extended intervals between zero crossings are shown to be consistent with the FD model, but it is not well supported by the AR(1) model. Also, since the FD model is also stochastic, regime-like behavior based on zero crossings does not necessarily require a deterministic oscillation model. The fact that FD models are more supportive of regime-like behavior allows us to make practical discrimination between AR(1) and FD models based upon auxiliary information. For example, the evidence of regimes in several biological systems in the North Pacific is strong, particularly for salmon (Mantua *et al.* 1997). Our FD model is consistent with the quite reasonable hypothesis that the physical environment in the North Pacific is a contributing factor to the regimes observed in these biological systems. Auxiliary information from the physical system can also be used to lend support for the FD model over the AR(1) model. The maximum likelihood analysis of Haines and Hannachi (1995) suggests that the PNA pattern, and thus the NP index, have preferred bimodal states. Another plausible mechanism for persistence or a long memory effect is ocean atmospheric feedback in the North Pacific (Latif and Barnett 1994). Feldstein (2000) suggests that interannual variability of the PNA pattern arises both from climate noise and external forcing, which might be consistent with the level of persistence suggested by the FD model.

b. Interpretation and Adequacy of Long Memory Models

From Table 1 we see that the estimated values of the FD parameter δ for both the NP and Sitka series are around 0.2. The allowable range of δ for stationary FD models with long memory dependence is $0 < \delta < 0.5$. As δ approaches zero, an FD process approaches white noise, which has 'no memory' in the sense that its random variables are pairwise uncorrelated. At the other extreme, as δ approaches a half, realizations from the FD process exhibit a strong long memory effect. To get a better idea of how to interpret δ, Figure 7 shows two columns of simulated FD series (thin curves), each with four rows. Each row corresponds to a different choice of δ. From top to bottom, these are $\delta = 0.02$ (the lower end of the 95% CI for δ for the NP index, which is quite close to

white noise), 0.17 (the estimated value for the NP index), 0.32 (the upper end of the 95% CI, which corresponds to a moderate long memory effect) and 0.45 (a value in the upper allowable range for δ corresponding to a strong long memory effect). All four processes have zero mean and unit innovations variance, so they only differ in the choice of δ. All four time series in a given column were formed using the *same* realization of white noise so that differences amongst these series can be attributed entirely to δ (i.e., we used the exact simulation method described in Davies and Harte (1987) and Wood and Chan (1994), which works by transforming $2N$ white noise deviates into a correlated series of length N – we have just used the same white noise sequence to form the four simulated FD series). Note that, as the degree of the long memory effect increases, we see a more regime-like structure in the series; i.e., there is a greater tendancy for the series to be above (or below) the process mean of zero for long stretches of time. To quantify this, let us consider five point running averages (the thick curves on each plot). As δ increases, the total number of runs in the five point running averages tends to decrease (24, 16, 16 and 12 in, respectively, top to bottom left-hand plots; 20, 15, 13 and 11 in right-hand plots), while the length of the longest run increases (12, 17, 18 and 23, left-hand plots; 11, 21, 24 and 25, right-hand plots). The results are not linear in δ. In fact, the best estimate of $\delta = 0.17$ for the NP index has similar behavior to $\delta = 0.32$, and is substantially different from the white noise model.

We can thus interpret δ as an indicator of how much regime-like structure there is in a time series: if δ is close to zero, there is very little tendency for the series to remain above its process mean for long stretches of time, whereas the opposite is true when δ is close to a half. If we consider that the time series has both short term (interannual) variability and long term memory, then even the modest value of $\delta = 0.17$ is enough to change the zero crossing behavior to provide a regime-like behavior in the five year running means. Thus, the estimated δ parameters for both the NP and Sitka series are significantly different from zero (i.e., they cannot be reasonably taken to be a realization of a white noise process), and the size of δ suggests there is moderate long memory structure, based on run statistics. The appeal of FD models is that they have a single parameter (δ) that can help us understand if a particular climatological series exhibits weak, moderate or strong

long memory characteristics. For the NP and Sitka series, we can conclude from the estimated δ that there are structures that are compatible with the notion of regimes, but both series cannot be characterized as being dominated by a single strong long memory process. In addition, by inspection of Figures 1, 4, and 7, there is a broad distribution of zero crossing intervals. If the FD model were the true underlying process for the NP index, then even though regimes are a major feature, prediction would be problematic.

5. Conclusions

We have compared a first order autoregressive (AR(1)) model with a fractionally differenced (FD) model applied to two North Pacific (NP) time series, the winter NP sea level pressure index which is centered on the Aleutian low region, and the winter average of the monthly termperature records from Sitka, Alaska. Both models reduce to white noise when one of their model parameters is zero. For both time series, this parameter is (just barely) significantly different from zero at a 95% level of confidence, and hence there is evidence to say that both time series have significant serial correlation. The AR(1) model has a rapid drop off of the autocovariance sequence, which essentially models the large interannual variability of the time series. The autocovariance sequence for the FD model has a similar drop for short lags, but also has a long tail of small but positive correlations at longer lags, which is termed 'long memory' in the statistical literature. The statistical analysis of the winter averaged NP index shows that the AR(1) and FD models fit equally well. A similar analysis using the longer Sitka air temperature series corroborates this result. Like the AR(1) model, fitting the FD model to a given time series involves the estimation of just three parameters; hence the models are equally parsimonious. The FD model has the additional property that, unlike the AR(1) model, it creates regime-like behavior in which the winter averaged NP index tends to remain above or below the mean for a number of years. This is true even when the low frequency variance is a relatively small percentage of the total, as it is for the NP and Sitka series.

To fit the AR(1) and FD models to climatic series, we have adopted a rigorous statistical approach appropriate for the problem at hand. This approach includes maximum likelihood estimation

of model parameters (adapted, in the case of the Sitka series, to handle missing values in a time series without the need for a questionable interpolation scheme); use of Monte Carlo experiments to verify large sample approximations to the variance of the estimated parameters; use of goodness of fit test statistics to evaluate the fitted models; and an evaluation of the performance of these test statistics in the presence of incorrect models. This approach should prove useful to investigators who wish to examine other climatic data sets from a short versus long memory perspective.

Based on synthetic time series derived from both the AR(1) and FD models, we show that it would take a time series of several hundred years to discriminate between the two models as being the underlying process for the North Pacific. In such a situation with relatively short time series and large interannual variability, we are left with the less attractive option of comparing models rather than claiming that one model is statistically more appropriate than another. Hence, in modeling climate variability in the North Pacific, it is necessary to rely on model-to-model and time series to time series comparisons, and to bring in additional information outside the time series in order to choose between models. For example, two distinct North Pacific time series, the NP index and the Sitka air temperatures, both have a fitted FD model with nearly the same parameter value, $\delta = 0.17$ and $\delta = 0.18$. Physical arguments can also be brought in as additional information. For example, in the North Pacific there are atmosphere-ocean models that suggest feedbacks on decadal scales. The PNA teleconnection pattern has been shown to have bimodal behaviors. Perhaps the strongest evidence for a FD model over an AR(1) model is from biological time series (Hare and Mantua, 2000). Well established regime behavior seen in the biology of the region, such as geographic changes in salmon populations, support evidence for shifts in the physical system near 1925, 1947, and 1977. However, the strongest statement that we can make from our analysis is that regime-like behavior for the North Pacific, based on the long memory model, cannot be ruled out based on statistical grounds.

One important point that our work draws attention to regards the interpretation of δ, the parameter in the FD model that determines its long memory characteristics. This parameter varies from $\delta = 0.0$ for white noise to just below 0.5 for a strong long memory effect. However, if we

consider regime-like behavior based on interval statistics for zero crossings, then the behavior of δ is nonlinear. The value of the parameter for the two North Pacific time series is $\delta \approx 0.17$, yet its behavior in terms of run lengths was similar to $\delta = 0.32$ and both of these values had behavior closer to $\delta = 0.45$ than $\delta = 0.02$. Apparently, the small displacement contributions from long periods is enough so that weak excursions at interannual scales do not cross the zero level. The δ parameter is a measure of the tendency to form regimes. Because the FD model is completely stochastic, it cannot be used to make deterministic predictions for the beginning and duration of regimes. Our results show, however, that North Pacific time series are consistent with moderate regime-like behavior, based on the FD model. The results of our comparison of the FD model with the AR(1) model leave room for further characterization and potential prediction of North Pacific climate processes.

Appendix A: Maximum Likelihood Estimation

Suppose that $\mathbf{U} \equiv [U_0, U_1, \ldots, U_{N-1}]^T$ is a vector of random variables that form a portion of a real-valued Gaussian stationary process with zero mean and ACVS $\{s_{U,\tau} : \tau = \ldots, -1, 0, 1, \ldots\}$. Let Σ be the covariance matrix for $\mathbf{U}$; i.e., the (j,k)th element of Σ is given by $s_{U,j-k}$, where $0 \le j, k \le N-1$. The joint probability density function for these RVs can be written as

$$f(\mathbf{U}) \equiv \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} e^{-\mathbf{U}^T\Sigma^{-1}\mathbf{U}/2}, \tag{12}$$

where $|\Sigma|$ and Σ^{-1} are, respectively, the determinant and inverse of Σ. Suppose now that $\{s_{U,\tau}\}$ and hence Σ are completely determined by a vector $\mathbf{a}$ of K unknown parameters, where typically $K \ll N$. Given $\mathbf{U}$, we can regard the right-hand side of Equation (12) as an implicit function of $\mathbf{a}$ known as the likelihood function:

$$L(\mathbf{a} \mid \mathbf{U}) \equiv \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} e^{-\mathbf{U}^T\Sigma^{-1}\mathbf{U}/2}.$$

The maximum likelihood (ML) estimator $\hat{\mathbf{a}}$ for $\mathbf{a}$ is the vector that maximizes $L(\mathbf{a} \mid \mathbf{U})$ as a function of $\mathbf{a}$; equivalently, the ML estimator is the vector that minimizes

$$l(\mathbf{a} \mid \mathbf{U}) \equiv -2\log\left(L(\mathbf{a} \mid \mathbf{U})\right) - N\,\log(2\pi) = \log(|\Sigma|) + \mathbf{U}^T\Sigma^{-1}\mathbf{U}.$$

When dealing with a time series with missing values (e.g., the Sitka series), we can reformulate the above by letting $\mathbf{U}$ just contain the random variables corresponding to the actual observations and by deleting all rows and columns Σ corresponding to the missing values. The ML estimators satisfy a number of optimality criteria and hence are generally to be preferred over other estimators, particularly when dealing with small sample sizes (see, e.g., §5.2 of Priestley 1981).

a. MLEs for an AR(1) Process

In the case of an AR(1) process, we take $\mathbf{U}$ to be $[\widetilde{X}_0, \widetilde{X}_1, \ldots, \widetilde{X}_{N-1}]^T$, where the recentered time series $\{\widetilde{X}_t\}$ is assumed to obey the model $\widetilde{X}_t = \phi\widetilde{X}_{t-1} + \epsilon_t$. The process $\{\epsilon_t\}$ is taken to be Gaussian white noise with mean zero and variance σ_ϵ^2. The ACVS and hence Γ depend on two parameters,

namely, ϕ and σ_ϵ^2. The ML estimator $\hat{\phi}$ for ϕ is the value of ϕ that minimizes the reduced (or profile) log likelihood function, namely,

$$l^{(\mathrm{AR})}(\phi) \equiv -\log(1-\phi^2) + N\log(C(\phi)/N) + N, \text{ where } C(\phi) \equiv \widetilde{X}_0^2(1-\phi^2) + \sum_{t=1}^{N-1}(\widetilde{X}_t - \phi\widetilde{X}_{t-1})^2$$

(for details, see, e.g., §9.8 of Percival and Walden 1993). Differentiation of the above yields

$$A^{(\mathrm{AR})}(\phi) \equiv \frac{\phi C(\phi)}{N} - (1-\phi^2)\left(\sum_{t=1}^{N-1}\widetilde{X}_t\widetilde{X}_{t-1} - \phi\widetilde{X}_t^2\right),$$

which is a cubic polynomial in ϕ. The desired estimator $\hat{\phi}$ is the root of the polynomial equation $A^{(\mathrm{AR})}(\phi) = 0$ that minimizes $l^{(\mathrm{AR})}(\phi)$. The corresponding ML estimator of σ_ϵ^2 is given by $\hat{\sigma}_\epsilon^2 \equiv C(\hat{\phi})/N$. When dealing with a time series with missing values, the above formulation does not apply, but we can make use of a Kalman filtering (state space) formulation of an AR(1) process to compute the ML estimators for ϕ and σ_ϵ^2 (see Jones 1980 for details).

b. MLEs for Fractionally Differenced Processes

In the case of an FD process, we take $\mathbf{U}$ to be $[\widetilde{Y}_0, \widetilde{Y}_1, \ldots, \widetilde{Y}_{N-1}]^T$, where the recentered time series $\{\widetilde{Y}_t\}$ is assumed to obey an FD process with parameters ϕ and σ_ε^2 (these fully determine the covariance matrix Γ). We can formulate a reduced log likelihood function for δ as follows. We first compute the partial autocorrelation sequence (PACS) $\phi_{t,t}$, $t = 1, \ldots, N-1$, which is given by $\phi_{t,t} = \frac{\delta}{t-\delta}$ (Hosking 1981). The PACS is used to recursively compute the coefficients of the best linear predictor of $\widetilde{Y}_t$ given $\widetilde{Y}_{t-1}, \ldots, \widetilde{Y}_0$ for $t = 2, \ldots, N-1$. These coefficients are given by

$$\phi_{t,k} = \phi_{t-1,k} - \phi_{t,t}\phi_{t-1,t-k}, \quad k = 1, \ldots, t-1,$$

and are used to form

$$e_t \equiv \widetilde{Y}_t - \sum_{k=1}^{t}\phi_{t,k}\widetilde{Y}_{t-k}, \quad t = 1, \ldots, N-1$$

(we define e_0 to be $\widetilde{Y}_0$). We also use the PACS is to compute a sequence $\{v_t\}$ relating $\mathrm{var}\{e_t\}$ to $\mathrm{var}\{e_0\} = \mathrm{var}\{\widetilde{Y}_t\}$ (the latter is given by Equation (8)):

$$v_t = \mathrm{var}\{e_0\}\prod_{n=1}^{t}(1-\phi_{n,n}^2), \quad t = 0, \ldots, N-1.$$

Given $\widetilde{Y}_t$, the sequences $\{\phi_{t,k}\}$, $\{e_t\}$ and $\{v_t\}$ are all implicit functions of δ and are entirely determined by it. Define $\hat{\varepsilon}_t \equiv e_t/\sqrt{v_t}$, and let

$$\hat{\sigma}_\varepsilon^2(\delta) \equiv \frac{1}{N}\sum_{t=0}^{N-1} \hat{\varepsilon}_t^2. \tag{13}$$

The reduced log likelihood takes the form

$$l^{(\mathrm{FD})}(\delta) \equiv N\,\log\left(\hat{\sigma}_\varepsilon^2(\delta)\right) + N\log\left(\frac{\Gamma(1-2\delta)}{\Gamma^2(1-\delta)}\right) + \sum_{t=1}^{N-1}(N-t)\,\log\left(1-\phi_{t,t}^2\right).$$

We can numerically minimize the above to obtain the ML estimate $\hat{\delta}$. After we have $\hat{\delta}$, we can obtain the ML estimate $\hat{\sigma}_\varepsilon^2$ for σ_ε^2 by substituting $\hat{\delta}$ into Equation (13). Again, we cannot use the above formulation for time series with missing values, but the ML procedure can be adjusted to handle this case (see Palma and Chan 1997 for details).

Appendix B: Goodness of Fit Tests

Here we describe the four statistical tests that we used to assess the adequacy of short and long memory models. In what follows, we let $\hat{e}_t$ stand for the residuals under either the AR model (i.e., $\hat{\epsilon}_t$) or the FD model (i.e., $\hat{\varepsilon}_t$).

a. Spectral Density Function Test

Let $\hat{S}(f_k)$ be the periodogram for the NP index at the Fourier frequency $f_k \equiv k/N$ as given in Equation (6). Let $S(f_k;\hat{\theta})$ be a theoretical SDF that depends on a vector $\hat{\theta}$ of estimated parameters. In the AR(1) and FD cases, the functional forms for $S(f_k;\hat{\theta})$ are given in, respectively, Equations (2) and (7), and we have, respectively, $\hat{\theta} = [\hat{\phi}, \hat{\sigma}_\epsilon^2]^T$ and $\hat{\theta} = [\hat{\delta}, \hat{\sigma}_\varepsilon^2]^T$. Letting M be the integer part of $\frac{1}{2}(N-1)$, the SDF test statistic is given by

$$T_1 \equiv \frac{NA}{4\pi B^2}, \text{ where } A \equiv \sum_{k=1}^{M}\left(\frac{\hat{S}(f_k)}{S(f_k;\hat{\theta})}\right)^2 \text{ and } B \equiv \sum_{k=1}^{M}\frac{\hat{S}(f_k)}{S(f_k;\hat{\theta})}.$$

Under the null hypothesis of a correct model, this test statistic is asymptotically normally distributed with mean $1/\pi$ and variance $2/(\pi^2 N)$ (for details, see Milhøj 1981 and §10.2 of Beran

1994). We can reject the null hypothesis at level of significance α when $\sqrt{N/2}(\pi T_1 - 1)$ exceeds the upper $(1-\alpha) \times 100\%$ percentage point $Q_1(1-\alpha)$ for the standard normal distribution; e.g., $Q_1(1-\alpha) \doteq 1.96$ when $\alpha = 0.05$. The critical level $\hat{\alpha}$ for T_1 is given by $\Phi(\sqrt{N/2}[\pi T_1 - 1])$, where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal random variable.

b. Cumulative Periodogram Test

Let $\hat{S}_{\hat{e}}(f_k)$ be the periodogram for $\hat{e}_t$ at the Fourier frequency $f_k \equiv k/N$, i.e., the right-hand side of Equation (6) with $\widetilde{X}_t$ replaced by $\hat{e}_t$. We form the normalized cumulative periodogram

$$\mathcal{P}_l \equiv \frac{\sum_{k=1}^{l} \hat{S}_e(f_k)}{\sum_{k=1}^{M} \hat{S}_e(f_k)}, \qquad l = 1, \ldots, M.$$

The test statistic T_2 is given by $\max\{D^+, D^-\}$, where

$$D^+ \equiv \max_{1 \le l \le M-1} \left(\frac{l}{M-1} - \mathcal{P}_l \right) \text{ and } D^- \equiv \max_{1 \le l \le M-1} \left(\mathcal{P}_l - \frac{l-1}{M-1} \right). \tag{14}$$

We reject the null hypothesis of white noise at the α level of significance if D exceeds the upper $\alpha \times 100\%$ percentage point $Q_2(1-\alpha)$ for D under the null hypothesis. To a good approximation, we have

$$Q_2(1-\alpha) \equiv \frac{C(1-\alpha)}{(M-1)^{1/2} + 0.12 + \frac{0.11}{(M-1)^{1/2}}},$$

where $C(0.9) = 1.224$, $C(0.95) = 1.358$ and $C(0.99) = 1.628$ (Stephens 1974). We can get some idea as to what the critical value $\hat{\alpha}$ is by comparing the computed T_2 to $Q_2(0.90)$, $Q_2(0.95)$ and $Q_2(0.99)$.

c. Portmanteau Tests

The portmanteau test is designed to see if the sample ACS of the residuals for lags $\tau = 1, \ldots, K$ is consistent with a hypothesis of zero mean white noise, where K is taken to be relatively small in relation to the sample size N (the sample ACS is defined as in Equation (5) with $\widetilde{X}_t$ replaced by $\hat{e}_t$). Here we consider two variations on the portmanteau test, namely, the Box–Pierce test statistic

T_3 and the Ljung–Box–Pierce test statistic T_4, given by, respectively,

$$T_3 = N \sum_{\tau=1}^{K} \hat{\rho}_\tau^2 \text{ and } T_4 = N(N+2) \sum_{\tau=1}^{K} \frac{\hat{\rho}_\tau^2}{N-\tau}$$

(Box and Pierce 1970; Ljung and Box 1978). For either test statistic, we reject the null hypothesis of white noise at significance level α when the statistic exceeds the $(1-\alpha)\times 100\%$ percentage point $Q_3(1-\alpha) = Q_4(1-\alpha)$ for the chi-square distribution with $K-1$ degrees of freedom. If we let $\chi^2_{K-1}(\cdot)$ represent the corresponding cumulative distribution function, then the critical levels for these tests are given by $\chi^2_{K-1}(T_3)$ and $\chi^2_{K-1}(T_4)$.

Appendix C: Performance of Test Statistics under Incorrect Models

Here we give some details on the Monte Carlo experiments used to determine the probability of rejecting an incorrect model using one of the goodness of fit test statistics discussed in Appendix B. Suppose first that the NP index is in fact a realization of an FD process with parameters $\hat{\delta}$ and $\hat{\sigma}_\varepsilon^2$ given as in Table 1. Given a particular sample size N, we can generate a realization from this process (Davies and Harte 1987; Wood and Chan 1994). We then fit an AR(1) model to this realization using the maximum likelihood method described in Appendix A and apply all four goodness of fit test statistics to assess the hypothesis that the AR(1) model is an adequate fit. We repeat this procedure M times and keep track of the number of times M_j that the test statistic T_j rejects the null hypothesis. Our estimate of the probability that the T_j will reject the null hypothesis is given by $\hat{p}_j \equiv M_j/M$. The statistics of the binomial distribution says that, for large M, the estimator $\hat{p}_j$ should be approximately normally distributed with a mean value given by the true rejection probablility p_j and a variance given by $p_j(1-p_j)/M$. We can thus estimate the standard error in $\hat{p}_j$ using $[\hat{p}_j(1-\hat{p}_j)/M]^{1/2}$. By letting $M = 2500$, we found that the estimated standard errors were no larger than 0.02. The left-hand plot of Figure 3 shows $\hat{p}_j$ as a function of sample size for the four test statistics. By reversing the roles of the AR(1) and FD processes, we obtain the right-hand plot.

	$\hat{\phi}$ (AR)	$\hat{\sigma}_\epsilon$ (AR)	$\hat{\delta}$ (FD)	$\hat{\sigma}_\varepsilon$ (FD)
NP	0.21	2.37	0.17	2.35
95% CI	$[0.02, 0.40]$	$[2.01, 2.67]$	$[0.02, 0.32]$	$[2.00, 2.66]$
Sitka	0.18	1.39	0.18	1.37
95% CI	$[0.02, 0.34]$	$[1.22, 1.54]$	$[0.05, 0.30]$	$[1.20, 1.52]$
Sitka (I)	0.29	1.33	0.24	1.30
95% CI	$[0.14, 0.43]$	$[1.18, 1.47]$	$[0.13, 0.36]$	$[1.15, 1.43]$

Table 1: Autoregressive (AR) and fractionally differenced (FD) process parameter estimates for the NP index, uninterpolated Sitka air temperature and interpolated Sitka air temperature series.

j	model	T_j	$Q_j(0.90)$	$Q_j(0.95)$	$Q_j(0.99)$	$\alpha = 0.05$ test result	$\hat{\alpha}$
1	AR	0.30	0.38	0.39	0.42	fail to reject	0.67
	FD	0.28	"	"	"	fail to reject	0.78
	WN	0.39	"	"	"	reject	0.05
2	AR	0.10	0.17	0.19	0.23	fail to reject	$\gg 0.1$
	FD	0.07	"	"	"	fail to reject	$\gg 0.1$
	WN	0.21	"	"	"	reject	≈ 0.03
3	AR	4.65	7.74	9.45	13.31	fail to reject	0.32
	FD	3.12	"	"	"	fail to reject	0.54
	WN	12.63	"	"	"	reject	0.01
4	AR	4.97	7.74	9.45	13.31	fail to reject	0.29
	FD	3.34	"	"	"	fail to reject	0.50
	WN	13.31	"	"	"	reject	0.01

Table 2: Model goodness of fit tests for the NP index.

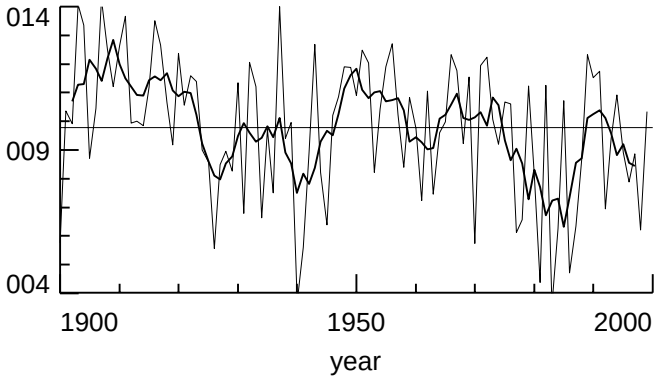

Figure 1: Plot of the NP index (thin curve) and a five year running average of the index (thick). The thin horizontal line depicts the sample mean (1009.8) for the index.

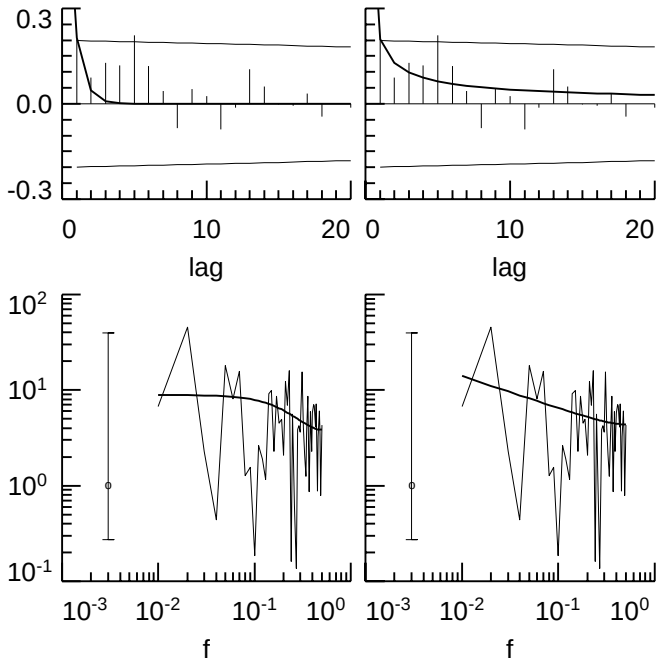

Figure 2: Sample autocorrelation sequence (ACS) and periodogram for the NP index, along with theoretical ACSs and spectral density functions (SDFs) for fitted AR process (left-hand plots) and fitted FD process (right).

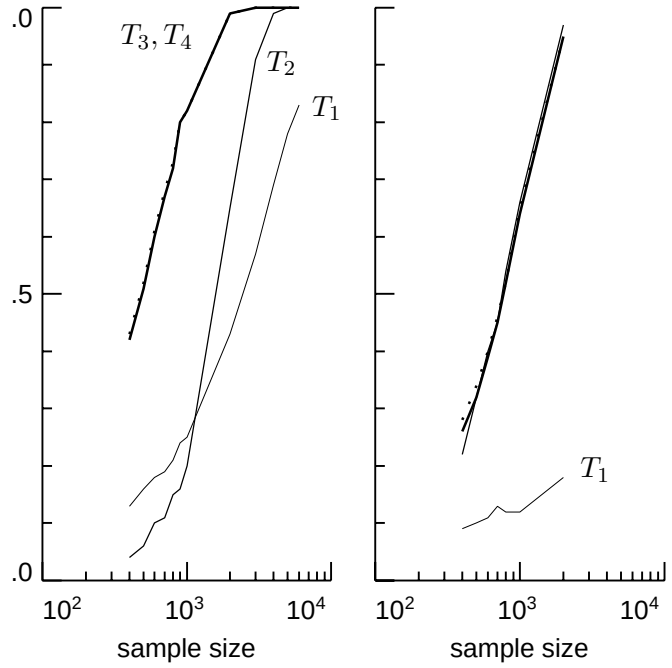

Figure 3: Probability (as a function of sample size) of rejecting the null hypothesis that a fitted model A is adequate for a realization of a process B when using the test statistics $T_1, \ldots, T_4$. In the left-hand plot, model A and process B are, respectively, an AR(1) model and an FD process with parameters δ and σ_ε^2 set to the values estimated for the NP index; in the right-hand plot, A and B are an FD model and an AR(1) process with ϕ and σ_ϵ^2 again set to the values estimated for the NP index. In both cases the best statistics for identifying that a particular model is not correct are the two portmanteau test statistics T_3 and T_4 (however, the cumulative periodogram test statistic T_2 is competitive with T_3 and T_4 when fitting an FD model to realizations of an AR(1) process).

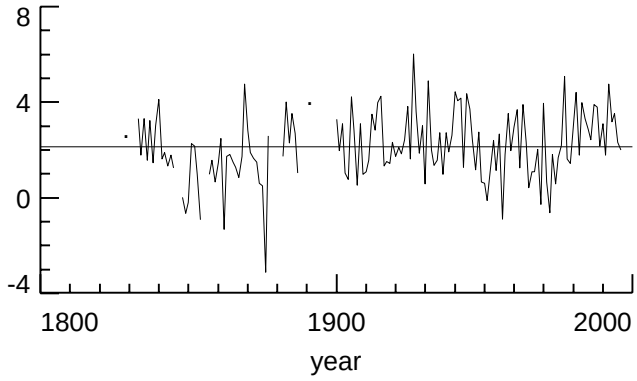

Figure 4: Plot of Sitka winter air temperatures (broken curve). The thin horizontal line depicts the sample mean (2.13) for the series.

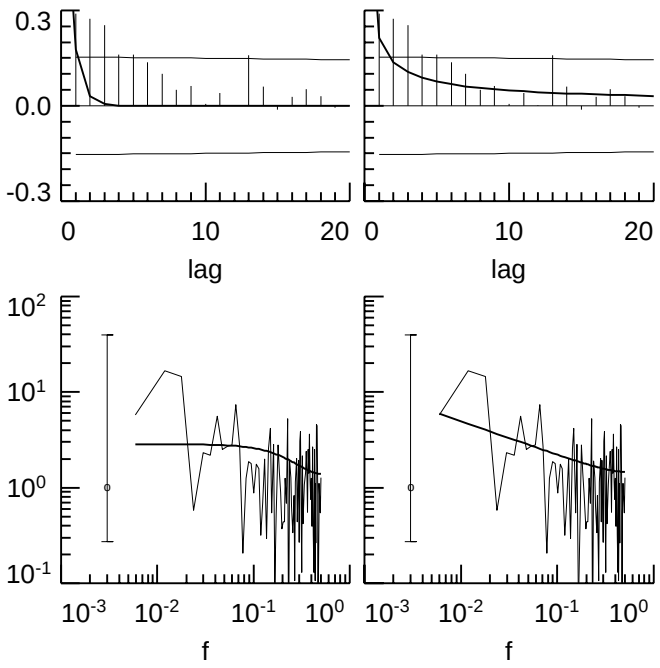

Figure 5: Sample autocorrelation sequence (ACS) and periodogram for Sitka winter air temperatures, along with theoretical ACSs and spectral density functions (SDFs) for fitted AR process (left-hand plots) and fitted FD process (right).

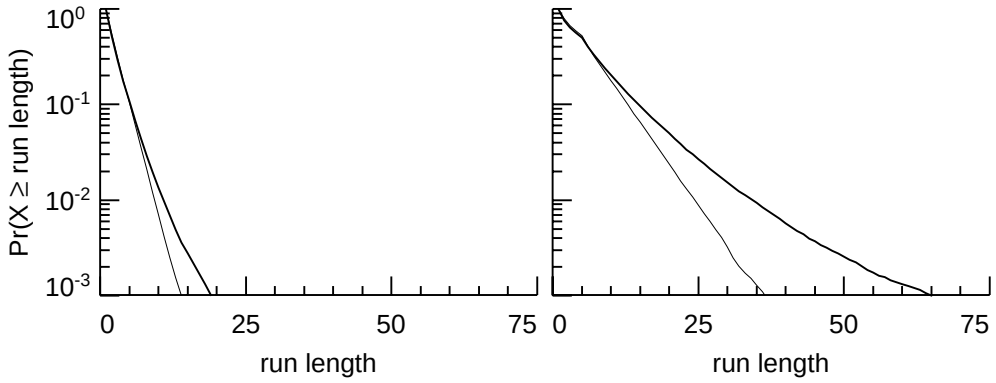

Figure 6: Probability of observing a run that is greater than or equal to a specified run length. The thin (thick) curves denote the AR (FD) process. The left-hand plot is for processes without smoothing, whereas the right-hand plot is for processes subjected to a five year running average.

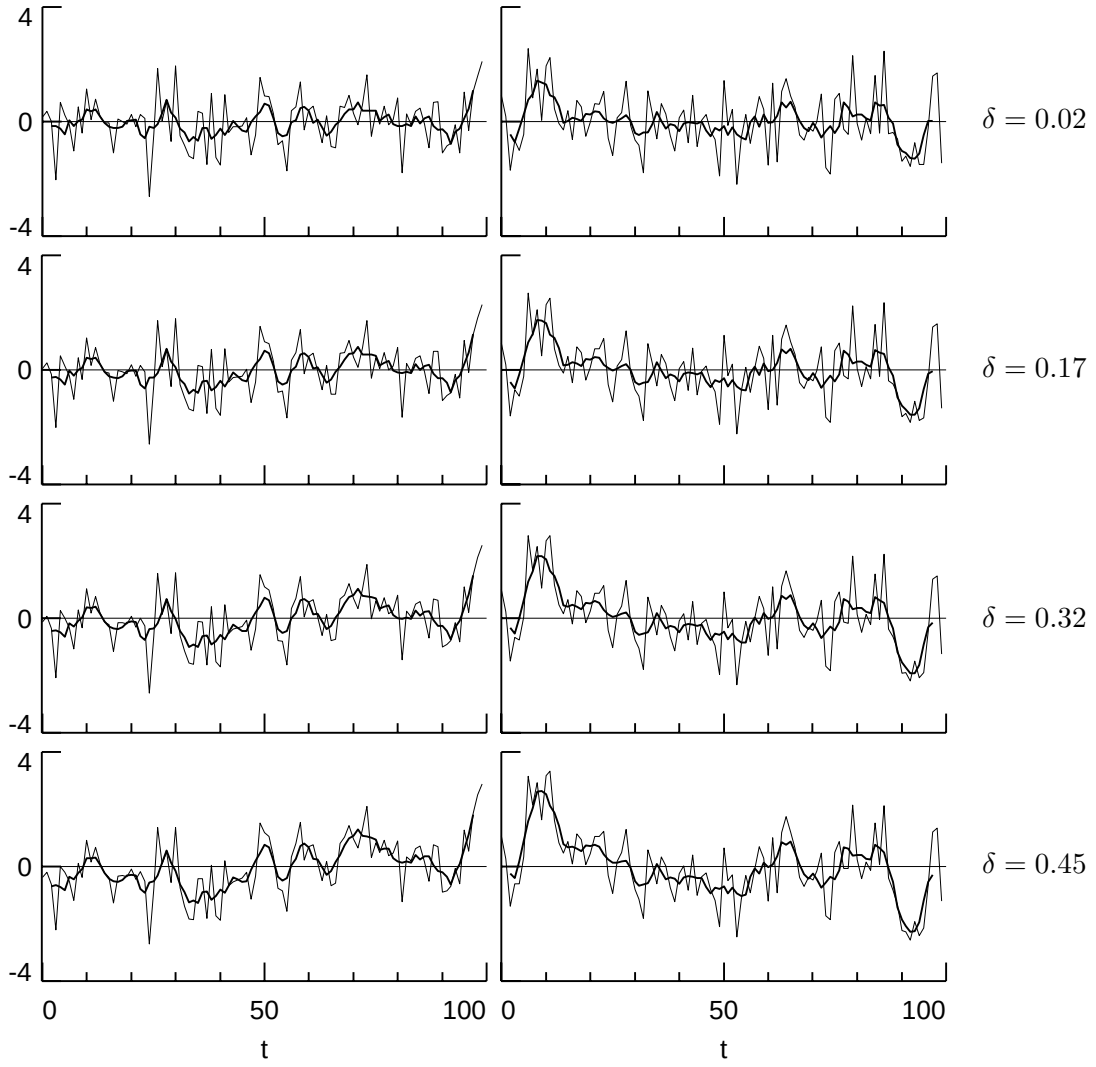

Figure 7: Simulated realizations of FD processes with different parameters δ (thin curves) along with five point running averages (thick).

REFERENCES

Abramowitz, M., and I. A. Stegun, editors, 1964: *Handbook of Mathematical Functions*. US Government Printing Office (reprinted in 1968 by Dover Publications), 1046 pp.

Andersson, M. K., 1998: On the effects of imposing or ignoring long memory when forecasting. *Working Paper Series in Economics and Finance No. 225,* Department of Economic Statistics, Stockholm School of Economics, 14 pp .

Beran, J., 1994: *Statistics for Long Memory Processes*. Chapman and Hall, 315 pp.

Box, G. E. P., and D. A. Pierce, 1970: Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *J. Amer. Stat. Assoc.*, **65**, 1509–1526.

Davies, R. B., and D. S. Harte, 1987: Tests for Hurst effect. *Biometrika*, **74**, 95–101.

Feldstein, S. B., 2000: The time scale, power spectra and climate noise properties of teleconnection patterns. *J. Climate*, **13**, 4430–4440.

Fuller, W. A., 1996: *Introduction to Statistical Time Series* (Second Edition). Wiley–Interscience, 698 pp.

Granger, C. W. J., and R. Joyeux, 1980: An introduction to long-memory time series models and fractional differencing. *J. Time Series Analy.* **1**, 15–29.

Haines, K., and A. Hannachi, 1995: Weather regimes in the Pacific from a GCM. *J. Atmos. Sci.*, **52**, 2444–2462.

Hare, S. R., and N. J. Mantua, 2000: Empirical evidence for North Pacific regime shifts in 1977 and 1989. *Prog. Oceanogr.*, **47**, 103–146.

Hosking, J. R. M., 1981: Fractional differencing. *Biometrika*, **68**, 165–176.

Jones, R. H., 1980: Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, **22**, 389–395.

Kay, S. M., 1981: Efficient generation of colored noise. *Proc. IEEE*, **69**, 480–481.

Latif, M., and T. P. Barnett, 1994: Causes of decadel climate variability over the North Pacific and North America. *Science*, **266**, 634–637.

Ljung, G. M., and G. E. P. Box, 1978: On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.

Mantua, N. J., S. R. Hare, Y. Zang and J. M. Wallace, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Am. Meteorol. Soc.*, **78**, 1069–1079.

Milhøj, A., 1981: A test of fit in time series models. *Biometrika*, **68**, 177–187.

Minobe, S., 1999: Resonance in bidecadal and pentadecadal climate oscillations over the North Pacific: Role in climate regime shifts. *Geophys. Res. Lett.*, **26**, 855–858.

Overland, J. E., J. M. Adams and N. A. Bond, 1999: Decadal variability of the Aleutian low and its relation to high-latitude circulation. *J. Clim.*, **12**, 1542–1548.

Overland, J. E., J. M. Adams and H. O. Mofjeld, 2000: Chaos in the North Pacific: spatial modes and temporal irregularity. *Prog. Oceanogr.*, **47**, 337–354.

Palma, W., and N. H. Chan, 1997: Estimation and forecasting of long-memory time series with missing values. *J. Forecast*, **16**, 395–410.

Palmer, T. N., 1999: A non-linear dynamical perspective on climate prediction. *J. Clim.*, **12**, 575–591.

Percival, D. B., and A. T. Walden, 1993: *Spectral Analysis for Physical Applicataions: Multitaper and Conventional Univariate Techniques*. Cambridge University Press, 583 pp.

Pierce, D. W., 2001: Distinguishing coupled ocean-atmosphere interactions from background noise in the North Pacific. *Prog. Oceanogr.*, **47**, in press.

Priestley, H. B., 1981: *Spectral Analysis and Time Series*. Academic Press, 890 pp.

Stephens, M. A., 1974: EDF statistics for goodness of fit and some comparisons. *J. Amer. Stat. Assoc.*, **69**, 730–737.

Trenberth, K. E., and D. A. Paolino, 1980: The Northern Hemisphere sea level pressure data set: Trends, errors and discontinuites. *Mon. Weather Rev.*, **108**, 855–872.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp., p. 71.

Wood, A. T. A., and G. Chan, 1994: Simulation of stationary Gaussian processes in $[0,1]^d$. *J. Comp. Graph. Stat,*, **3**, 409–432.

Wunsch, C., 2000: The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. *Bull. Am. Meteorol. Soc.*, **80**, 245–255.