"Eyeballing" Trends in Climate Time Series: A Cautionary Note

Donald B. Percival[1] **and D. Andrew Rothrock**[1]

April 16, 2004

[1]Applied Physics Laboratory, Box 355640, University of Washington, Seattle, WA 98195–5640, USA

Correspondence to Dr. Donald B. Percival, Applied Physics Laboratory, Box 355640, University of Washington, Seattle, WA 98195–5640, USA; Phone: (206)–543–1368; Fax: (206)–543–6785; E-mail: dbp@apl.washington.edu

Abstract

In examining a plot of a time series of a scalar climate variable for indications of climate change, we might pick out what appears to be a linear trend commencing near the end of the record. We demonstrate that visual determination of the starting time of the trend can lead us to incorrectly declare a trend to be significant when we base our assessment on standard linear regression analysis; in fact a presumed level of significance of 5% can be smaller than the actual level by up to an order of magnitude. We suggest an alternative procedure that is more appropriate for assessing the significance of a trend whose starting point is identified visually.

1 Introduction

There is currently a strong scientific and societal interest in the issue of climate change. It is not uncommon that one examines a simple time series of a scalar climatic variable and asks whether it contains a trend that might indicate whether, by how much, and in what direction climate has changed. We offer here some thoughts about the pitfalls in estimating trends and their significance from selected portions of time series.

Our example is one familiar in our own field of research: the North Atlantic Oscillation (NAO) index. It indicates the strength of the Icelandic low, a persistent feature of northern hemispheric surface atmospheric pressure. It is of interest particularly because its simple definition as the pressure difference from the center to the edge of the low pressure cell allows one to establish a long historical record. The two sites of pressure observations for the original NAO index are Ponta Delgada in the Azores and Reykjavik, Iceland (Hurrell 1995); by using observations from Gibraltar and Iceland (Jones et al. 1997), the beginning of the record can be pushed back to 1824, giving a record for one and three quarters centuries. We use this latter record here. Contiguous monthly values for this index from November 1824 through October 2000 are available from the Web site for the Climatic Research Unit, University of East Anglia. By reflecting the strength of the Icelandic cyclonic circulation, this index is an indicator of the intensity of heat and moisture advection from the North Atlantic into Eurasia and from Eurasia into the arctic. It also indicates the strength of sea ice and fresh water outflow from the Arctic Ocean to the North Atlantic Ocean, and hence of the North Atlantic's control of the global ocean's thermohaline circulation. Because of its coarse characterization of these climatic processes, this index and its interannual variability are commonly cited and hence provide a familiar example with which to illustrate the point of this paper.

In Section 2, we show by a standard trend analysis that the NAO index has increased

during the last thirty years at a rate of 0.05 yr^{-1} and that this rate is significant at the 95% level. We also illustrate that in Gaussian white noise processes, as the NAO index appears to be a sample of, one will find "significant" trends over several decades far more often than the 5% suggested by the 95% level of confidence. The flaw is in selecting the trial interval after viewing the longer data record rather than *a priori*, as is the premise of the statistical theory. In Section 3, we show how one could construct a more stringent test that yields a more appropriate assessment of the significance of a linear trend. We consider in Section 4 the more pathological case of time series with weak but undetectable autocorrelation; in this case even more stringent tests are necessary to keep from wrongly identifying trends. There is a concluding discussion in Section 5.

Before we begin, some comments are in order about prior literature that is related to our main point and to our analysis of the NAO index. The flaw that is at the heart of our discussion has been discussed extensively in the statistical literature under various guises. For example, in his classic work on the analysis of variance, Scheffé (1959) has an extensive discussion about making multiple comparisons between sample means of subsets of the data after viewing the data. The flaw has also been discussed recently in this journal by Lund and Reeves (2002) in the context of the detection of undocumented change points in a time series. Their work considers the proper statistical evaluation of a test for hypothesized changes between two trends, whereas we concentrate on the task of testing for a single trend at the end of a series. Their conclusions are quite similar to ours, as is their remedy of constructing a more stringent test with the help of computer experiments. Their more stringent test is based upon a null hypothesis of Gaussian white noise, but they speculated that further adjustments to their test would be needed for 'heavily correlated errors.' In fact, we find (Section 4) that even weak autocorrelations are enough to change substantially the actual level of significance.

The NAO index has been analyzed in a number of papers, but Wunsch (1999) does so in the context of examining it as a realization of a stationary process. He shows that simulated time series from stationary processes that are reasonable models for the NAO often exhibit 'regimes' (i.e., long stretches during which the time series is consistently above or below its long term mean). While his discussion focuses on these regimes, he briefly notes there is '... a period, particularly since about 1960, of an apparent trend' in the actual NAO series, but also points out that portions of the simulated series also have similar 'trends,' concluding that 'one must ... be wary of apparent trends.' The particular NAO series he investigated is the winter-average Lisbon minus Iceland pressure difference from 1864 to 1996, which is evidently constructed somewhat differently from our NOA series and has 44 fewer values. These differences are evidently why his conclusions are slightly different from what we state in the next section. In particular, whereas his NAO series is consistent with a weak power-law (red) process having a 'near-white spectral density,' and is described as 'somewhat non-Gaussian,' our series appears to be indistinguishable from a realization of Gaussian white noise. His analysis provides good motivation for entertaining weakly correlated anomalies, as we do in Section 4.

2 False Trends in the Case of Uncorrelated Anomalies

To motivate our discussion, let us consider the NAO index X_l consisting of $N = 177$ yearly winter-time atmospheric pressure values for the years $l = 1824, \ldots, 2000$. For a particular year l, we use the average of the January, February and March values for year l and the value for the preceding December $(l-1)$. The left-hand plot of Figure 1 shows this index versus time. The right-hand plot gives its sample autocorrelation sequence (ACS) for lags τ

from one to twenty years, i.e.,

$$\hat{\rho}_\tau = \frac{\sum_{l=1824}^{2000-\tau}(X_l - \overline{X})(X_{l+\tau} - \overline{X})}{\sum_{l=1824}^{2000}(X_l - \overline{X})^2}, \quad \tau = 1, \ldots, 20,$$

where $\overline{X} = \sum_l X_l/N \doteq 0.4759$ is the average of the 177 values. The lines above and below the sample ACS depict upper and lower 95% confidence limits based upon the assumption that the NAO index is a sample from a Gaussian (normally distributed) white noise process; i.e., the true autocorrelations ρ_τ at nonzero lags are all zero (Fuller 1996). If we were to observe many different samples of size 177 from such a process, then the sample ACS at any particular lag should fall between these limits about 95% of the time. The fact that all of the sample autocorrelations fall within these limits suggests we cannot reject the hypothesis that there is no significant correlation in the NAO index. In addition, a plot (not shown) of the quantiles of the NAO index versus the quantiles from a Gaussian distribution indicates that the index follows this distribution quite closely (Chambers et al. 1983). This analysis of the NAO index thus indicates it to be indistinguishable from a sample of Gaussian white noise.

Mindful of the well-known hypothesis that the climate has significantly changed in the recent past, let us visually examine the end of the NAO index in Figure 1. Arguably there is a linear trend with a positive slope starting in 1969 (the dotted vertical line on the figure is placed at 1968). To assess whether or not this trend is statistically significant, we might be tempted to do the following. We entertain a model consisting of a line observed in the presence of Gaussian white noise; i.e., we write

$$X_l = a + bl + \epsilon_l, \quad l = 1969, \ldots, 2000, \tag{1}$$

where a is an unknown intercept, b is an unknown slope, and ϵ_l represents a random anomaly associated with the observation for year l. We assume that these anomalies are a sample of Gaussian white noise with mean zero and unknown variance σ_ϵ^2. With this formulation,

we can estimate the line via least squares (Draper and Smith 1998). The estimated line is drawn on Figure 1. The estimator for the slope b takes the form

$$\hat{b} = \frac{\sum_l (l - 1984.5) X_l}{\sum_l (l - 1984.5)^2} \doteq 0.0516 \ \text{year}^{-1},$$

where the summations run from $l = 1969$ to $l = 2000$, and the value 1984.5 is the mean value of the years 1969, ... , 2000. We can calculate, say, a 95% confidence interval for the unknown true slope b using the formula

$$\hat{b} \pm t_{30}(0.975) \frac{\hat{\sigma}_\epsilon}{\sqrt{\sum_l (l - 1984.5)^2}},$$

where $t_{30}(0.975) \doteq 2.042$ is the upper 97.5% percentage point of student's t-distribution with 30 degrees of freedom, and

$$\hat{\sigma}_\epsilon^2 \equiv \frac{1}{30} \sum_l (X_l - \hat{a} - \hat{b}l)^2 \doteq 1.3419$$

is an estimator of σ_ϵ^2, while $\hat{a} = \frac{1}{32} \sum_l X_l - \hat{b} \cdot 1984.5 \doteq -101.87$ is the least squares estimator of a. This yields a 95% confidence interval for b of the form $[0.006, 0.097]$. Because this interval does not contain the value zero, we seem to have evidence (at the 0.05 level of significance) that the slope is indeed statistically different from zero. Moreover, if we do a careful study of the residuals $e_l \equiv X_l - \hat{a} - \hat{b}l$, we find that there is no compelling reason to doubt the validity of our assumptions about the anomalies ϵ_l. It appears that we have evidence for the claim that the NAO index has an upward linear trend over the last 32 years.

There is, however, a source of concern here, namely, that we did not pick the starting point of the trend *a prior*, but rather made this choice *after* we had examined the data. The procedure that we just used to assess the significance of a linear trend really *requires* that we set up the testing procedure *prior* to any examination of the data. To illustrate this subtle but crucial point, let us conduct the following experiment.

Let W_l be a Gaussian white noise process with zero mean. This process is a sequence of uncorrelated random variables (each with an expected value of zero) and hence by definition

does not contain a linear trend. We can create a realization of a portion $W_{1824}, \ldots, W_{2000}$ of this process covering the same years as the NAO index by sampling 177 deviates from a Gaussian random number generator. One such realization is shown in Figure 2(a). What happens if we decide to search for a linear trend somewhere over, say, the most recent 10 to 50 years? To automate this search, we can fit linear regression models to the last m years of this simulated series, where we allow m to vary from 10 to 50. The model for a particular m takes the form

$$W_l = a_m + b_m l + \epsilon_{m,t}, \quad l = 2000 - (m-1), 2000 - (m-2), \ldots, 2000.$$

The fitted regression lines for the cases $m = 10, 20, 30, 40$ and 50 are drawn on Figure 2. There are 41 fitted regression lines in all for 41 different segments, and we can test for a significant slope for each segment using the same procedure as outlined above; i.e., we compute an appropriate 95% confidence interval and see if the interval contains the value zero or not. If any one of these 41 tests indicates a significant slope, we would claim (incorrectly) that there is a linear trend over the corresponding segment. For the example shown in Figure 2(a), it happens that all 41 tested slopes are declared to be insignificant, so we would claim (correctly) that there is not a trend in any of the segments we have examined from this simulated series. Figure 2(b) shows a second realization, for which our conclusion would be different. Here we reject the null hypothesis for 18 of the 41 tested slopes, including two of the 5 cases depicted on the plot ($m = 20$ and 40). We would now claim (incorrectly) that there is a trend over several examined segments.

How often can we expect to claim there is a trend when there really isn't one, as happened with the series in Figure 2(b)? We can address this question by repeating our experiment a large number of times (100,000). When we do so, we find at least one significant trend slightly more than 38% of the time; i.e., we incorrectly identify a significant trend approximately 7.5 times more often than what would be indicated by the level of significance for an individual

test (5%). The implications of this result on our claim that the NAO index has a significant upward trend are alarming. If this index were a sample from a white noise process and if we could repeatedly sample from this process, we would find a trend not just in 5% of the samples, but rather in 38% of them. This casts serious doubt on the significance of a linear trend in the NAO index.

3 More Realistic Assessment of Significance

Here we consider a more stringent procedure that offer better protection against falsely identifying trends. The proper assessment of the significance of a trend over a stretch of time that has been preferentially picked out by eye raises some difficult questions, some of which are related to what statisticians call the "multiple comparison" problem. Reasonable analytic solutions to this problem are difficult to obtain. In the present case, we propose a simple method for providing a more realistic assessment of significance. The method relies upon computer experiments and hence can be readily adapted to situations that deviate in detail from our motivating example of the NAO index.

The computer experiment we discussed in Section 2 was based on the assumption that we can equate a visual inspection for a linear trend at the end of the NAO index to a series of statistical tests over the most recent 10 to 50 years. We found that, if we perform individual tests for the null hypothesis of a zero slope at a level of significance $\alpha_I = 0.05$ and if we declare there to be a significant slope if any one of the 41 tests rejects the null hypothesis, the overall level of significance α_O is unreasonably high, namely, $\alpha_O = 0.38$. If we were to decrease α_I, we can expect the corresponding α_O to decrease also. The idea then is to perform a series of computer experiments to determine what value of α_I we would need to use so that the overall significance of the test would be, say, $\alpha_O = 0.05$. The thickest (bottom) curve in Figure 3 shows a plot of α_O versus α_I in which we varied α_I from 0.0001

to 0.05. The thin horizontal line indicates when $\alpha_O = 0.05$. This line cuts the thickest curve when $\alpha_I = 0.004$. Hence, in order to have a test with an overall significance of $\alpha_O = 0.05$, we need to perform the 41 individual tests at a level of significance of $\alpha_I = 0.004$ (i.e., rather than using 95% confidence intervals, we increase the confidence level to 99.6%).

Let us now reconsider the example of the NAO index. If we fit 41 least squares lines starting at years 1951, ... , 1991 and ending in 2000 and if we test the null hypothesis of a zero slope at a level of significance of $\alpha_I = 0.004$ using each of the 41 estimates slopes, we find that we cannot reject the null hypothesis in any of the 41 cases. We conclude that our preferentially picked linear trend is not significant at the 0.05 level of significance.

4 False Trends in the Presence of Undetected Weakly Correlated Anomalies

In Section 2 we entertained a white noise model in our experiment to demonstrate the problems that can arise when we attempt to assess the significance of a linear trend at the end of a time series. This choice for a model was motivated by the fact that we could not reject the null hypothesis of white noise for the NAO index. We now demonstrate that the shortness of the NAO index (177 values in all) limits our ability to assess the hypothesis of uncorrelatedness. It is thus of interest to redo our experiments using *correlated* models that might be appropriate for the NAO index.

One of the most commonly used models in climate research for a correlated time series is a first order stationary autoregressive (AR) process U_l (von Storch and Zwiers 1999). This process is defined via the equation

$$U_l = \mu + \phi(U_{l-1} - \mu) + \varepsilon_l,$$

where μ is the mean value of the process; ϕ is a parameter that is less than unity in magnitude and is equal to the unit lag autocorrelation ρ_1 for the process; and ε_l is a Gaussian white

noise process with mean zero and variance σ_ε^2. The theoretical ACS for this AR process is given by $\phi^{|\tau|}$ at lag τ; i.e., the ACS decays to zero exponentially. When $\phi = 0$, the AR process reduces to white noise. The popularity of this model stems in part from the fact that it is related to a stochastic version of a first order differential equation.

Suppose now that the NAO index is actually a realization of an AR process with a small unit lag autocorrelation (i.e., $\rho_1 = \phi$ is small). Given a sample size of only 177, can we expect to be able to detect a small amount of autocorrelation in the series? To address this question, we can generate a realization of length $N = 177$ from an AR process with a given nonzero ϕ and then subject it to the Ljung–Box portmanteau test for white noise (Ljung and Box 1978; Brockwell and Davis 1991). This test is based upon the test statistic

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

where K is typically chosen to be much smaller than N, and $\hat{\rho}_\tau$ is the sample ACS for the realization (here we set $K = 10$, but we also obtained similar results using $K = 5$ and 20). The test consists of comparing T_K to an upper percentage point (say 95%) of a chi-square distribution with K degrees of freedom. If T_K exceeds this percentage point, we have evidence for rejecting the white noise hypothesis. By repeating this procedure many times (100,000), we can determine the probability that the portmanteau test will be able to correctly reject the null hypothesis that this sample of 177 values is from a white noise process. For $\phi = 0.1$ and 0.2, we find this probability to be, respectively, 0.12 and 0.41. Thus, if we were to examine many different time series, each of which is a realization of a first order AR process with $\phi = 0.2$, the portmanteau test will correctly reject the null hypothesis of white noise for about 41% of these series, but the majority of these series (59%) will be declared to be indistinguishable from white noise. The situation is even more unfavorable when $\phi = 0.1$. If we strengthen the correlation by increasing ϕ, the ability of the portmanteau test to detect autocorrelation improves. For $\phi = 0.25$ and 0.3, the probabilities

of correctly rejecting the white noise hypothesis are, respectively, 0.62 and 0.81. We also note that the poor performance of the portmanteau test when $\phi = 0.1$ and 0.2 is not unique to this particular test. Similar results hold for other tests that can be used to detect autocorrelation, including the cumulative periodogram test, the turning point test, the difference-sign test and the rank test (Brockwell and Davis 1991).

To allow for the possibility that the NAO index consists of undetected weakly correlated anomalies, let us repeat the experiment described in Section 2 by replacing $W_{1824}, \ldots, W_{2000}$ with a realization from an AR processes with $\phi = 0.1$. As was true in the case of white noise, this process by definition does not contain a linear trend. With the experiment so modified we now incorrectly find a significant trend about 49% of the time; i.e., we are now incorrect approximately 10 times more often than indicated by the 5% level of significance for an individual test. If we now redo the entire experiment letting $\phi = 0.2$, we incorrectly find a significant trend about 59% of the time.

As for the corrective procedure described in Section 3, since the portmanteau test is unlikely to detect weak autocorrelation, we extend the computer experiments to AR processes, leading to the middle ($\phi = 0.1$) and upper ($\phi = 0.2$) curves in Figure 3. We see that, under these two scenarios, we need to set $\alpha_I = 0.002$ and $\alpha_I = 0.0008$ to achieve $\alpha_O = 0.05$.

In summary, if the NAO index were in fact a sample of a weakly correlated AR process, we are unlikely to be able to detect this correlation due to the limited number of observations. The presence of this undetected weak correlation has the undesirable effect of increasing the probability of our incorrectly identifying a significant trend at the end of the series to the point that we will be finding a trend more likely than not!

5 Discussion

Statistical significance as it is commonly presented in introductory textbooks on statistics is cast in terms of a well-defined repeatable experiment designed to test a scientific hypothesis. For example, we might entertain the null hypothesis that a particular coin is fair and design an experiment in which we count the number of observed "heads" after flipping the coin a certain number of times. We can then evaluate the null hypothesis by comparing the results of our experiment with statistics dictated by the binomial distribution.

The analysis of climate time series such as the NAO index does not fit well into this simple mindset; i.e., while we can flip a coin as many times as we like to obtain multiple realizations, we only have a single realization of the NAO index. Because the study of the climate is empirical rather than experimental, we often find ourselves entertaining scientific hypotheses *after* we have already obtained and studied the data in some detail. The main point of this cautionary note is that the notion of statistical significance can be quite subtle in this setup and that caution is called for in the use of statistical tests. We need to beware of the effect of preselection ("eyeballing" in our example) when evaluating the significance of certain results. To this end, and in keeping with recommendations by Wunsch (1999) and Lund and Reeves (2002), we advocate the use of computer experiments – in the spirit of the ones presented in this paper – to guard against unwarranted claims of statistical significance. The key here is to come up with some way of modeling (at least roughly) the preselection procedure. In our case, we were able to model preselection as a series of elementary tests that then forms the basis for obtaining a more realistic overall assessment of significance. Other ways of modeling preselection in our example can be entertained, but the point that we wish to make is that it is incumbent upon us to model preselection in some manner and to temper claims of statistical significance by thinking carefully about the subtleties raised by preselection.

Acknowledgements

We would like to thank R. Moritz, H. Stern, R. Colony, the editor, and an anonymous reviewer for constructive comments that improved the quality of this paper. We gratefully acknowledge the support of this work by the EOS/IDS Program of the National Aeronautics and Space Administration, Grant No. NAG5–9334.

REFERENCES

Brockwell, P. J. and R. A. Davis, 1991: *Time Series: Theory and Method* (Second Edition). Springer, 577 pp.

Chambers, J. M., W. S. Cleveland, B. Kleiner and P. A. Tukey, 1983: *Graphical Methods for Data Analysis*. Duxbury Press, 395 pp.

Draper, N. R. and H. Smith, 1998: *Applied Regression Analysis* (Third Edition). John Wiley & Sons, 706 pp.

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

Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation and relationships to regional temperature and precipitation. *Science*, **269**, 676–679.

Lund, R., and J. Reeves, 2002: Detection of undocumented changepoints: a revision of the two-phase regression model. *Journal of Climate*, **15**, 2547–2554.

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

Scheffé, N., 1959: *The Analysis of Variance*. John Wiley & Sons, 477 pp.

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

Wunsch, C., 1999: The Interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. *Bulletin of the American Meteorological Society*, **80**, 245–255.

FIGURE CAPTIONS

Figure 1. (a) NAO index and (b) its sample autocorrelation sequence (ACS) for lags of one to twenty years. The vertical dotted line in (a) marks the year 1968, while the thick line shows a linear least squares fit for the data from 1969 to 2000. In (b), the curves above and below the sample ACS are upper and lower 95% confidence limits under the assumption that the NAO index is a realization of Gaussian white noise.

Figure 2. Two time series simulated from a white noise process, along with regression lines fit over the last 10, 20, 30, 40 and 50 years of data.

Figure 3. Plot of overall significance α_O versus assumed significance α_I for individual tests. As discussed in Section 4, the curves are for time series whose lag one autocorrelations are $\phi = 0$ (i.e., white noise), 0.1 and 0.2.

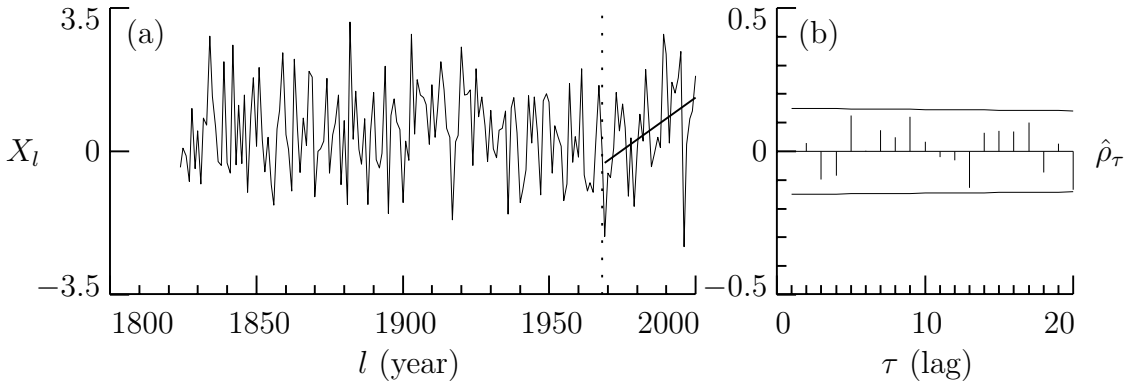

Figure 1: (a) NAO index and (b) its sample autocorrelation sequence (ACS) for lags of one to twenty years. The vertical dotted line in (a) marks the year 1968, while the thick line shows a linear least squares fit for the data from 1969 to 2000. In (b), the curves above and below the sample ACS are upper and lower 95% confidence limits under the assumption that the NAO index is a realization of Gaussian white noise.

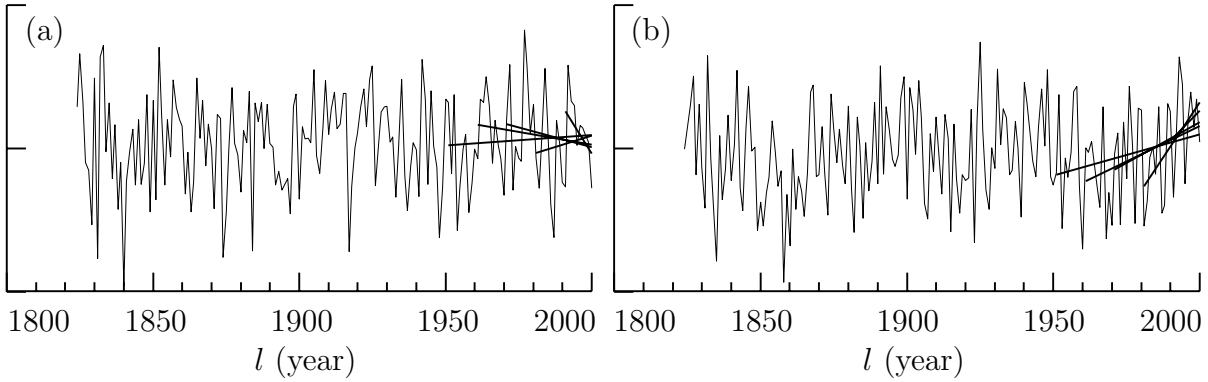

Figure 2: Two time series simulated from a white noise process, along with regression lines fit over the last 10, 20, 30, 40 and 50 years of data.

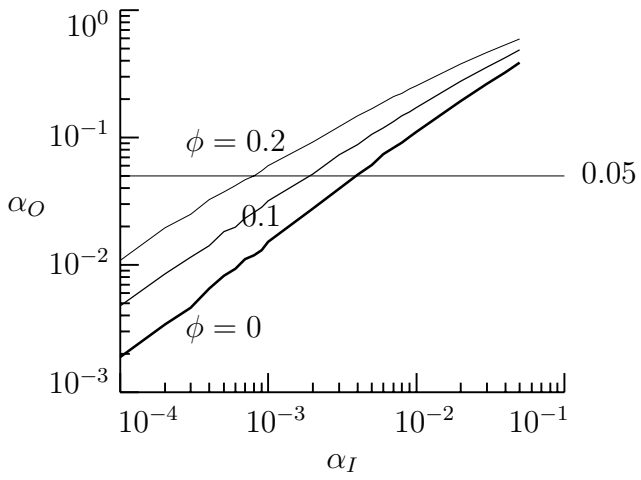

Figure 3: Plot of overall significance α_O versus assumed significance α_I for individual tests. As discussed in Section 4, the curves are for time series whose lag one autocorrelations are $\phi = 0$ (i.e., white noise), 0.1 and 0.2.