MODELING NORTH PACIFIC CLIMATE TIME SERIES

DONALD B. PERCIVAL*, JAMES E. OVERLAND†, AND
HAROLD O. MOFJELD†

Abstract. We present a case study in modeling the North Pacific (NP) index, which is a time series related to atmospheric pressure variations at sea level. We consider three statistical models, namely, a Gaussian stationary autoregressive process, a Gaussian stationary fractionally differenced (FD) process, and a 'signal plus noise' process consisting of a square wave oscillation with a pentadecadal period embedded in Gaussian white noise. Each model depends upon three parameters, so all three models are equally simple. Statistically each model fits the NP index equally well. The fact that this index consists of just a hundred observations makes it unrealistic to expect to be able to clearly prefer one model over the other. Although the models fit equally well, their implications for the long term behavior of the NP index can be quite different in terms of, e.g., generating regimes of characteristic lengths (i.e., stretches of years over which the NP index is predominantly either above or below its long term average value). Because we cannot determine a preferred model statistically, we are faced with either entertaining multiple models when considering what the long term behavior of the NP index is likely to be or using physical arguments to select one model. The latter approach would arguably favor the FD process because it has an interpretation as the synthesis of first order differential equations involving many different damping constants.

Key words. Autoregressive process, North Pacific index, fractionally differenced process, time series analysis.

AMS(MOS) subject classifications. 62P12, 60G10, 60G15.

1. Introduction. There has been considerable interest in recent years in understanding how the climate in the North Pacific (NP) influences fish and mammals in the region. For example, the dramatic drop in the abundance of Stellar sea lions in the Gulf of Alaska over the last fifty years has been hypothesized to be due in part to NP climate variability. One particular measure of this climate variability is the Aleutian low sea level pressure field averaged over the months November to March [16, 12]. This series is known as the NP index and is plotted in Figure 1 (thin curve) for the years 1900 to 1999. Our goal is to investigate the nature of interdecadal variability in this climate time series, but the shortness of this series (one hundred points in all) presents major difficulties in our ability to say with some degree of confidence what the NP index will look like in the future. Faced with a lack of data, one approach is to investigate various models. The idea is to see what models are reasonable fits to the observed series and then to investigate what each model implies about the future behavior of the NP index. We cannot expect to identity one 'true' model for the index

*Applied Physics Laboratory, Box 355640, University of Washington, Seattle, Washington 98195–5640, dbp@apl.washington.edu.

†Pacific Marine Environmental Laboratory, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, Washington 98115–6349.

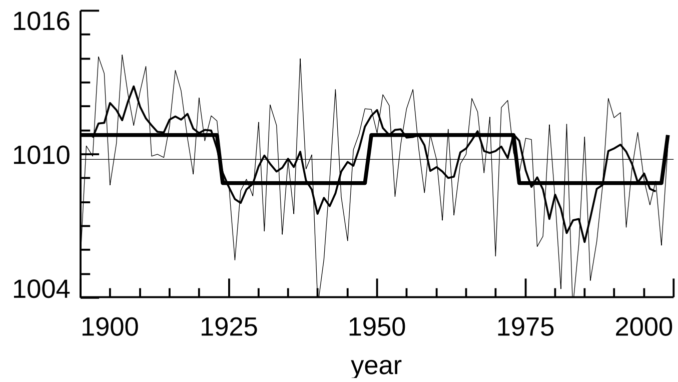

FIG. 1. *Plot of the NP index (thin curve), a five year running average of the index (thicker) and a pentadecadal square wave oscillation used as a model for the index (thickest). There is also a thin horizontal line on the plot. The height of this line depicts the sample mean (1009.8 millibars) for the index.*

based upon the small amount of available data, but we can investigate what implications the different models have on interesting aspects of the index.

The question then arises as to which models we should consider for the NP index. There are many viable candidates. In [13], we compared two purely stochastic models, namely, a 'short memory' model specified by a Gaussian first order autoregressive (AR(1)) process, and a 'long memory' model given by a Gaussian fractionally differenced (FD) process. In [14], we used the matching pursuit algorithm to investigate evidence for a hypothesis by Minobe [11] that the NP index is subject to pentadecadal and bidecadal oscillations in the form of rapid transitions (or 'regime shifts') that 'cannot be attributed to a single sinusoidal-wavelike variability.' This work leads to us to consider a simple 'signal plus noise' model consisting of a square wave oscillation (SWO) observed in the presence of Gaussian white noise. Our intent here is to study the relative merits of this model and the two purely stochastic models.

The remainder of this article is organized as follows. We briefly describe our three models in §2, after which we discuss their parameter estimation in §3. We assess how well each of the models fits the NP index in §4. Given the small amount of available data, we cannot hope to definitively discriminate amongst the three models on statistical grounds, a fact that we quantify in §5. We discuss in §6 the different implications these three models have for extrapolations and for the notion of regime shifts. We state our conclusions in §7.

2. Statistical models for the NP index. In the subsections below, we discuss two Gaussian stationary processes and a 'signal plus Gaussian noise' process as models for the NP index. The stationary processes are an AR(1) process and an FD process, while the 'signal plus noise' process consists of an SWO observed in the presence of Gaussian white noise. All three processes are fully specified by three parameters and hence can be considered to be 'equally simple.' One of the parameters for each process is related to the overall mean value of the NP index, while a second is critical in setting its overall variance. The final parameter controls either (i) the shapes of both the autocovariance sequence (ACVS) and the spectral density function (SDF) for the stationary processes or (ii) the amplitude of the SWO. An important distinguishing characteristic of each process is expressed in their ACVSs. The ACVS for an AR(1) process dies down quickly (exponentially) to zero, so this process is said to have 'short memory;' on the other hand, for an FD process, the ACVS dies down slowly (hyperbolically) and is said to be associated with 'long memory.' For the 'signal plus noise' process, if we force it into the mold of a stationary process by introducing a random phase for the SWO (i.e., location of first zero crossing), the resulting ACVS is periodic and hence never dies down to zero.

2.1. Short memory model. For this model we regard the NP index as a realization of portion $X_0, X_1, \ldots, X_{N-1}$ of a stationary Gaussian AR(1) process $\{X_t\}$; i.e.,

$$X_t - \mu_X = \phi(X_{t-1} - \mu_X) + \epsilon_t = \sum_{k=0}^{\infty} \phi^k \epsilon_{t-k}, \tag{2.1}$$

where $\mu_X = E\{X_t\}$ is the process mean; $\{\epsilon_t\}$ is a Gaussian white noise process with mean zero and variance σ_ϵ^2; and $|\phi| < 1$. We note that, if $\phi = 0$, then $\{X_t\}$ reduces to Gaussian white noise. The ACVS for an AR(1) process is given by

$$s_{X,\tau} \equiv \operatorname{cov}\{X_t, X_{t+\tau}\} = \sigma_\epsilon^2 \frac{\phi^{|\tau|}}{1-\phi^2}, \tag{2.2}$$

where $\tau \in \mathbb{Z}$ (the set of all integers). Its SDF is

$$S_X(f;\theta_X) = \frac{\sigma_\epsilon^2}{1+\phi^2 - 2\phi\cos(2\pi f)}, \tag{2.3}$$

where $\theta_X \equiv \{\phi, \sigma_\epsilon^2\}$, and $|f| \le 1/2$. An AR(1) process is widely used in the physical sciences as a default model for correlated time series [17]. It is suggested as a viable model for the NP index by the fact that, when presented with all possible autoregressive/moving average models of orders ranging from zero to ten, the default order selection criterion used by the ITSM software package [5] picked the AR(1) process as the best model [7]. An AR(1) process is related to a discretized first order stochastic differential equation with a single damping constant that is related to ϕ. In [17], a measure of the decorrelation time (or integral time scale) of an AR(1) process is given by

$$\tau_D \equiv 1 + 2\sum_{\tau=1}^{\infty} \frac{s_{X,\tau}}{s_{X,0}} = \frac{1+\phi}{1-\phi}; \tag{2.4}$$

i.e., the subseries $X_{n\lceil \tau_D \rceil}$, $n \in \mathbb{Z}$, can be regarded as a reasonable approximation to a white noise process.

2.2. Long memory model. For this model we regard the NP index as a realization of a portion $Y_0, Y_1, \ldots, Y_{N-1}$ of a stationary Gaussian FD process $\{Y_t\}$; i.e,

$$Y_t - \mu_Y = \sum_{k=0}^{\infty} \frac{\Gamma(1+\delta)}{\Gamma(k+1)\Gamma(1+\delta-k)} (-1)^k (Y_{t-k} - \mu_Y) \tag{2.5}$$

$$= \sum_{k=0}^{\infty} \frac{\Gamma(1-\delta)}{\Gamma(k+1)\Gamma(1-\delta-k)} (-1)^k \varepsilon_{t-k}, \tag{2.6}$$

where $\mu_Y = E\{Y_t\}$ is the process mean; $\{\varepsilon_t\}$ is a Gaussian white noise process with mean zero and variance σ_ε^2; and $|\delta| < 1/2$. If $\delta = 0$, then $\{Y_t\}$ reduces to a white noise process. The process exhibits long memory if $\delta > 0$. The ACVS and SDF for an FD process are given by

$$s_{Y,\tau} = \sigma_\varepsilon^2 \frac{\sin(\pi\delta)\Gamma(1-2\delta)\Gamma(\tau+\delta)}{\pi\Gamma(\tau+1-\delta)} \tag{2.7}$$

and

$$S_Y(f; \theta_Y) = \frac{\sigma_\varepsilon^2}{|2\sin(\pi f)|^{2\delta}}, \tag{2.8}$$

where $\theta_Y \equiv \{\delta, \sigma_\varepsilon^2\}$. Given $s_{Y,0}$ (i.e., the process variance), we can compute $s_{Y,1}, s_{Y,2}, \ldots$ recursively using the simple formula

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

For large τ, we have $s_{Y,\tau} \propto |\tau|^{2\delta-1}$ approximately, while, for small f, we have $S_Y(f) \propto 1/|f|^{2\delta}$ approximately. An FD model is suggested for the NP index by the fact that, when presented with all autoregressive/fractionally differenced models with the autoregressive order ranging from zero to ten, the ITSM order selection criterion [5] picked the FD process as the best model [7]. An FD process can be synthesized using an aggregation of first order differential equations involving many different damping constants [1]. The measure of the decorrelation time given by Equation (2.4) is infinite for long memory FD processes.

2.3. Square wave oscillation plus noise model. Recently Minobe [11] analyzed the NP index and found evidence for 'regime' shifts (see his paper for citations to prior work on this hypothesis). Loosely speaking, a regime is a stretch of years over which the index is essentially either above or below the sample mean of the entire record. Consider, for example, the index from 1901 to 1923 (thin curve in Figure 1), over which it is essentially greater than its sample mean (thin horizontal line), with the only exceptions being the years 1905 and 1919. This positive regime with a duration of 23 years is more clearly seen in a five year running mean (thicker curve), which lies strictly above the sample mean up until 1924. Continuing on, we find this running mean to be essentially (but not strictly) below the sample mean from 1924 to 1946, indicating a negative regime of 23 years. This pattern arguably persists up to the end of the recorded data, providing evidence for a pentadecadal oscillation [11].

A regime shift is the transition between regimes of different polarities. Part of the current interest in the notion of regimes is that the shifts occur rapidly, signaling a change in conditions that can influence the relative abundances of fish and mammal populations. To quote from Minobe [11],

> 'Although the regime shifts in the present century are characterized by a pentadecadal timescale, rapid transitions from one regime to another cannot be attributed to a single sinusoidal-wavelike variability. The rapid-transition nature of 20th century regime shifts suggests that the pentadecadal variability is characterized by a non-sinusoidal variation such as a rectangular wave ....'

In [14], we informally assessed Minobe's hypothesis using matching pursuit. The basic idea behind matching pursuit is to construct a 'dictionary' of vectors with characteristics that might be useful in describing a time series. We constructed a dictionary consisting of (i) sinusoidal vectors from a discrete Fourier transform, (ii) SWOs with periods ranging from two up to one hundred years in combination with all possible alignments (i.e., phase shifts) and (iii) wavelet and scaling vectors from a Haar wavelet transform spanning two up to one hundred years (again with all possible alignments). When presented with this dictionary, matching pursuit found the best single approximating vector to be a square wave oscillation with a pentadecadal period; i.e., given the choice between square wave and sinusoidal oscillations, matching pursuit preferentially favored the former over the latter, in agreement with Minobe's hypothesis. In addition, the alignment of the selected pentadecadal oscillation was such that its regime shifts are in quite good agreement with those given in Minobe [11].

In view of this result from matching pursuit, we will also consider the NP index to be a realization of a portion $Z_0, Z_1, \ldots, Z_{N-1}$ of an SWO process defined by

$$Z_t = \mu_Z + \beta D_{t+1} + e_t, \tag{2.9}$$

where μ_Z and β are unknown parameters; $\{e_t\}$ is a Gaussian white noise process with mean zero and variance σ_e^2; and

$$D_t \equiv \begin{cases} +0.1, & \text{when } t = 0, \ldots, 24; \\ -0.1, & \text{when } t = 25, \ldots, 49; \text{ and} \\ D_{t \bmod 50}, & \text{otherwise} \end{cases} \tag{2.10}$$

(by definition, '$t \bmod 50$' is equal to $t + 50n$, where n is the unique integer such that $0 \le t + 50n \le 49$). Note that, if $\beta = 0$, then $\{Z_t\}$ becomes a white noise process. When $\beta \neq 0$, an intuitive measure for the decorrelation time is fifty since the subseries Z_{j+50k}, $k \in \mathbb{Z}$, for a fixed $j \in \mathbb{Z}$ is exactly a white noise process.

3. Estimation of model parameters. Each of the statistical models for the NP index described in §2 has three parameters: μ_X, ϕ and σ_ϵ^2 for the AR(1) process $\{X_t\}$; μ_Y, δ and σ_ε^2 for the FD process $\{Y_t\}$; and μ_Z, β and σ_e^2 for the SWO process $\{Z_t\}$. The parameters μ_X, μ_Y and μ_Z capture the overall level of the index and can be estimated via the sample mean; i.e.,

$$\hat{\mu}_X = \frac{1}{N} \sum_{t=0}^{N-1} X_t, \;\; \hat{\mu}_Y = \frac{1}{N} \sum_{t=0}^{N-1} Y_t \;\text{ and }\; \hat{\mu}_Z = \frac{1}{N} \sum_{t=0}^{N-1} Z_t.$$

For the AR(1) and FD models, the sample mean is a suboptimal estimator, but is commonly used in time series analysis because there is little practical loss of efficiency.

After computing the sample mean, we use it to recenter the NP index, yielding

$$\widetilde{X}_t \equiv X_t - \hat{\mu}_X, \;\; \widetilde{Y}_t \equiv Y_t - \hat{\mu}_Y \;\text{ and }\; \widetilde{Z}_t \equiv Z_t - \hat{\mu}_Z.$$

As an approximation, we consider $\widetilde{X}_t$, $\widetilde{Y}_t$ and $\widetilde{Z}_t$ to be AR(1), FD and SWO processes with $\mu_X = \mu_Y = \mu_Z = 0$. There are now two parameters left to estimate for each process. We estimate these using the maximum likelihood (ML) method ([13] has details about this approach for the AR(1) and FD models). Let $\hat{\phi}$, $\hat{\sigma}_\epsilon^2$, $\hat{\delta}$, $\hat{\sigma}_\varepsilon^2$, $\hat{\beta}$ and $\hat{\sigma}_e^2$ denote the ML estimators. Large sample theory says that these estimators are all (at least) approximately normally distributed and unbiased (both statements are exact for $\hat{\beta}$, and $\hat{\sigma}_e^2$ is also exactly unbiased). The variances of $\hat{\phi}$ and $\hat{\delta}$ are approximately given by, respectively, $(1 - \phi^2)/N$ and $6/\pi^2 N$, whereas the variance of $\hat{\beta}$ is given exactly by σ_e^2 (the dependence on N is implicit here because the SWO of Equation (2.10) is defined to have unit norm over N observations). The variances of $\hat{\sigma}_\epsilon^2$, $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_e^2$ all have a similar form, namely, $2\sigma_\epsilon^4/N$, $2\sigma_\varepsilon^4/N$ and $2\sigma_e^4/N$, respectively (these are approximations for the AR(1) and FD processes but exact for the SWO process). Monte Carlo experiments indicate that the large sample approximations we have quoted for the AR(1) and FD processes are in fact very accurate for $N \ge 100$ [13].

TABLE 1

Autoregressive (AR), fractionally differenced (FD) and square wave oscillator (SWO) process parameter estimates for the NP index, along with associated 95% confidence intervals (CIs)

model	parameter	95% CI	σ	95% CI
AR	$\hat{\phi} \doteq 0.21$	$[0.02, 0.40]$	$\hat{\sigma}_\epsilon \doteq 2.37$	$[2.01, 2.67]$
FD	$\hat{\delta} \doteq 0.17$	$[0.02, 0.32]$	$\hat{\sigma}_\varepsilon \doteq 2.35$	$[2.00, 2.66]$
SWO	$\hat{\beta} \doteq -10.09$	$[-14.51, -5.67]$	$\hat{\sigma}_e \doteq 2.21$	$[1.88, 2.50]$

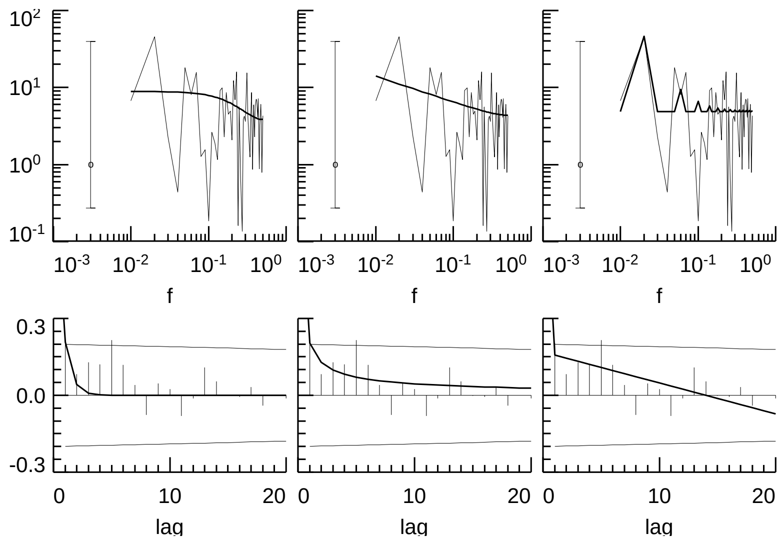

FIG. 2. *Spectral density functions (thick curves, top row) for the fitted AR, FD and SWO models (left-hand, middle and right-hand columns, respectively) and the corresponding periodogram for the NP index (thin curves, top row), along with theoretical autocorrelation sequences (ACSs, thick curves, bottom row) for the models and the corresponding sample ACS (depicted as deviations from zero in each plot in the bottom row). The two thin curves above and below the zero line in each lower plot indicate limits within which individual sample ACVS estimates should fall 95% of the time when the true process is white noise. The error bar in each of the upper plots gives a similar assessment of the inherent variability in the periodogram.*

Table 1 shows the parameter estimates for the three models, along with associated 95% confidence intervals (CIs) for the unknown parameters based upon ML theory. Based upon the CIs, we can reject the white noise hypotheses $\phi = 0$, $\delta = 0$ and $\beta = 0$ at the 0.05 level of significance. The estimated residual standard deviations are remarkably similar for the three fitted models, with the one for the SWO model being the smallest. The fitted SWO model $\hat{\mu}_Z + \hat{\beta} D_{t+1}$ is shown in Figure 1 as the thickest curve.

4. Assessment of fitted models. As a first step in assessing how well the three fitted models capture the characteristics of the NP index, let us compare plots of the theoretical SDFs corresponding to the fitted models with the periodogram for the index (top row of Figure 2). By definition the SDFs for the AR(1) and FD processes are given by Equations (2.3) and (2.8), respectively. With ϕ set equal to $\hat{\phi}$ and δ set to $\hat{\delta}$, the theoretical SDFs for the fitted models are given by the solid curves in the upper left-hand and middle plots. For the SWO process, we can define a theoretical SDF to be the expected value of the periodogram for $\widehat{Z}_t \equiv \hat{\beta} D_{t+1} + e_t$ conditional on regarding $\hat{\beta}$ as fixed. By definition, this periodogram is given by

$$\hat{S}_{\widehat{Z}}(f_k) \equiv \frac{1}{N} \left| \sum_{t=0}^{N-1} \widehat{Z}_t e^{-i2\pi f_k t} \right|^2, \quad f_k \equiv k/N, \;\; k = 1, \ldots, N/2, \tag{4.1}$$

and its expected value is given by

$$E\{\hat{S}_{\widehat{Z}}(f_k)\} = \frac{1}{N} \left| \sum_{t=0}^{N-1} \hat{\beta} D_{t+1} e^{-i2\pi f_k t} \right|^2 + \sigma_e^2.$$

With σ_e^2 set to its ML estimate, the above is plotted as the thick curve in the upper right-hand plot. The periodogram for the NP index is given by Equation (4.1) with $\widetilde{X}_t$ substituted for $\widehat{Z}_t$ and is shown on all three upper plots as the thin jagged curves. An assessment of the sampling variability in the periodogram is given by the confidence intervals (CIs) shown in the left-hand portion of each plot in the top row. These CIs emanate from a circle. If we move the CI so that the circle is centered on a particular value of $\hat{S}_{\widetilde{X}}(f_k)$, then we have a 95% CI for the true SDF at frequency f_k.

If we take into account the sampling variability inherent in the periodogram, all three theoretical SDFs are viable summaries of the spectral content in the NP index. While the fitted AR(1) and FD SDFs are broad-band summaries, the SDF for the SWO model locks onto the largest value in the periodogram, associated with a period of $1/f_2 = 50$ years. There are additional peaks in the SWO SDF visible at certain harmonics of f_2. The most prominent is the second harmonic, i.e., $3f_2 = f_6$, which is associated with a period of $1/f_6 \doteq 16.7$ years. The fact that there appears to be some spectral content surrounding this second harmonic in the periodogram lends some additional (but statistically questionable) support for Minobe's hypothesis that a pure sinusoidal oscillation is inadequate to describe the NP index. Note that the background level from which these harmonics arise is determined by the variance σ_e^2 of the white noise process $\{e_t\}$ in the model.

Let us now compare plots of the theoretical autocorrelation sequences (ACSs) with the sample ACS for the NP index (bottom row of Figure 2). The ACSs for the AR(1) and FD processes are given by, respectively, $\rho_{X,\tau} \equiv s_{X,\tau}/s_{X,0}$ and $\rho_{Y,\tau} \equiv s_{Y,\tau}/s_{Y,0}$, where $s_{X,\tau}$ and $s_{Y,\tau}$ are given in Equations (2.2) and (2.7). The inverse Fourier transform of $E\{\hat{S}(f)\}$, $|f| \leq 1/2$, can be used to form a theoretical ACVS for the SWO process, from which the corresponding ACS follows. The theoretical ACSs for the fitted AR(1), FD and SWO models are the solid curves in, respectively, the left-hand, middle and right-hand plots. Letting $\widetilde{X}_t$ stand for recentered NP index, the sample ACS is given by

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

The sample ACS is shown in each plot as deviations from the zero line. An assessment of the sampling variability in the sample ACS is given by the upper and lower thin curves in each plot, which depict 95% confidence intervals (CIs) under the assumption that the NP index is a realization of a white noise process (see Corollary 6.3.6.2 of [8] for details).

Each model captures certain aspects of the sample ACS for the NP index. The AR(1) ACS captures $\hat{\rho}_{\widetilde{X},1}$ almost perfectly, but then damps down to zero very quickly. The FD ACS also matches $\hat{\rho}_{\widetilde{X},1}$ very well, but then decays to zero slowly and hence visually captures the sample ACS a bit better. The SWO ACS underestimates $\hat{\rho}_{\widetilde{X},1}$ somewhat and, like the FD ACS, has much larger values than the AR(1) ACS for lags below about $\tau = 10$. When we take into account the sampling variability in the sample ACS, however, all three models seem qualitatively reasonable.

Let us now look at three test statistics that give us quantitative assessments of the fitted models. The first statistic T_1 compares the periodogram $\hat{S}_{\widetilde{X}}(f_k)$ for the NP index to the fitted $S(f_k; \hat{\theta})$ from a particular model [10, 1]:

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

Here $S(f_k; \hat{\theta})$ is taken to be either $S_X(f_k; \hat{\theta}_X)$ of Equation (2.3) for the AR(1) model or $S_Y(f_k; \hat{\theta}_Y)$ of Equation (2.8) for the FD model, where $\hat{\theta}_X \equiv \{\hat{\phi}, \hat{\sigma}_\epsilon^2\}$ and $\hat{\theta}_Y \equiv \{\hat{\delta}, \hat{\sigma}_\varepsilon^2\}$ (the theory behind T_1 developed in [10] does not extend to the SDF we defined for the SWO model). Under the null hypothesis that the model corresponding to $S(f_k; \hat{\theta})$ is correct, T_1 is asymptotically normal with mean $1/\pi$ and variance $2/(\pi^2 N)$. We reject the null hypothesis at a level of significance of α when $\sqrt{N/2}(\pi T_1 - 1)$ exceeds $Q_1(1-\alpha)$, which is the upper $(1-\alpha) \times 100\%$ percentage point for the standard normal distribution.

The other two test statistics make use of residuals from each model. For the AR(1), FD and SWO models, we denote the residuals as, respectively, $\hat{\epsilon}_t$, $\hat{\varepsilon}_t$ and $\hat{e}_t$ This notation emphasizes the fact that the residuals can be regarded as estimates of the white noise processes ϵ_t, ε_t and e_t

involved in each model (see Equations (2.1), (2.6) and (2.9)). There are details in [13] about how to compute ϵ_t and ε_t for the AR(1) and FD models (sometimes called the observed innovations). For the SWO model, we have

$$\hat{e}_t \equiv \widetilde{Z}_t - \hat{\beta} D_{t+1}.$$

If a particular model is adequate for the NP index, then the residuals from the fitted model should resemble a sample from a white noise process. The cumulative periodogram test statistic assesses this resemblance by determining if the periodogram for the residuals is consistent with the white noise assumption [2, 4]. This test statistic is defined as

$$T_2 = \max\left\{\max_l\left(\frac{l}{\lfloor\frac{N-1}{2}\rfloor - 1} - \mathcal{P}_l\right), \max_l\left(\mathcal{P}_l - \frac{l-1}{\lfloor\frac{N-1}{2}\rfloor - 1}\right)\right\},$$

where $\mathcal{P}_l$ is the normalized cumulative periodogram for, say, $\hat{\epsilon}_t$:

$$\mathcal{P}_l \equiv \frac{\sum_{k=1}^{l} \hat{S}_{\hat{\epsilon}}(f_k)}{\sum_{k=1}^{\lfloor\frac{N-1}{2}\rfloor} \hat{S}_{\hat{\epsilon}}(f_k)}$$

(analogous expressions hold for $\hat{\varepsilon}_t$ and $\hat{e}_t$). We reject that null hypothesis of white noise at the α level of significance if T_2 exceeds

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

where $C(0.9) = 1.224$, $C(0.95) = 1.358$ and $C(0.99) = 1.628$ [15].

The last test statistic determines if the residuals are consistent with the white noise hypothesis by examining their sample ACS. Given a positive integer K (taken to be small compared the sample size N), the Box–Pierce portmanteau test statistic [3] is defined for, e.g., the AR(1) residuals $\hat{\epsilon}$ as

$$T_3 = N \sum_{\tau=1}^{K} \hat{\rho}^2_{\hat{\epsilon},\tau},$$

where $\rho_{\hat{\epsilon}_t,\tau}$ is the sample ACS for $\hat{\epsilon}_t$ (similar expressions hold for $\hat{\varepsilon}_t$ and $\hat{e}_t$). We reject the null hypothesis of white noise at a level of significance α if T_3 exceeds $Q_3(1-\alpha)$, which is the $(1-\alpha)\times 100\%$ percentage point for the chi-square distribution with $K-1$ degrees of freedom. In keeping with recommendations in the literature, we set $K = N/20 = 5$, but we also looked at $K = 10$ and obtained virtually the same results. (We note that there is a variation on T_3 known as the Ljung–Box–Pierce portmanteau test statistic [9], which takes the form

$$T_4 = N(N+2) \sum_{\tau=1}^{K} \frac{\hat{\rho}^2_{\hat{\epsilon}_t,\tau}}{N-\tau}.$$

TABLE 2

Model goodness of fit tests for the NP index. In the column reporting the result of using an $\alpha = .05$ level test, 'accept' should be interpreted as shorthand for 'fail to reject the null hypothesis.'

j	model	T_j	$Q_j(0.90)$	$Q_j(0.95)$	$Q_j(0.99)$	$\alpha = .05$ test	$\hat{\alpha}$
1	AR	0.30	0.38	0.39	0.42	accept	0.67
	FD	0.28	"	"	"	accept	0.78
	WN	0.39	"	"	"	reject	0.05
2	AR	0.10	0.17	0.19	0.23	accept	$\gg 0.1$
	FD	0.07	"	"	"	accept	$\gg 0.1$
	SWO	0.10	"	"	"	accept	$\gg 0.1$
	WN	0.21	"	"	"	reject	≈ 0.03
3	AR	4.65	7.74	9.45	13.31	accept	0.32
	FD	3.12	"	"	"	accept	0.54
	SWO	2.83	"	"	"	accept	0.59
	WN	12.63	"	"	"	reject	0.01

The results that we got using T_3 and T_4 were virtually identical.)

For all three test statistics T_j, we reject the 'model is adequate' hypothesis when T_j is 'too big' as quantified by a percentage point from a distribution under the null hypothesis. Table 2 shows the results of these goodness of fit tests for the AR(1), FD and SWO models, along with an additional model that regards the NP index as a realization of a white noise process. This model is denoted as 'WN' in the table and has 'residuals' that are taken to be the NP index itself. At the 0.05 level of significance, all the test statistics reject the hypothesis that the NP index is white noise, but they all fail to reject the adequacy of the AR(1), FD and SWO models; i.e., statistically, all three models are viable. The table also gives an indication of the observed level of significant $\hat{\alpha}$ (i.e., the smallest α for which we would end up rejecting the null hypothesis). For all three test statistics and for all three models, $\hat{\alpha}$ is so large that we cannot reject the null hypothesis at any reasonable level of significance.

Finally, let us comment upon the Gaussian assumption that we have made for each model. Quantile-quantile plots [6] of the residuals from the three models indicate some possible departures from Gaussianity. These departures are not severe, but are of concern since it is unclear how they impact the analysis presented here. This topic is certainly appropriate for future research.

5. Model discrimination. The fact that the AR(1), FD and SWO models are viable for the NP index from a statistical point of view raises the question as to whether or not we could reasonably hope to distinguish amongst these models given the fact that we only have one hundred values for the NP index. To address this question, we consider the following experiment. For the sake of argument, let us assume that the fitted FD model is in fact exactly correct for the NP index. Using procedures outlined in [13], we can generate simulated time series of a desired length $N' \geq N$ from this fitted model. We can then fit an AR(1) model to each simulated FD series and evaluate the fitted AR(1) model using each of our three test statistics T_j. By repeating the above a large number of times (where 'large' is here taken to be 2500), we can estimate the probability that T_j will (correctly) reject the null hypothesis that the AR(1) model is the correct one. This experiment will tell us how much power T_j has in being able to say that the AR(1) model is incorrect. We can also try to fit an SWO model to the same simulated FD series to get an idea as to how much power each test statistic has in this case. We can repeat the above for a variety of sample sizes N'. We can also repeat all of the above by generating simulation from either the fitted AR(1) model or the fitted SWO model and then fitting the other two models to the simulated time series.

Figure 3 shows plots of the estimated probabilities versus sample size N' that, using a level of significance $\alpha = 0.05$, the three test statistics T_j will reject a fitted model A when the true model is B. For completeness, we also estimated these probabilities when A and B were the same model – theoretically these probabilities should be close to $\alpha = 0.05$. From this figure, we can see that, at best, we have only a 30% chance of rejecting the null hypothesis when we have only one hundred observations. The best scenario is when we use the T_1 test statistic on an AR(1) fit to a time series generated by an SWO process. If the SWO model is indeed correct for the NP index, we should have a reasonable chance of rejecting the AR(1) and FD models by the end of the current century. We would need two to three hundred years worth of data in order to have a 50% chance of rejecting the SWO model when either of the other two models is correct. The test statistics are less successful at discriminating between the AR(1) and FD models, where we would need about five hundred years of the NP index in order to have even a 50% chance of correctly rejecting the null hypothesis. Finally, it is of interest to note that no one test statistic performs uniformly better than the other two, which emphasizes the need to use them all.

6. Model implications. We have seen that, based upon our three test statistics T_j, there is no statistical reason to prefer either an AR(1), FD or SWO model for the NP index. In addition, all three models depend on just three parameters and hence can be regarded as equally simple, so the principle of parsimony would not prefer any of the models (this ignores the fact that the SWO model was picked out by matching pursuit rather than *a priori*). Even though all three models seem to match the NP index equally well, the models can have implications that are quite different and potentially important, as the following demonstrates.

As a starting point for our discussion, the left-hand plots in the bottom

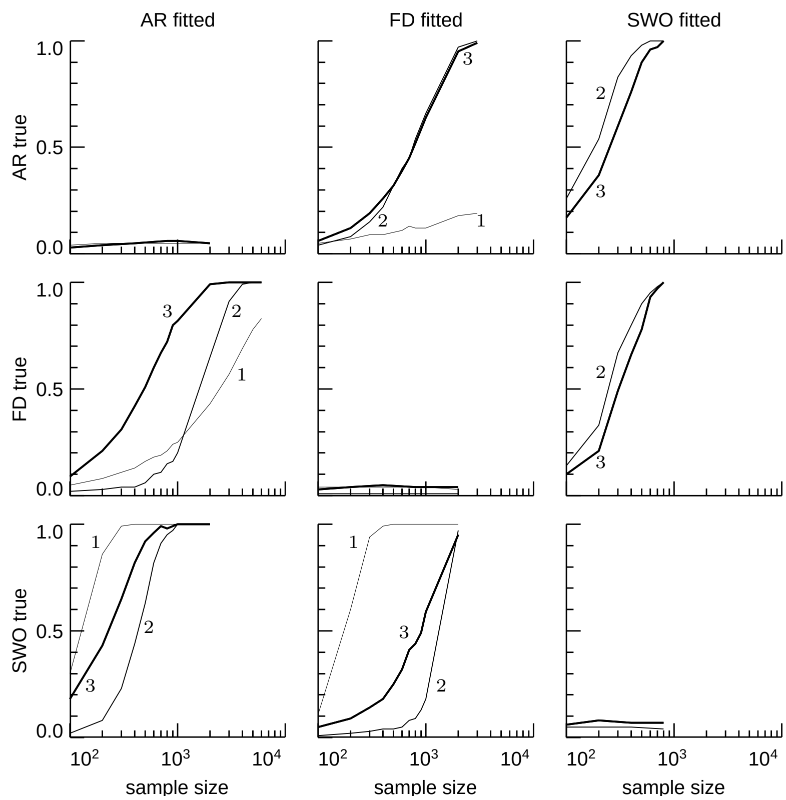

FIG. 3. *Probability (as a function of sample size) of rejecting the null hypothesis at a* 0.05 *level of significance that a fitted model A is adequate for a realization of a process B when using the test statistics* T_1*,* T_2 *and* T_3*. For the plots in the left- to right-hand columns, the fitted models A is, respectively, an FD, AR(1) and SWO model. The same ordering is used for the process B for the plots in the top to bottom rows.*

three rows of Figure 4 show simulated series from each of the three models. From the second down to fourth rows, the series are from, respectively, the AR(1), FD and SWO models. For comparison, the top row shows the recentered NP index itself. Each simulated series is a thousand years long, and the three series are visually rather similar. The right-hand plots show the periodograms corresponding to each series. With the passage of time, the square wave oscillation in the SWO model determines the dominant feature in the periodogram, but note that there is also considerable power at low frequencies in the simulated AR(1) and FD processes. The periodic nature of the SWO and the low frequency power in the other two models can all contribute to the generation of regimes. We thus pose the following two questions. First, how well do these three models support the notion of regimes? Second, are there any differences in the regimes generated by

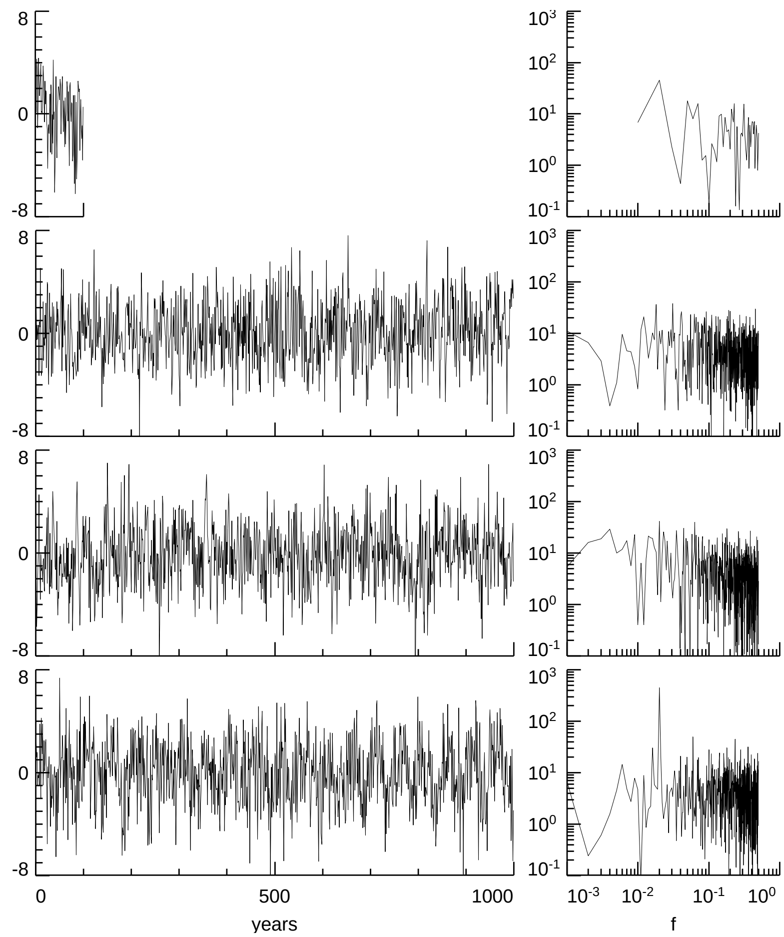

FIG. 4. *Comparison of recentered NP index (left-hand plot in top row) and its periodogram (right-hand) with thousand year simulations of this index and their periodograms using the fitted AR(1) model (second row), FD model (third row) and the SWO model (bottom row).*

these models? Answers to these questions are important in understanding the implications of using any one of these models to say something about expected future patterns of the NP index.

To address these questions, we consider the following computer experiment. For the sake of argument, suppose that the NP index is actually a realization of an FD process. To account for the fact that our parameter estimate $\hat{\delta}$ is not perfect but is subject to sampling error, we start by generating a deviate $\tilde{\delta}$ from the presumed distribution for $\hat{\delta}$, namely, a Gaussian distribution with mean $\hat{\delta}$ and variance $6/(\pi^2 N) = 6/(\pi^2 100)$. We next create a simulated series of length 1024 from an FD process with

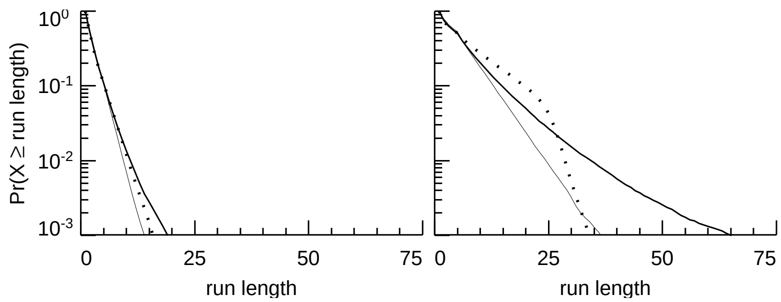

FIG. 5. *Probability of observing a run that is greater than or equal to a specified run length. The thin, thick and dotted curves correspond to the AR, FD and SWO models. The left-hand plot is for processes without smoothing, whereas the right-hand plot is for processes subjected to a five year running average.*

zero mean and parameter $\tilde{\delta}$ (the value we use for σ_ε^2 doesn't matter). We then tabulate the sizes of the observed regimes in the series, where the size of a regime is taken to be the number of contiguous values that are either entirely above or entirely below zero. We also do the same for a five year running average of the simulated series. If we repeat the above procedure a thousand times and if we do the same for the fitted AR(1) and SWO models (making use of the large sample distributions for $\hat{\phi}$ and $\hat{\beta}$), we can empirically determine the probability that a regime has a size greater than or equal to a specified run length.

The results of this experiment are shown in Figure 5. The left- and right-hand plots show the estimated probabilities for, respectively, the original and the five year averaged series. The AR(1), FD and SWO models are indicated by the thin, thick and dotted curves. The more interesting of the two plots is the right-hand one, for which a regime is defined in terms of run lengths of five year running averages. We see that intermediate regime sizes are most likely to occur under the SWO model and that quite large regime sizes are most likely under the FD model. The AR(1) model is the least supportive of the idea of extensive regimes. For example, in comparison to the AR(1) model, a regime whose size is greater than or equal to 23 years (i.e., the 'typical' size indicated by our inspection of Figure 1) is four times more likely under the FD model and is eight times more likely under the SWO model. Thus, even though all three models are viable from a statistical point of view, these models can lead to quite different statements about what the NP index will look like in the long term.

7. Conclusions. We have shown that, from a statistical point of view, the AR(1), FD and SWO models are all adequate descriptions of the NP index despite the fact that these models are quite different from one another. Our inability to identify a best model is directly related to the fact that we only have one hundred observations of the index. We would need several hundred more observations before we would have a reasonable chance of selecting a single best model. All three models include a white noise process as a special case, and one of the parameters in each of the models can be used to test the hypothesis of white noise. In all three cases, we can reject the white noise hypothesis. In terms of their ACSs, the ACS for the AR(1) model has the most rapid decay toward zero. The one for the FD model has a long tail of small positive correlations, while the SWO model has an oscillating ACS. The FD model is more supportive of the notion of extensive regimes than the other two models, but the SWO model supports regimes with sizes dictated by the half-period of its oscillation (25 years). In the absence of a viable physical explanation for a pentadecadal oscillation in the NP index, loose physical considerations might favor the FD model as the best single choice because this model corresponds to an aggregation of first order differential equations, each presumably associated with geophysical phenomena with different time scales. In lieu of picking a single best model, another strategy would be to consider the implications of all three statistically feasible models in any forecasts about the future behavior of the NP index.

Acknowledgments. The work of the first author was supported in part by grants from the Cooperative Institute for Arctic Research (CIFAR), Stellar Sea Lion Research, and from the Pacific Marine Environmental Laboratory, National Oceanic and Atmospheric Administration. We would like to thank the organizers of the IMS Workshop on 'Time Series Analysis and Applications to Geophysical Systems' for the invitation to participate and to thank the other participants for their many helpful comments about our work. The paper is Contribution No. 0000 from NOAA/Pacific Marine Environmental Laboratory.

REFERENCES

[1] J. Beran, *Statistics for Long-Memory Processes*, Chapman and Hall, NY, 1994.

[2] G. E. P. Box, G. M. Jenkins and G. C. Reinsel, *Time Series Analysis: Forecasting and Control (Third Edition)*, Prentice Hall, Englewood Cliffs, NJ, 1994.

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

[4] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods (Second Edition)*, Springer, NY, 1991.

[5] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting (Second Edition)*, Springer, NY, 2002.

[6] J. M. Chambers, W. S. Cleveland, B. Kleiner and P. A. Tukey, *Graphical Methods for Data Analysis*, Duxbury Press, Boston, 1983.

[7] R. A. Davis, private communication, 2001.

[8] W. A. Fuller, *Introduction to Statistical Time Series (Second Edition)*, Wiley–Interscience, NY, 1996.

[9] G. M. Ljung and G. E. P. Box, *On a measure of lack of fit in time series models*, Biometrika **65** (1978), pp. 297–303.

[10] A. Milhøj, *A test of fit in time series models*, Biometrika **68** (2001), pp. 177–187.

[11] S. Minobe, *Resonance in bidecadal and pentadecadal climate oscillations over the North Pacific: Role in climate regime shifts*, Geophys. Res. Lett. **26** (2001), pp. 855–858.

[12] J. E. Overland J. M. Adams and N. A. Bond, *Decadal variability of the Aleutian low and its relation to high-latitude circulation*, J. Climate **12** (1999), pp. 1542–1548..

[13] D. B. Percival, J. E. Overland and H. O. Mofjeld, *Interpretation of North Pacific variability as a short- and long-memory process*, J. Climate **14** (2001), pp. 4545–4559.

[14] ———, *Using matching pursuit to assess circulation changes over the North Pacific*, (2002), to be submitted.

[15] M. A. Stephens, *EDF statistics for goodness of fit and some comparisons*, J. Amer. Stat. Assoc. **69** (1974), pp. 730–737.

[16] K. E. Trenberth and D. A. Paolino, *Hemisphere sea level pressure data set: Trends, errors and discontinuities*, Mon. Weather Rev. **108** (1980), pp. 855–872.

[17] H. von Storch and F. W. Zwiers, *Statistical Analysis in Climate Research*, Cambridge University Press, Cambridge, UK, 1999.