

THE UNIVERSITY  
OF STATISTICS

DEPARTMENT OF STATISTICS  
UNIVERSITY OF NORTH CAROLINA



LONG-RANGE DEPENDENCE AND GLOBAL WARMING

by

Richard L. Smith

June 1992

Mimeo Series #2077

DEPARTMENT OF STATISTICS  
Chapel Hill, North Carolina

INSTITUTE OF STATISTICS MIMEO SERIES

UNIVERSITY OF NORTH CAROLINA  
DEPARTMENT OF STATISTICS  
CHAPEL HILL, NC 27599

2000 FAN, J.: Asymptotic normality for deconvolving kernel density estimators, Sept. 1989. *Sankhya*, to appear.

2001 MURPHY, S.: Time-dependent coefficients in a Cox-type regression model, Sept. 1989. (Dissertation)

2002 SEN, P.K.: Statistical functionals, stopping times and asymptotic minimum risk property, Sept. 1989. (18 pages)

2003 CHAKRAVARTI, I.M.: Pseudo-dimensions, Sept. 1989. *Probabil Behadr, ed., Wiley, 1990, to app*

2004 NAKAMURA, M.: Transformations, Sept. 1989. (Dissertation)

2005 FAN, J.: On the estimation of q

2006 BELE, S.: Some comments on c, Oct. 1989. *Amer. J. Anthropology*

2007 CHU, C.K. & MARRON, J.S.: (Sept. 1989. (18 pages)

2008 CHU, C.K.: Some results in non

2009 STROMBERG, A.J.: Robust eff (Dissertation)

2010 FAN, J.: Adaptively local 1-dim

2011 FAN, J.: Global behavior of dec

2012 BROOKS, M. & MARRON, J.S.: bandwidth for kernel estimates of

2013 CSORGO, S. & DODUNEKOVA, R.: Inhibit, Dec. 1989. (28 pages)

2014 CSORGO, S., HAEUSLER, E. & MASON, D.: The quantile-transform approach to the asymptotic distribution of modulus trimmed sums, Dec. 1989. (14 pages)

2015 CSORGO, S. & MASON, D.: Intermediate sums and stochastic compactness of maxima, Dec. 1989. (21 pages)

2016 CSORGO, S. & MASON, D.: Intermediate- and extreme-sum processes, Dec. 1989. (18 pages)

2017 CSORGO, S., HAEUSLER, E. & MASON, D.: The quantile-transform - empirical-process approach to limit theorems for sums of order statistics, May 1990. (52 pages)

MIMEO	Richard L. Smith
SERIES	LONG RANGE DEPENDENCE AND GLOBAL WARMING
#2077	
NAME	
NAME	
DATE	

The Library of the Department of Statistics  
North Carolina State University

2018 CROWELL, J. & SEN, P.K.: On stopping times for fixed-width confidence bands, Dec. 1989. (15 pages)

2019 JOHNSON, N.L. & KOTZ, S.: Non-existence of certain polynomial regressions in random sum theory, Jan. 1990. (6 pages)

2020 CARLSTEIN, E. & BELE, S.: Nonparametric change-point estimation for data from an ergodic sequence, Jan. 1990. (9 pages)

2021 REN, J.I. & SEN, P.K.: On Hadamard differentiability of statistical functional process, Jan. 1990. (21 pages)

2022 SHEPP, L., SIMONS, G. & YAO, Y.-C.: On a problem of ammunition rationing, Mar. 1990. (22 pages)

...: Nonparametric regression with errors in variables, Mar. 1990.

... J.S. & PARK, B.U.: A simple root n bandwidth selector, Apr. 1990.

...: Optimal estimation of the renewal function, optimal block replacement confidence bands, Apr. 1990. (Dissertation)

...: Central limit theorems for the renewal and cumulative processes, ... bandwidth selection, May 1990. (15 pages)

...: Regression estimators and nonparametric minimax efficiency, May 1990.

...: Law of large numbers for the bootstrap mean, June 1990. (12 pages)

...: Association schemes, orthogonal arrays and codes from non-degenerate metrics in finite projective geometries, June 1990. (9 pages)

... E.: Two-sample bootstrap tests: When to mix? July 1990. (17 pages)

... P.K.: Bootstrap confidence bands for the renewal function, July 1990.

...: Smoothing parameter selection in hazard rate estimation, Sept. 1990.

...: Generalized two sample Kolmogorov-Smirnov test involving a possibly censored sample, Sept. 1990. (10 pages)

2035 FAN, J.: Deconvolution with supersmooth distributions, Sept. 1990. (21 pages)

2036 FAN, J., TRUONG, Y. & WANG, Y.: Nonparametric function estimation involving error-in-variables, Sept. 1990. (15 pages)

2037 JI, C.: Sieve estimators for pair-interaction potentials and local characteristics in Gibbs random fields, Oct. 1990. (91 pages)

# LONG-RANGE DEPENDENCE AND GLOBAL WARMING

RICHARD L. SMITH

Department of Statistics,  
University of North Carolina,  
Chapel Hill N.C. 27599-3260, U.S.A

Presented at the SPRUCE conference, Lisbon, Portugal, April 1992; to appear in *Statistics in the Environment*, ed. V. Barnett and F. Turkman, to be published by John Wiley, Chichester.

The Library of the Department of Statistics  
North Carolina State University

## Summary

One of the scientific issues connected with “global warming” is the detection of trends in observed series of temperature data. This is important for providing evidence that global warming has actually occurred during the twentieth century, but also and perhaps more crucially as a means of validating general circulation models (GCM’s) which are the main means of predicting future temperatures rises that will occur if the emission of greenhouse gases is allowed to continue unhindered. However, the interpretation of trends in observed temperature data is complicated by numerous factors, including the fact that measured trends are different for different parts of the earth’s surface, and that they are different for different time periods. For example, global temperatures showed a steady rise from about 1910 to 1940, but there followed a period when temperatures were static or even decreasing until about 1975, since when a further rise has taken place. These kinds of effects raise the possibility that the observed temperature fluctuations are merely part of long-term natural variability and cannot be taken as indicative of any man-made activity.

In this paper, we address these issues from the point of view of time series with long-range dependence, i.e. a spectral density which is proportional to  $\omega^{1-2H}$  as  $\omega \rightarrow 0$  where  $\omega$  is frequency and the parameter  $H$  (the Hurst coefficient) is between 0.5 and 1. A method of estimating  $H$  directly from the periodogram is proposed and applied to the temperature data. The effect of such dependence on trend estimates is then analysed.

The methods are applied to a number of series, (a) the Central England temperature data of Manley (1974), (b) U.S.A. data from the Historical Climatological Network (HCN), and (c) global averages from the report of the Intergovernmental Panel on Climate Change (IPCC). The proposed method is compared with a more straightforward technique based on high-order autoregressive series. In all cases the standard error appears to be higher when calculated under the long-range dependence model than under an autoregressive model. However, the effect is different for different series, and greatest for the IPCC global data. The results for the Central England and U.S.A. data do not show significant global warming over the twentieth century. For the global data, the results are consistent with other statistical studies that have claimed that the global rise is significant, but they do confirm the importance of taking long-range dependence into account.

## 1. GLOBAL WARMING AND THE DETECTION OF TRENDS IN TEMPERATURE SERIES

I begin by reviewing some basic facts about global warming. These are taken from numerous publications in recent years of which a comprehensive summary is in the report of the Intergovernmental Panel on Climate Change (1990), henceforth IPCC. There is a natural greenhouse effect which keeps the Earth warmer than it otherwise would be. Human activities are substantially increasing the proportions in the atmosphere of a group of gases collectively known as greenhouse gases. These include carbon dioxide, methane, chlorofluorocarbons (CFC's) and nitrous oxide. Some gases are more effective than others at changing climate, but carbon dioxide is generally considered the most important single factor. For this reason, the level of carbon dioxide is widely used as a measure of the concentration of the greenhouse gases in the atmosphere.

A major scientific issue connected with global warming is to determine the magnitudes of the temperature effects resulting from an actual or hypothesised increase in greenhouse gases. To facilitate this, atmospheric scientists have created a number of mathematical models collectively known as general circulation models (GCM's). These try to take account of the many interactions between the physical and dynamical processes in the global climate. However, many of the descriptions of these processes are quite crude and consequently there is substantial uncertainty attached to the predictions of the GCM's.

The results of GCM's are often compared in terms of their predictions of the rise in temperature associated with a doubling of atmospheric carbon dioxide (the climate sensitivity). This is an artificial criterion — for example, it assumes equilibrium conditions, which certainly will not be attained for a very long time. Nevertheless, it serves for making comparisons. The range of results from current models for this scenario is from 1.9 to 5.2°C. Most results are close to 4.0°C. After taking into account all the model results and comparisons with observational evidence over the last century (which has tended to support values in the lower portion of this range), the IPCC adopted the value of 2.5°C as its best estimate of the climate sensitivity. When translated into actual predictions of global warming on the assumption that greenhouse gases continue to rise at the currently projected rate (the “business-as-usual scenario”), this leads to a prediction of an overall global warming of 1.8°C by 2030 and 4°C by 2100.

Observed temperature series thus play an important role in calibrating the GCM's. There is evidence of an actual rise in temperatures in the range 0.3-0.6°C over the last century. However, the rise has not been steady. Much of the warming since 1900 has been concentrated in two periods, the first between 1910 and 1940 and the second since 1975. During the period from 1940-1975 the Northern Hemisphere was actually cooling while the Southern Hemisphere remained at a constant temperature, while some regions have continued to cool until very recently. This substantial temporal and spatial variability shows that future temperature changes are likely to differ substantially from a global average. “Global mean temperature alone is an inadequate indicator of greenhouse-gas-induced climate change. Identifying the causes of any global mean temperature change

requires examination of other aspects of the changing climate, particularly its spatial and temporal characteristics.... Any changes to date could be masked by natural variability and other (possibly man-made) factors, and we do not have a clear picture of these." (IPCC 1990, page *xxix*.)

From these remarks we can see that a major *statistical* issue connected with global warming is the separation of effects due to natural variability, such as may for example be represented by a stationary time series model, and those that are due to a trend in the data. In this paper I consider particularly how these issues will work out in the case of *long-range dependence*, which may be represented by a power-law spectral density proportional to  $\omega^{1-2H}$  as  $\omega \rightarrow 0$  where  $\omega$  is frequency and the parameter  $H$  (known as the Hurst coefficient after the pioneering study of the River Nile by the English physicist H.E. Hurst in the 1950's) is between 0.5 and 1. The correlations from such a process decrease with lag  $k$  at a rate  $k^{2H-2}$ , i.e. a polynomial decrease in  $k$  compared with the exponential decrease that occurs with more familiar processes such as the ARMA family.

In the present paper I discuss methods for estimating  $H$  based on the spectral density (to be developed in more theoretical detail in the forthcoming paper of Samarov, Smith and Taqqu (1992)), and consider the effects of long-range dependence on the estimation of a linear trend.

## 2. LONG-RANGE DEPENDENCE AND THE ESTIMATION OF A LINEAR TREND.

In this section I review some basic relations among the autocovariances, spectral density and variances of sample means of a stationary time series, and derive an extension to the asymptotic variance of the least-squares estimator of the slope of a linear regression.

Let  $\{Y_n, n = 1, \dots, N\}$  denote a sample of  $N$  observations from a stationary time series with mean 0. Let  $\gamma_k = E\{Y_n Y_{n+k}\}$  denote the  $k$ 'th autocovariance and  $f(\omega)$ ,  $-\pi < \omega < \pi$  the spectral density. The relations between the autocovariances and the spectral density are

$$f(\omega) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \gamma_n e^{-in\omega} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} \gamma_n \cos(n\omega), \quad (2.1)$$

$$\gamma_n = \int_{-\pi}^{\pi} e^{in\omega} f(\omega) d\omega = \frac{1}{2} \int_0^{\pi} \cos(n\omega) f(\omega) d\omega. \quad (2.2)$$

The spectral density at a particular frequency  $\omega$  may be estimated by means of the periodogram

$$I_N(\omega) = \frac{2}{N} \left| \sum_{n=1}^N Y_n e^{in\omega} \right|^2 = \frac{2}{N} \left[ \sum_{n=1}^N Y_n^2 + 2 \sum_{k=1}^{N-1} \sum_{m=1}^{N-k} Y_m Y_{m+k} \cos(k\omega) \right] \quad (2.3)$$

whose mean

$$2\gamma_0 + 4 \sum_{k=1}^{N-1} \left(1 - \frac{k}{N}\right) \gamma_k \cos(k\omega)$$

converges as  $N \rightarrow \infty$  to

$$2\gamma_0 + 4 \sum_{k=1}^{\infty} \gamma_k \cos(k\omega) = 4\pi f(\omega). \quad (2.4)$$

The formulae (2.1)-(2.4) have been spelt out because there seems to be no unanimity about the definition of either the spectral density or the periodogram, published formulae often differing from the above by factors of  $2\pi$  or  $4\pi$ . They are, however, consistent with those in Olshen (1967) and Cox (1984), and we take these as our starting point.

Suppose now we have long-range dependence, which we may define by either of the relations

$$\gamma_k \sim ak^{2H-2}, \quad k \rightarrow \infty, \quad (2.5)$$

$$f(\omega) \sim b\omega^{1-2H}, \quad \omega \downarrow 0 \quad (2.6)$$

where  $\frac{1}{2} < H < 1$ . The relations (2.5) and (2.6) are also meaningful for other values of  $H$  outside this range, but it is for these values of the parameter  $H$  that a process is stationary with long-range dependence and therefore on these that we concentrate.

The relation between the positive constants  $a$  and  $b$  is

$$b = \frac{a}{\pi} \Gamma(2H - 1) \sin(\pi - \pi H) \quad (2.7)$$

which may be derived from the asymptotic relationships

$$\sum_{n=1}^{\infty} n^{-\beta} \cos(nx) \sim x^{\beta-1} \Gamma(1 - \beta) \sin(\pi\beta/2),$$

$$\sum_{n=1}^{\infty} n^{-\beta} \sin(nx) \sim x^{\beta-1} \Gamma(1 - \beta) \cos(\pi\beta/2),$$

both valid as  $x \downarrow 0$  for  $0 < \beta < 1$  (Zygmund 1959, Section V.2). See also Beran (1989), p. 261-2, after correcting for a factor  $2\pi$ .

One problem that has been studied in this area is the effect of long-range dependence on the estimation of a sample mean  $\bar{Y}_N = (1/N) \sum_{n=1}^N Y_n$ . We have

$$\text{Var } \bar{Y}_N = \frac{\gamma_0}{N} + \frac{2}{N} \sum_{k=1}^{N-1} \left(1 - \frac{k}{N}\right) \gamma_k.$$

Under (2.5) we have, for large  $N$ ,

$$\begin{aligned} \sum_{k=1}^N \left(1 - \frac{k}{N}\right) \gamma_k &\sim \frac{a}{N} \int_0^N (N-x)x^{2H-2} dx \\ &= \frac{aN^{2H-1}}{2H(2H-1)} \end{aligned}$$

from which

$$\text{Var } \bar{Y}_N \sim \frac{a}{H(2H-1)} N^{2H-2} = \frac{b\pi}{H\Gamma(2H)\sin(\pi-\pi H)} N^{2H-2}. \quad (2.8)$$

This equation is the same as the formulae on pages 192 and 194 of Samarov and Taqqu (1988), and in a different form with Beran (1989).

In the present paper our focus is on an extension of this idea, to the case when  $\{Y_n\}$  is regressed on a covariate  $\{x_n\}$ . The ordinary least squares estimator of the slope  $\beta$  is

$$\hat{\beta} = \frac{\sum Y_n(x_n - \bar{x}_N)}{\sum (x_n - \bar{x}_N)^2}.$$

For simplicity we shall assume  $\bar{x}_N = 0$ . The variance of  $\hat{\beta}$  is then

$$\left(\sum_{n=1}^N x_n^2\right)^{-2} \left\{ \gamma_0 \sum_{n=1}^N x_n^2 + 2 \sum_{k=1}^{N-1} \gamma_k \sum_{m=1}^{N-k} x_m x_{m+k} \right\}.$$

Let us specialize to the case of a linear trend,  $x_n = n - (N+1)/2$ ,  $1 \leq n \leq N$ . We then approximate

$$\sum x_n^2 \sim \int_0^N \left(x - \frac{N}{2}\right)^2 dx = \frac{N^3}{12}$$

and

$$\sum_{k=1}^{N-1} \gamma_k \sum_{m=1}^{N-k} x_m x_{m+k} \sim a \int_0^N x^{2H-2} \int_0^{N-x} \left(y - \frac{N}{2}\right) \left(x + y - \frac{N}{2}\right) dy dx.$$

After some manipulation the right-hand integral can be shown to equal

$$\frac{a(1-H)}{8H(1+H)(2H-1)} N^{2H+2}$$

and it follows that

$$\text{Var } \hat{\beta} \sim \frac{36a(1-H)}{H(1+H)(2H-1)} N^{2H-4} = \frac{36b\pi(1-H)}{H(1+H)\Gamma(2H)\sin(\pi-\pi H)} N^{2H-4}. \quad (2.9)$$

The formulae (2.8) and (2.9) may also be written in the form

$$\text{Var } \bar{Y}_N \sim \theta_1 b N^{2H-2}, \quad \text{Var } \hat{\beta} \sim \theta_2 b N^{2H-4} \quad (2.10)$$

where the constants  $\theta_1$  and  $\theta_2$  are given (as a function of  $H$ ) in the following table:

Table 1: Variation of $\theta_1$ and $\theta_2$ with $H$		
$H$	$\theta_1$	$\theta_2$
0.5	6.28	75.40
0.55	6.08	63.54
0.6	6.00	53.97
0.65	6.04	46.16
0.7	6.25	39.72
0.75	6.68	34.38
0.8	7.48	29.91
0.85	8.96	26.15
0.9	12.13	22.98
0.95	21.98	20.29
1.0		18.00
1.1		14.38
1.2		11.74
1.3		9.81
1.4		8.44
1.5		7.54

Note that, although the value of  $\theta_1$  becomes infinite as  $H \rightarrow 1$ , the value of  $\theta_2$  remains well-defined for  $H \geq 1$ . It is not clear whether the result is meaningful in this case — the derivation just given is not then valid, since the process is non-stationary and neither the autocovariances or the spectral density are well defined.

The preceding analysis is a special case of a more general result due to Yajima (1988) in the case of polynomial regression and Yajima (1991) for general regressors. In particular it can be shown, with some work, that the formula (2.9) is a special case of Theorem 2.2 of Yajima (1988). The derivation given here has the advantage of being a direct calculation. Yajima (1988) also considered the efficiency of the least squares estimate as compared with the best linear unbiased estimate (BLUE). I have recomputed his Table 1 to consider the asymptotic efficiency (expressed as a ratio of variances of the BLUE to the LSE) of  $\hat{\beta}$ . I obtain efficiencies of 1, 0.991, 0.969, 0.942, 0.915 and 0.889 for  $H = 0.5$  to 1 in increments of



0.1 (corresponding to  $d = 0, 0.1, \dots, 0.5$  in Yajima's table). These seem to me to confirm that the efficiency is near 1 so that there is no need to do the extra work required to compute the BLUE. On the other hand, taking the opposite viewpoint, these figures also quantify the potential benefit of a fully model-based approach against the approach adopted in this paper.

### 3. CENTRAL ENGLAND TEMPERATURE DATA.

In this section we consider a monthly series of mean temperatures in central England 1659-1973, collected by Manley (1974), and updated to 1989. It should be pointed out that there are numerous sources of possible nonstationarity due to changes in measurement methods, and in particular, Manley drew attention to the fact that the data values before 1723 were much less reliable than those following. However, in this analysis we consider the original data set as reported by Manley.

Figure 1 shows a plot of the whole data set. These are smoothed by calculating a 49-month centred moving average; it is this moving average that is plotted. The most obvious feature is the remarkable dip in temperature around 1700. In the light of the reservations just mentioned, there must be some doubt about the accuracy of the precise values, though presumably not about the fact that there was a spell of very cold weather at that time. At the opposite end, there is some suggestion of a temperature rise over the period 1920-1960, but the evidence for that is not clear-cut.

Figure 2 shows *monthly* values from 1907-1973. In this case the seasonal effect was eliminated by subtracting, from the data for each month of the year, the overall average for that month. There is no visual evidence of any trend.

In the spirit of Section 1 we can consider a model of the form

$$Y_n = \alpha + \beta x_n + \eta_n \quad (3.1)$$

where  $Y_n$  is the observed value in month  $n$  and  $x_n = (n - 402.5)/1200$ . The value 402.5 was subtracted from  $n$  so that the  $N = 804$  values of  $x_n$  are centered about 0, and the result was divided by 1200 to express it as a change in temperature per century (the most widely used unit). The residual  $\eta_n$  may be taken as a stationary time series. A natural short-range dependent model for  $\{\eta_n\}$  is the AR( $p$ ) model:

$$\eta_n = \sum_{j=1}^p a_j \eta_{n-j} + \epsilon_n \quad (3.2)$$

where  $\{\epsilon_n\}$  is a white noise sequence which we take to be independent  $N(0, \sigma^2)$  for unknown  $\sigma^2$ . The simplest way to fit the model defined by (3.1) and (3.2) is by conditional maximum likelihood: for some  $p_0 \geq p$  use equations (3.1) and (3.2) to express  $\{\epsilon_n, n > p_0\}$  as a function of the data and parameter values, and then minimize  $D_{\bullet} = \sum_{j=p_0+1}^N (\epsilon_j/\sigma)^2 +$

$(N - p_0) \log(\sigma^2)$  where  $N$  is total sample size and  $D$  is twice the negative log likelihood, or deviance. In the analysis to follow I have considered values of  $p$  up to 10 and  $p_0 = 10$ , the value of  $p_0$  being fixed so as to make direct comparison between fits for different values of  $p$ . This method is called conditional maximum likelihood because the likelihood being calculated is the conditional distribution of the last  $N - p_0$  data values given the first  $p_0$ .

A simple estimate assuming independence (i.e.  $p = 0$ ) produced a value  $\hat{\beta} = 0.275$  ( $^{\circ}\text{C}$  per century) with standard error 0.17. This is reasonably consistent with the global warming estimate of 0.3-0.6, but the standard error casts considerable doubt on the significance of this. This is confirmed when higher value of  $p$  are fitted: for  $p = 1, 2, 3$  the estimates of  $\beta$  are 0.274, 0.274 and 0.273 to three decimal places, with standard errors 0.22, 0.22, 0.23. There is a sharp drop in deviance when passing from AR(0) to AR(1), but only a small drop thereafter, and the Akaike Information Criterion selects  $p = 1$  as the most suitable value.

#### 4. LONG-RANGE DEPENDENCE IN ENGLISH DATA.

We now consider the effect of possible long-range dependence on these data. For the purpose of the present section, it was considered desirable to analyse as long a series as possible. Therefore the full data set (up to 1973) was used without any allowance for possible nonstationarity. In further analysis, it would be preferable to consider the robustness of the following analysis against various forms of departure from stationarity.

The first step was to form the periodogram (2.3). This was evaluated at the Fourier frequencies  $\{\omega_j = 2\pi j/N, 1 \leq j \leq N/2\}$ . In Figure 3, the first 500 log periodogram ordinates are plotted against frequency, and in Figure 4 against log frequency. There is a prominent spike at frequency  $1/12$  (the annual seasonal effect) but no other indication of deterministic seasonality. In addition, the behavior of the log periodogram near frequency 0 does suggest long-range dependence.

According to well-known theory,  $I_n(\omega_j)$  for  $j = 1, 2, \dots$  are approximately independent exponential random variables with means  $4\pi f(\omega_j)$ . When combined with the assumption (2.6) this suggests a regression of  $\log I_n(\omega_j)$  on  $\log \omega_j$  for  $1 \leq j \leq n$ , where  $n \ll N$ . Such a procedure was proposed by Geweke and Porter-Hudak (1983). Subsequent work (Samarov, Smith and Taqqu 1992) has suggested the following refinements of the Geweke-Porter-Hudak procedure:

(a) Geweke and Porter-Hudak suggested fitting the regression by least-squares. However, the approximate exponentiality of periodogram ordinates suggests that a superior method would be a maximum likelihood fit based on the Gumbel error distribution function  $1 - \exp(-e^x)$ , which arises from the logarithm of a standard exponential random variable.

(b) The value of  $n$  needs to be much smaller than  $N$  and carefully chosen to balance bias and variance.

(c) In addition, it may be desirable to restrict the regression to Fourier frequencies  $m \leq j \leq n$  with  $m > 1$ . This was suggested by Figure 5 and the accompanying discussion in Haslett and Raftery (1989), in which the linearity of the plot of log periodogram against log frequency broke down in the very smallest ordinates.

As an indication of the sensitivity of the results to  $m$  and  $n$ , we show both the least squares and maximum likelihood estimates for the data under discussion, for a variety of values of  $m$  and  $n$ .

First, let us arbitrarily fix  $n = 100$  and try several values of  $m$ . The tabulated parameter is the slope  $1 - 2H$ .

Table 2: Slope estimates and standard errors for fixed  $n$ .

$m$	$n$	LS Est	Std. Err.	ML Est	Std. Err.
1	100	-.58	.13	-.52	.10
2	100	-.62	.14	-.55	.11
3	100	-.59	.15	-.54	.12
4	100	-.51	.16	-.43	.13
5	100	-.48	.17	-.41	.14
6	100	-.47	.18	-.41	.15
7	100	-.49	.19	-.44	.16
8	100	-.51	.20	-.46	.16
9	100	-.49	.20	-.47	.17
10	100	-.46	.21	-.47	.17

It can be seen that the values tend to settle down from  $m = 4$  onwards, suggesting that the first three Fourier frequencies are high-leverage points which have a distorting effect on the results. Similar behaviour has been observed in a number of other series to which this technique has been applied. The difference between the maximum likelihood and least squares estimates reflects, largely, the effect of a few very small periodogram ordinates which are downweighted by a Gumbel fit.

Henceforth the value  $m = 5$  was fixed, and a number of values of  $n$  tried (see Table 3 to follow).

In this case there seems a steady drift in slope as  $n$  increases, with no indication of where to draw the line.

However, it is also possible to fit an extension of the model (2.6) to include a next-order term:

$$\log f(\omega) = \log b + (1 - 2H) \log \omega + c\omega^2. \quad (4.1)$$

With intelligent application, this can provide further indication of the correct choice of  $n$ . In this case, fitting (4.1) by maximum likelihood with  $n = 200, 250$  and  $300$  respectively produced estimates of  $1 - 2H$  of  $-.34, -.32, -.30$  with standard errors  $0.17, 0.15, 0.07$ . Given that many of the maximum likelihood estimates in the above table are around these values, there seems to be strong evidence that the true value of  $1 - 2H$  is around  $-0.30$ . i.e.  $H$  is about  $0.65$ .

Table 3: Slope estimates and standard errors for fixed  $m$ .

$m$	$n$	LS Est	Std. Err.	ML Est	Std. Err.
5	50	-.55	.25	-.54	.26
5	60	-.49	.21	-.50	.21
5	70	-.52	.20	-.44	.19
5	80	-.59	.20	-.48	.18
5	90	-.49	.18	-.46	.15
5	100	-.48	.17	-.41	.15
5	110	-.42	.16	-.32	.13
5	120	-.53	.18	-.34	.14
5	130	-.50	.17	-.36	.13
5	140	-.47	.16	-.33	.12
5	150	-.34	.15	-.23	.11
5	160	-.36	.14	-.26	.10
5	170	-.33	.14	-.25	.09
5	180	-.33	.13	-.27	.09
5	190	-.28	.12	-.25	.09
5	200	-.29	.12	-.27	.09
5	250	-.31	.10	-.31	.08

Fixing  $n = 250$  and again trying different  $m$  produced the following:

Table 4: Slope estimates and standard errors for  $n = 250$ .

$m$	$n$	LS Est	Std. Err.	ML Est	Std. Err.
1	250	-.39	.09	-.39	.06
2	250	-.39	.09	-.40	.06
3	250	-.37	.10	-.39	.07
4	250	-.33	.10	-.32	.07
5	250	-.31	.10	-.31	.08
6	250	-.30	.11	-.31	.08
7	250	-.30	.11	-.32	.08

Again there is stability for  $m \geq 4$ .

The final estimates used  $m = 5$ ,  $n = 250$ , leaving  $1 - 2\hat{H} = -0.31$ , and hence  $\hat{H} = 0.655$ , and also  $\hat{b} = 0.035$ , after correcting for a factor  $4\pi$  to allow for the fact that the periodogram estimates  $4\pi f(\omega)$  rather than  $f(\omega)$  itself.

To conclude this section, we discuss the effect of long-range dependence on the trend estimates reported in Section 2. Note that this assumes that the model just fitted for the whole data series is also applicable to the 1907-1973 data series analyzed in Section 2. For the reasons outlined in the introduction, this may not be a valid assumption. For  $H = 0.655$  we have  $\theta_2 = 45.5$ , and hence for  $N = 800$  (the actual number of data points used in Section 2) the approximate variance in (2.10) becomes  $2.47 \times 10^{-8}$ , or standard deviation  $1.57 \times 10^{-4}$ . When multiplied by 1200 to be consistent with the correct units, the estimated standard error of the parameter  $\beta$  becomes 0.19, comparable with the earlier calculation of around 0.22.

By analogy with the method of Beran (1989), it should be possible to “studentize” the estimate of  $\beta$  by taking into account the distribution of  $(\hat{H}, \hat{b})$ . This will presumably increase the effective standard error of  $\hat{\beta}$  still further.

## 5. U.S.A. DATA

The U.S.A. data are part of the Historical Climatological Network (HCN) and comprise daily maximum and minimum temperatures over a network of 95 stations over the continental U.S.A. Most of the series started from between 1890 and 1930 and contain some missing values. For the purpose of this analysis, the following were computed for each station:

- (i) Monthly average daily averages (note that the daily average is *defined* to mean the average of the daily maximum and daily minimum values),
- (ii) Monthly average daily maxima,
- (iii) Monthly average daily minima,
- (iv) Monthly maximum daily maxima (i.e. the absolute maximum temperature for each month),
- (v) Monthly minimum daily minima.

All five series were expressed as *anomalies* from the monthly averages of 1950-1979, in other words, for each of the 12 months of the year, the 1950-1979 average for that month was subtracted from the observations. This is standard practice in analyzing climate series and has the advantage of removing overall differences due to different mean temperatures at different locations, and also removing the dominant seasonal effect, though the reader should be cautioned that neither spatial or seasonal effects are eliminated by the crude

device of subtracting off the means! However, in the following analysis we shall proceed as if they are, and shall consider in particular the U.S. average series formed by averages each of the series corresponding to (i)-(v) over the 95 stations.

Two recent papers in particular have drawn attention to fresh issues of global warming based on analysis of (different portions of) the same data set. Karl, Heim and Quayle (1990) fitted linear trends with autoregressive errors to the data, and then used simulation to assess the standard errors of the estimated trend coefficients. They concluded that trend estimates based on average temperatures, though consistent with those produced by IPCC, were not significant as judged by the standard errors. They pointed out that this could have serious implications for public policy: if the predictions of the GCM's are correct, then according to their calculations it will be another 30 years or so before the observed temperature changes are sufficient to be judged significant using statistical techniques (by which time, by implication, it may be too late to do anything about it). However, they also suggested that the conclusion may be different when applied to maximum or minimum temperatures. In further development, Karl, Kukla *et al.* (1991) suggested specifically that daily minima may be rising, while daily maxima remain roughly constant. This is consistent with a rough notion of the greenhouse effect acting to prevent nighttime cooling while not having any particular influence on daytime temperatures. This approach makes clear the need to analyse the entire distribution of temperatures, as opposed to just mean values, which has been the main focus of attention in the global warming literature to date.

In Figure 5, 49-month moving averages are plotted for each of the five series. All of these show the same pattern of a sharp rise between 1920 and 1940, followed by a steady decline up to 1975 and a recent rise. Visual inspection of the plots suggests that the top two (corresponding to series (v) and (iii)) do indeed show the greatest rate of warming since 1975 and the smallest rate of cooling during 1940-1975, consistent with Karl, Kukla *et al.* (1991).

The autoregressive model of Section 2 was fitted to each of the five series for the period 1895-1987. In each case, inspection of log likelihoods suggests that either an AR(1) or an AR(2) model is a perfectly good fit to the data — for consistency, we shall take AR(2) in each case. For these data sets, the estimated linear trends (with standard errors in parentheses) are 0.63 (0.22), 0.81 (0.25), 0.42 (0.21), 0.21 (0.27) and 0.87 (0.30) respectively for the series (i)-(v). The overall rise, represented by the first figure, does appear to be significant, but the results also suggest that the sharpest rise is in the minimum daily minima, consistent with the hypothesis of Karl, Kukla *et al.* (1991).

Let us now consider the effect of long-range dependence in this case. For brevity we only consider the overall mean series (i). The periodogram is plotted against frequency in Figure 3 and against log frequency in Figure 4. From the latter it is hard to see any evidence of long-range dependence (we are looking for a negative slope on the left-hand side of the plot) but if this does happen at all, it seems to be on the log frequency range up to about 0.3, which is at about the 120'th Fourier frequency. In the notation and

terminology of Section 3, the estimate was the ML estimate with  $m = 1$ ,  $n = 120$ . Note that the detailed exploration of Section 3, to select the best values of  $m$  and  $n$ , was not repeated here, though a brief study of alternatives showed no great sensitivity to the precise choice of these parameters. On the basis of this we deduce a slope in the log-log periodogram of  $-0.1452$  (standard error 0.08), which suggests that there is indeed some evidence for long-range dependence, though not so strong as to be definitely identified by a significance test. This leads to an estimate of  $H = 0.5726$ , and also of  $b = 0.5704$ . After applying (2.10) with  $\theta_2 = 59$ ,  $n = 1116$ , we deduce a standard error of the slope of the regression to be about 0.31. The standard error in this case is somewhat larger than that obtained from the autoregressive fit (0.22), and serves to reinforce doubts about the significance of the estimated trend.

As things stand, it is my view that there are too many uncertainties in this analysis to make any definitive conclusions. The most obvious drawback of this analysis is that the trend is clearly not linear (see Figure 5). Indeed, repeating the same analysis for the years 1920-1987 produces evidence of an overall cooling in all five series! However, in the spirit of Karl, Heim and Quayle (1991), the analysis does permit discussion of what rises would need to be observed to be judged statistically significant. For this, it is relevant that all the standard errors quoted are of the same order of magnitude as the claimed overall greenhouse effect of the IPCC estimates.

## 6. GLOBAL DATA

Having failed to detect a truly significant trend in either the English or the U.S. data, we now consider some of the IPCC data, divided into (i) mean overall (land and sea) temperatures for the Northern Hemisphere 1854-1989, (ii) mean overall (land and sea) temperatures for the Southern Hemisphere 1854-1989, (iii) mean land temperatures for the Northern Hemisphere 1851-1989, (iv) mean land temperatures for the Southern Hemisphere 1851-1989. The four data sets, again smoothed by a 49-month moving average, are shown in Figure 6. These confirm the general pattern described at the beginning of the paper, i.e. a sharp rise in temperature between 1920 and 1940, followed by a levelling off or even a cooling up to 1975, and then a further rise. The pattern is not quite the same in all four series. In the following analysis we concentrate on series (i).

In this case the attempt to fit an autoregressive model is not so successful. Models of order up to 20 were fitted with no clear identification of a model. However, all the fitted models showed an estimated trend of about 0.40 with a standard error of 0.02-0.03, which ought to put the significance of the estimated trend beyond doubt. Consider, however, the periodogram in Figures 3 and 4. In this case the evidence for long-range dependence seems clear, and in Figure 4 this appears to persist up to a log frequency of about  $-0.8$ , which is again near the 120'th Fourier frequency. Based on  $m = 1$ ,  $n = 120$  and a ML fit, the estimated slope of the log-log periodogram plot is  $-0.806$  (standard error .11), corresponding to  $H = 0.903$ , and also an estimate of  $b$  of 0.0033. In this case we have  $\theta_2 = 22.8$  (and  $n = 1632$ ), which leads via (2.10) to an estimated standard error of 0.10 for

the rate of global warming in °C per 100 years. In this case the choice of  $m$  and  $n$  seems very highly subjective, but other values lead to similar standard errors for rate of global warming even if rather different values of  $H$ . For example  $m = 4$ ,  $n = 100$  yields a slope of  $-0.751$  (standard error 0.15), which corresponds to  $H = 0.875$  and  $\theta_2 = 24.5$ . In this case we also have the estimate of  $b = 0.0038$  which leads to a standard error of .09 for rate of global warming. Analysis with  $m = 5$ ,  $n = 50$  leads to a final standard error of .07 via the estimate  $H = 0.79$ .

Although these results are somewhat sensitive to the choice of  $m$  and  $n$ , the following points do seem clear: (i) there is significant evidence of long-range dependence (i.e. that  $H > 0.5$ ), and (ii) the standard error is well in excess of its value computed under an autoregressive model. The overall conclusion, that there is significant evidence of global warming, is unchanged by this analysis, though the increase in standard error does show that this is more in doubt than previously.

Bloomfield (1992) considered the use of autoregressive and long-range-dependent models for the IPCC temperature series studied here, and also for the Hansen and Lebedeff (1987, 1988) data set. In this case autoregressive models were fitted to *annual* data, the best order being claimed as  $p = 4$  for the IPCC data and  $p = 8$  for the Hansen and Lebedeff data. Bloomfield also fitted the fractional ARIMA model of Hosking (1981), which is a parametric model satisfying (2.6). Using this model, also, there was very clear evidence of long-range dependence. For the IPCC data, the estimated rates of global warming were around  $0.4^\circ\text{C}$  per century, with a standard error of about 0.1, which is consistent with our results. The somewhat shorter Hansen and Lebedeff data produced a warming estimate of  $0.6^\circ\text{C}$  per century with a standard error again around 0.1 - the difference between the two series being partly explainable in terms of the differences in period covered. Haslett and Raftery (1989) contained a very detailed discussion of fractional ARIMA models. A more recent paper by Bloomfield and Nychka (1992) has compared a number of different time series approaches with particular regard for the spectrum of the series and its effect on the uncertainties of estimated trend. Yet another approach was taken in a paper of Kuo, Lindberg and Thomson (1990), who used frequency-domain nonparametric methods for the estimation of trend. For the Hansen-Lebedeff data, they again obtained estimates of trend and its standard error which were comparable with those of Bloomfield.

## 7. DISCUSSION

The conclusions of this paper are necessarily highly tentative. There remain doubts about both the reasonableness of fitting linear trends to series with such erratic behaviour, and about the details of the proposed estimation. My main purpose has been to present the method as an alternative to the ones currently available for assessing the uncertainties about global warming as estimated from observational data. It does seem clear, however, that long-range dependence is a legitimate cause for concern, and that the standard errors obtained under an assumption of long-range dependence can be considerably larger than



those obtained under an autoregressive model. For this reason, I do not believe that the autoregressive model provides an acceptable method for assessing these uncertainties.

The comparison with the fractional differencing method, mentioned briefly at the end of Section 6 but really well established as a method since the paper of Haslett and Raftery (1989), is more problematical. The fractional differencing model is a finite-parameter model whose estimation procedures via maximum likelihood are by now well understood, whereas the method proposed here still leaves open many theoretical questions. On the other hand, there are advantages to a nonparametric method. One can argue that since the principal formula (2.9), for the asymptotic variance of the least squares estimator, depends only on the lower tail of the spectral density, then for reasons of parsimony the method of estimation should also depend only on the lower tail, a property not satisfied by the fractional ARIMA approach. In other words, the method given here may be more robust. This is to be judged against the efficiency of the LSE as compared with the BLUE (the latter representing what is possible when the assumed ARIMA model is correct) that was mentioned at the end of Section 2.

The issues raised by Karl, Kupka *et al.* (1991) have drawn attention to the need to consider trends in the whole distribution rather than just the mean values. This is intended to be a focus of future work. Particular emphasis will be given to extremal properties of the distributions.

The analysis based on spatial averages also has the disadvantages of not allowing any modelling of spatial features, such as spatial correlations or the undoubtedly true fact that the rate of global warming is different in different locations. This raises the possibility of a comprehensive spatial model. A model combining spatial features with long-range dependence, within the fractional differencing framework, was proposed for wind speed data by Haslett and Raftery (1989), and this is also expected to be a focus of future work.

## ACKNOWLEDGEMENTS

For the data sources I thank: Ms. T. Bishop, formerly of the University of Surrey (Manley data), Dr. T.R. Karl of the National Climatic Data Center, Asheville, N.C. (U.S.A. data) and Dr. P.D. Jones of the Climatic Research Unit, University of East Anglia (global data). Ms. A. Grady helped with data processing. I also thank Peter J. Robinson and Peter Bloomfield for their remarks. The work was partly supported by the NSF under contract DMS-9115750 as part of the Mathematics-Geosciences initiative (jointly with P.J. Robinson).

## REFERENCES

Beran, J. (1989), A test of location for data with slowly decaying serial correlations. *Biometrika* **76**, 261-269.

- Bloomfield, P. (1992), Trends in global temperature. *Climatic Change* 21, 1-16.
- Bloomfield, P. and Nychka, D. (1992), Climate spectra and detecting climate change. *Climatic Change*, to appear.
- Geweke, J. and Porter-Hudak, S. (1983), The estimation and application of long-memory time series models. *J. Time Series Anal.* 4, 221-238.
- Hansen, J. and Lebedeff, S. (1987), Global trends of measured surface air temperature. *J. Geophys. Res.* 92, 13345-13372.
- Hansen, J. and Lebedeff, S. (1988), Global surface air temperatures: Update through 1987. *Geophys. Res. Lett.* 15, 323-326.
- Haslett, J. and Raftery, A.E. (1989), Space-time modelling with long-memory dependence: assessing Ireland's wind power resource (with discussion). *Applied Statistics* 38, 1-50.
- Hosking, J.R.M. (1981), Fractional differencing. *Biometrika* 68, 165-176.
- Intergovernmental Panel on Climate Change (1990), *Climate Change: The IPCC Scientific Assessment*. Cambridge University Press, Cambridge.
- Karl, T.R., Heim, R.R. and Quayle, R.G. (1991), The greenhouse effect in central North America: If not now, when? *Science* 251, 1058-1061.
- Karl, T.R., Kukla, G., Razuvayev, V.N., Changery, M.J., Quayle, R.G., Heim, R.R. Easterling, D.R. and Fu, C.B. (1991), Global warming: evidence for asymmetric diurnal temperature change. *Geophysical Research Letters* 18, 2253-2256.
- Kuo, C., Lindberg, C. and Thomson, D.J. (1990), Coherence established between atmospheric carbon dioxide and global temperature. *Nature* 343, 709-714.
- Manley, G. (1974), Central England temperatures: monthly means 1659 to 1973. *Quart. J. R. Met. Soc.* 100, 389-405.
- Samarov, A. and Taqqu, M. (1988), On the efficiency of the sample mean in long-memory noise. *Journal of Time Series Analysis* 9, 191-200.
- Samarov, A., Smith, R.L. and Taqqu, M.S. (1992), Estimating the intensity of long-range dependence through regression. Forthcoming.
- Yajima, Y. (1988), On estimation of a regression model with long-memory stationary errors. *Ann. Statist.* 16, 791-807.
- Yajima, Y. (1991), Asymptotic properties of the LSE in a regression model with long-memory stationary errors. *Ann. Statist.* 19, 158-177.

## Figure Captions

*Figure 1.* Smoothed plot of annual mean temperatures for Central England data series ( $^{\circ}\text{C}$ ) 1659-1989.

*Figure 2.* Plot of monthly mean temperatures for Central England data series expressed as deviations from overall monthly means ( $^{\circ}\text{C}$ ), 1907-1973.

*Figure 3.* Log periodogram vs. frequency for Central England series (top), USA (middle), global (bottom).

*Figure 4.* Log periodogram vs. log frequency for Central England series (top), USA (middle), global (bottom).

*Figure 5.* USA average temperature series expressed as anomalies from monthly means 1950-1979. The five plotted series are all based on 49-month moving averages. Top to bottom: monthly minima of daily minima, monthly averages of daily minima, monthly averages of daily averages, monthly averages of daily maxima, monthly maxima of daily maxima.

*Figure 6.* Global temperatures relative to 1950-1979 monthly averages. Plotted are 49-month moving averages. Top to bottom: Land and sea averages, Northern hemisphere, 1854-1989; Land and sea averages, Southern hemisphere, 1854-1989; Land only averages, Northern hemisphere, 1851-1989; Land only averages, Southern hemisphere, 1851-1989.

Figure 1: England Data 1659–1989

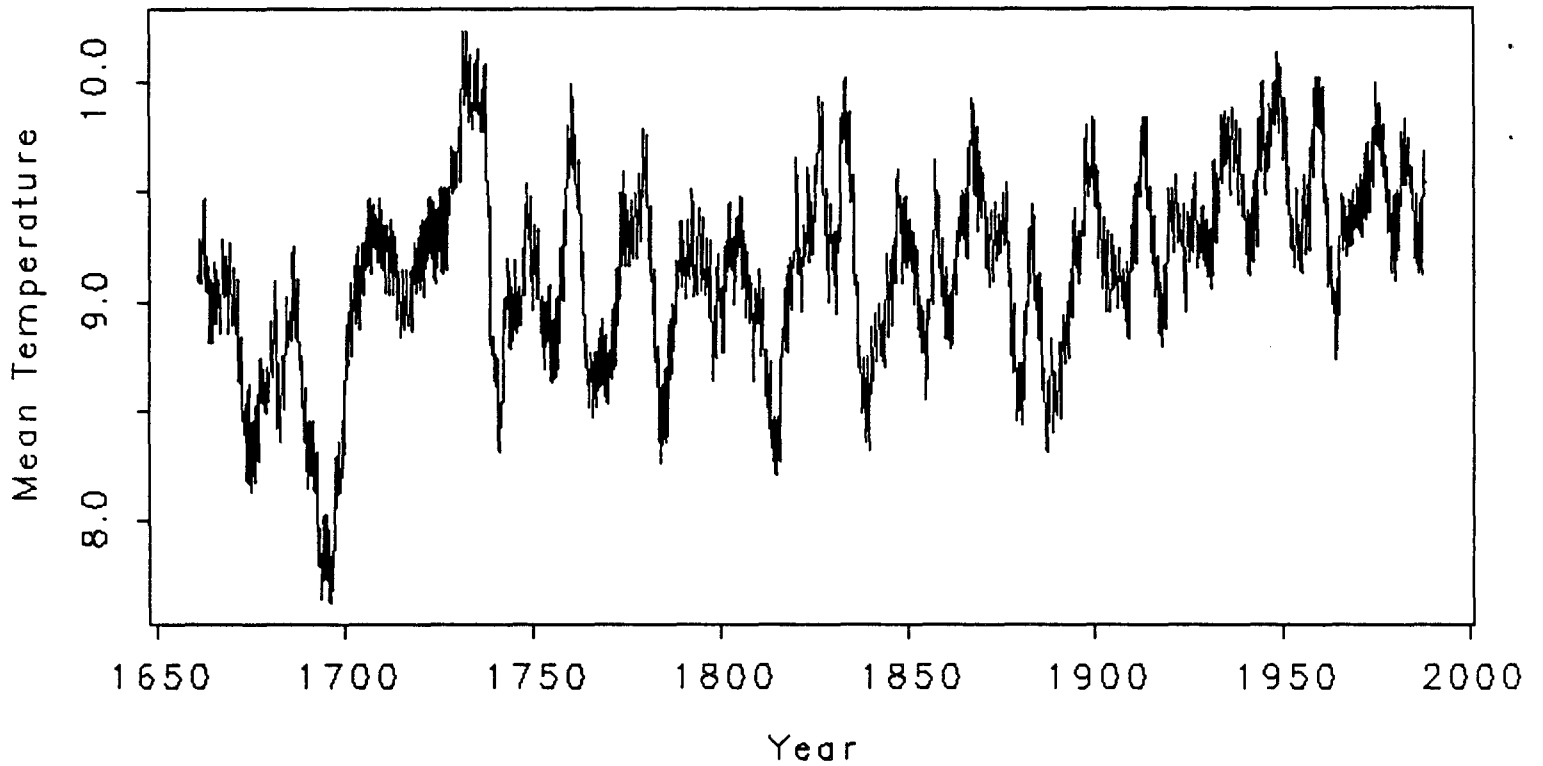


Figure 2: Monthly Data 1907–1973

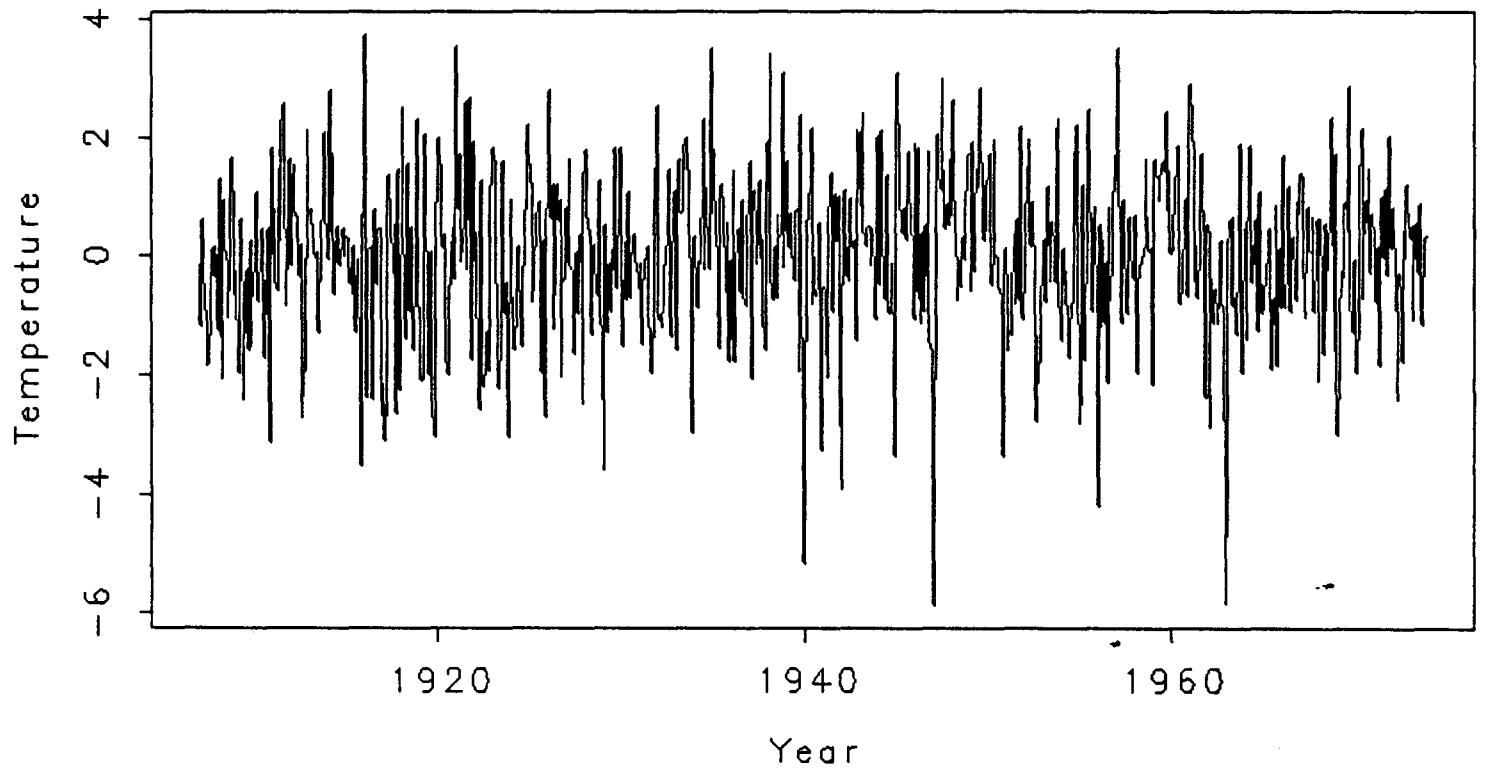


Figure 3: Periodograms of Temperature Series

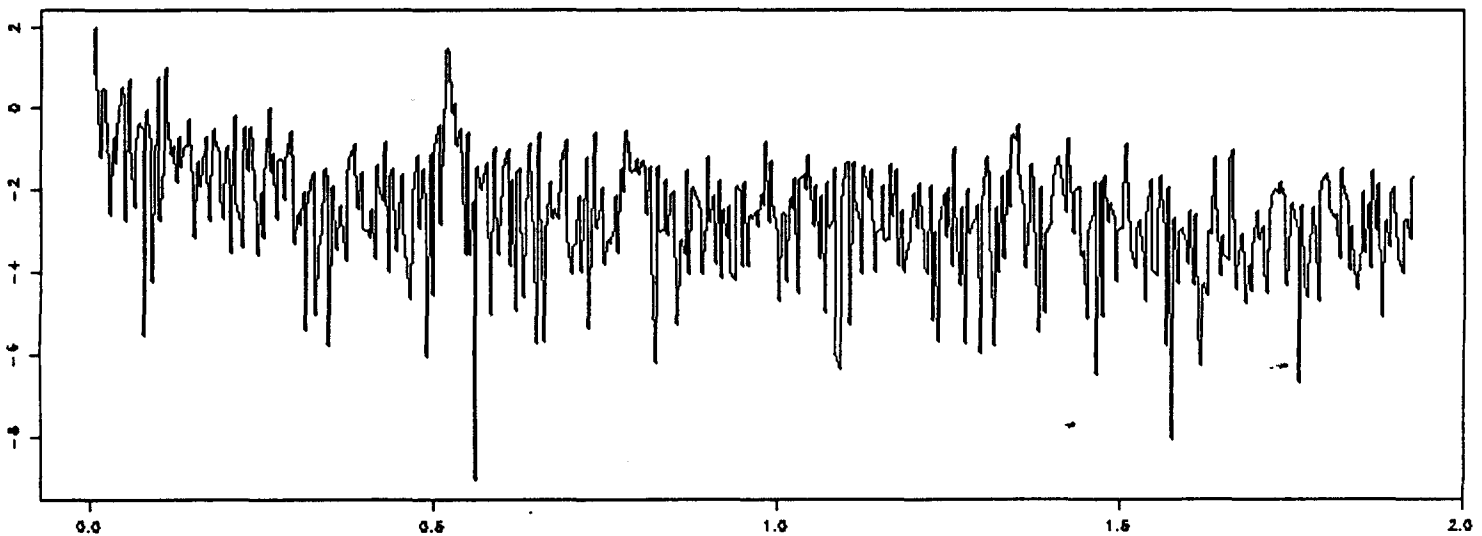
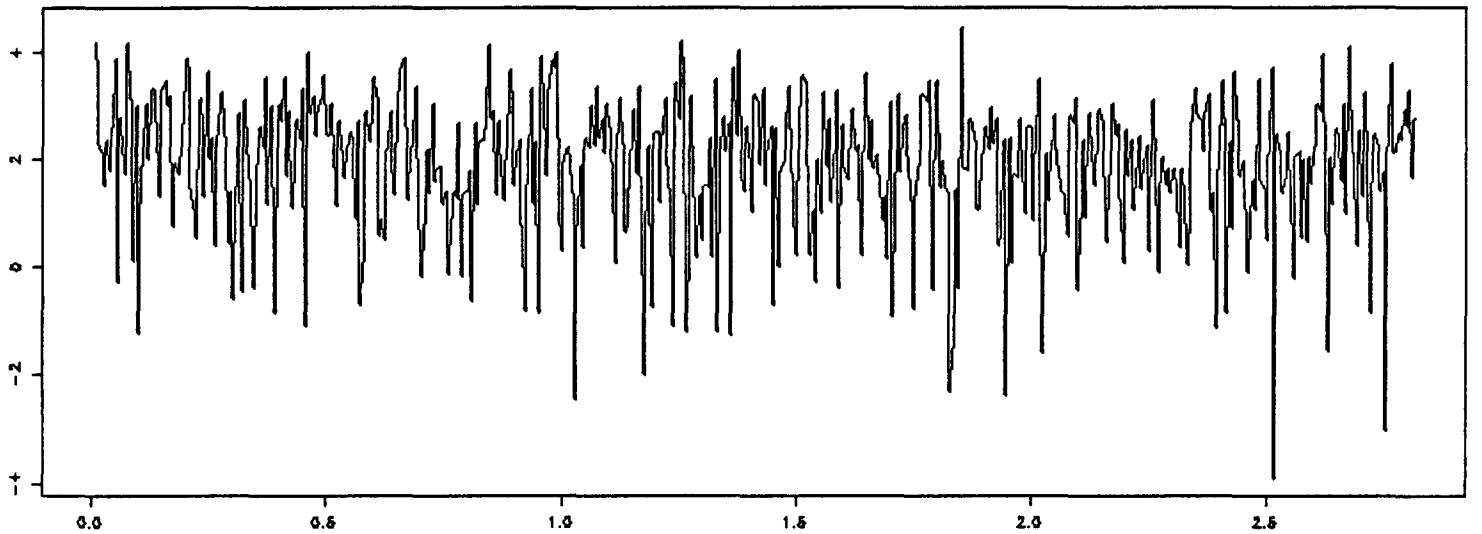
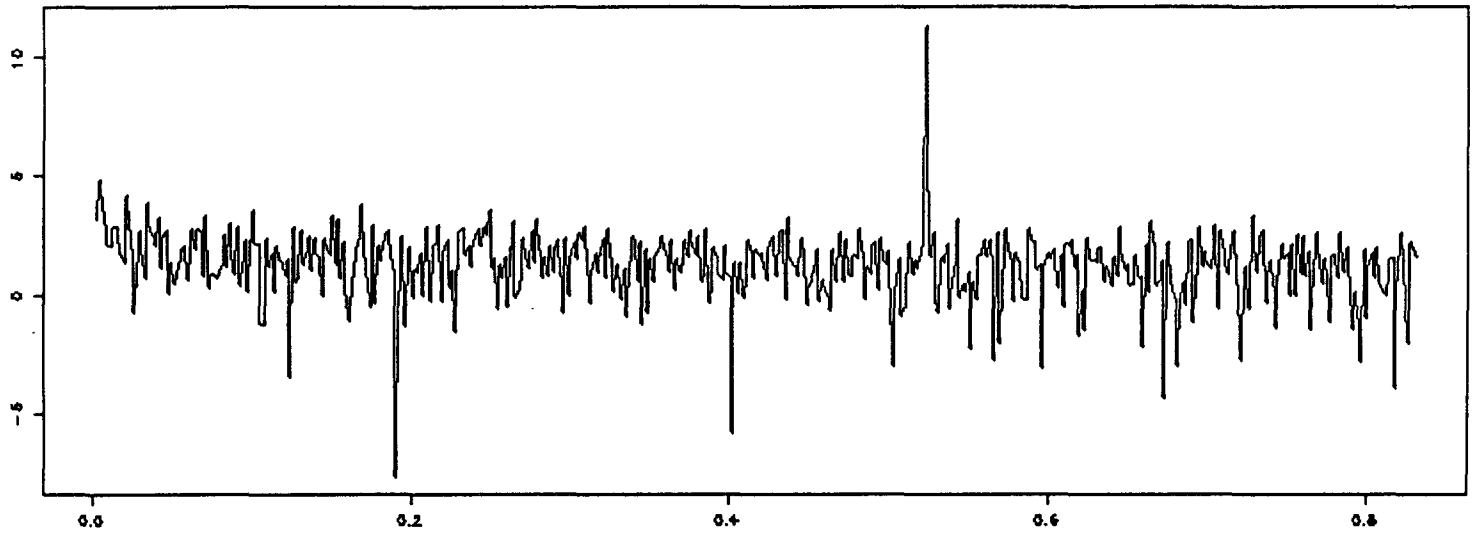


Figure 4: Periodograms on Log-Log Scales

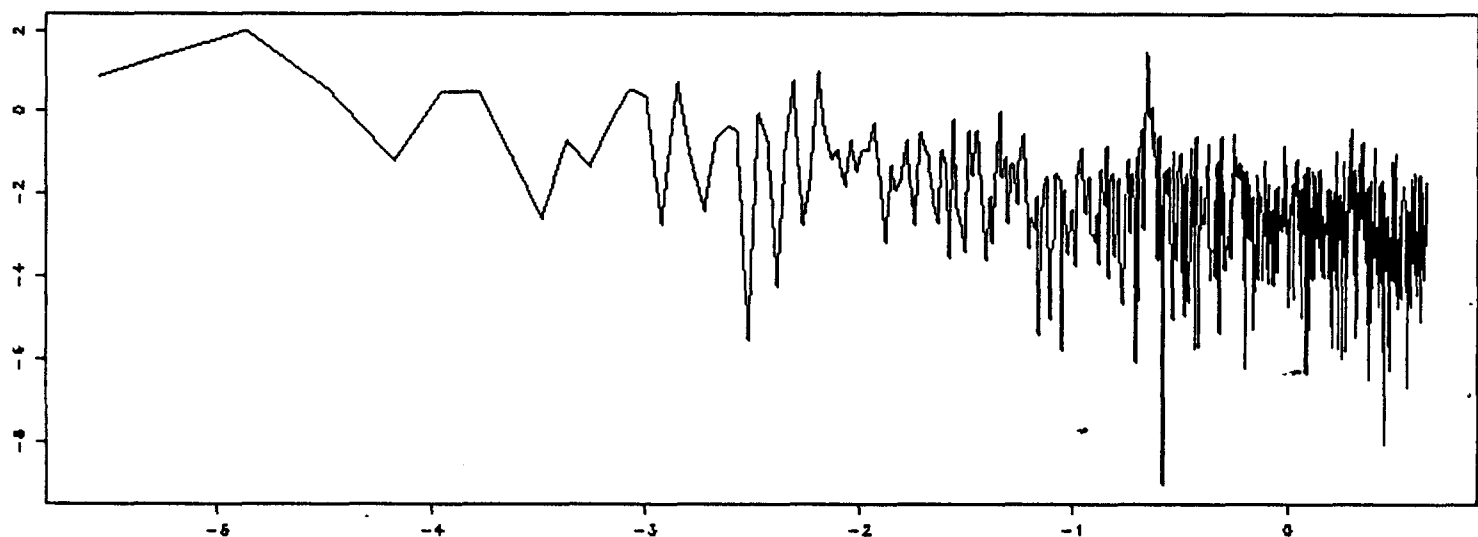
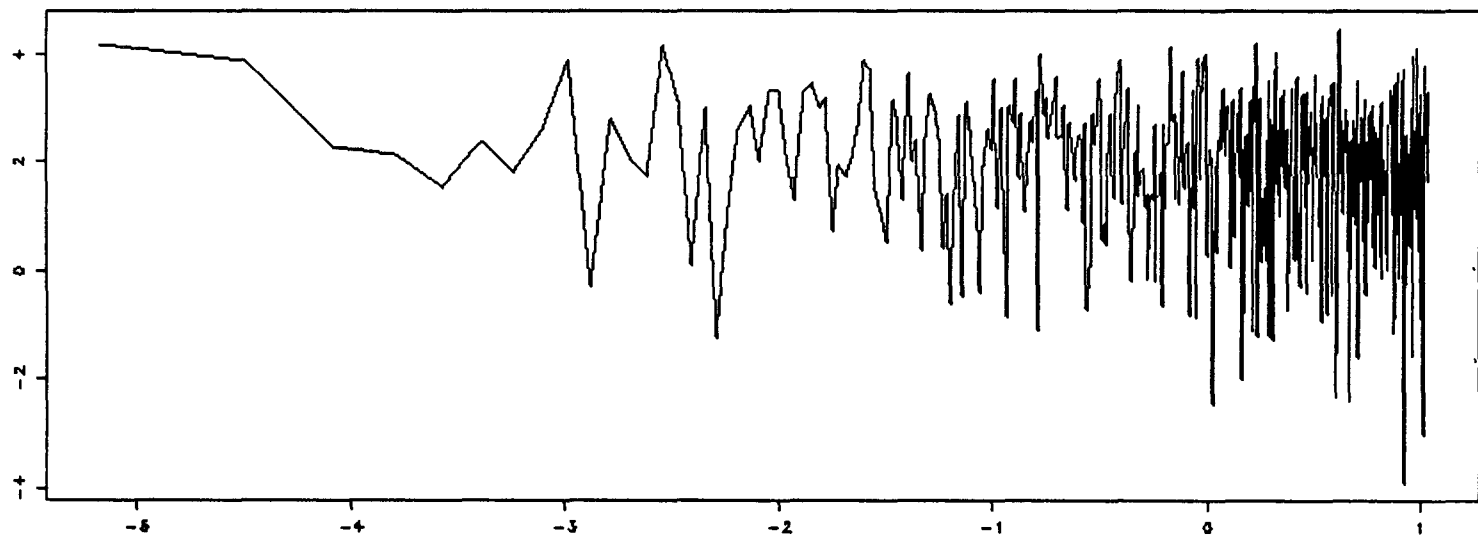
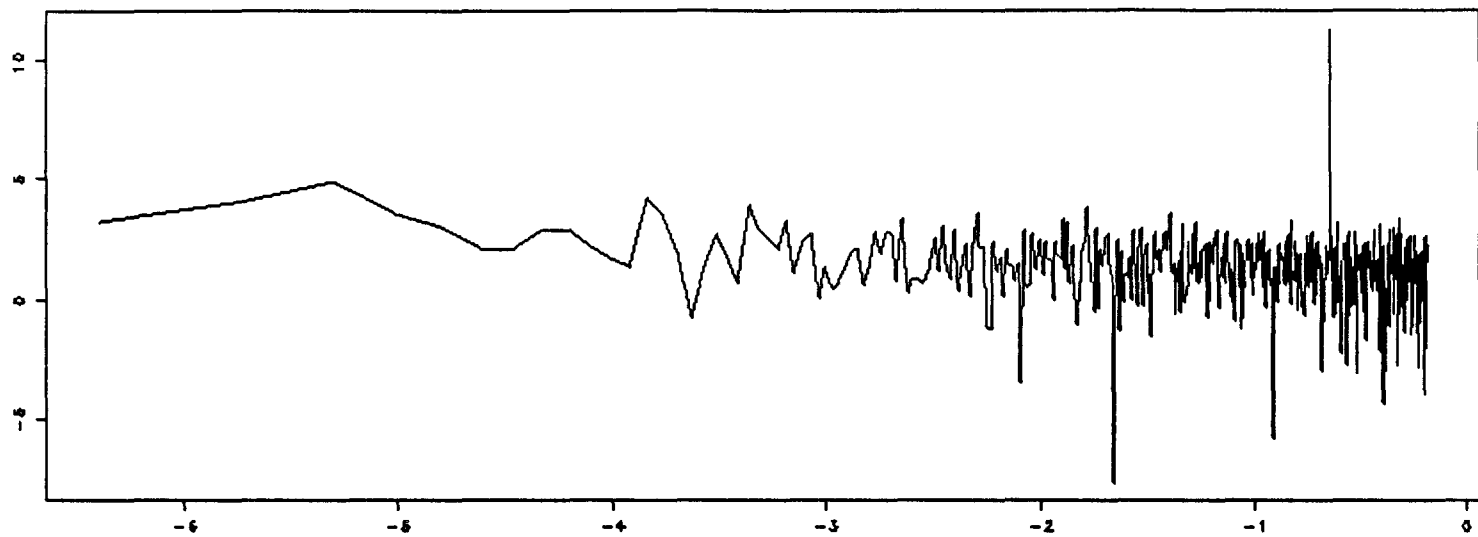


Figure 5: USA Data

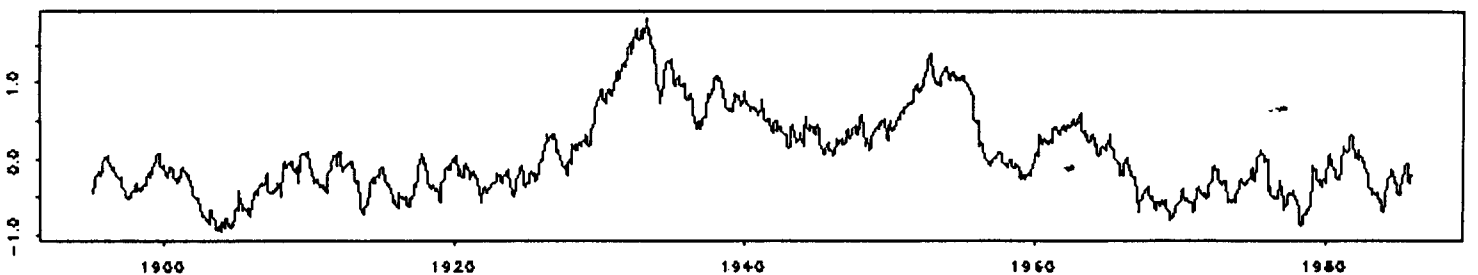
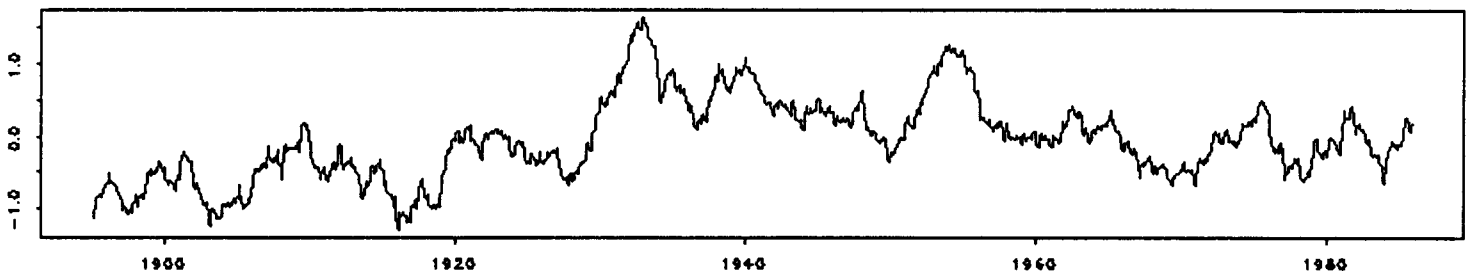
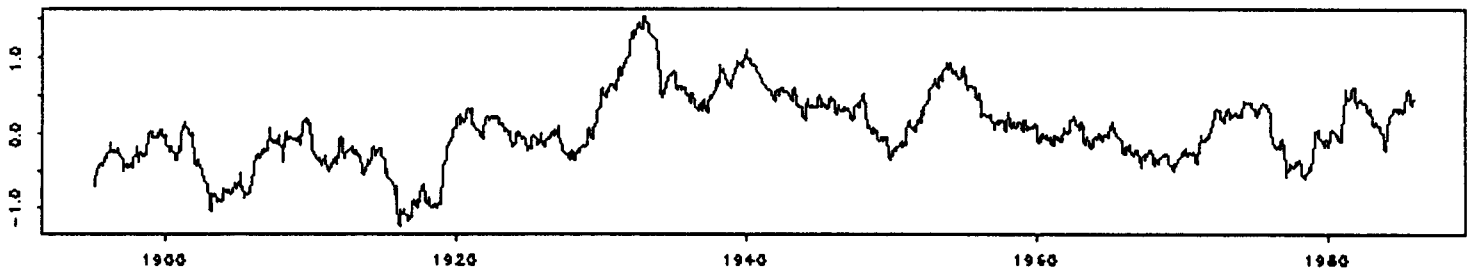
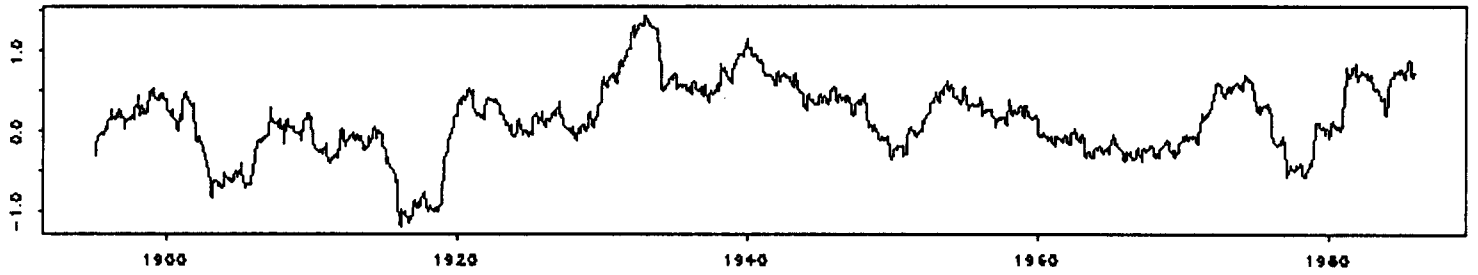
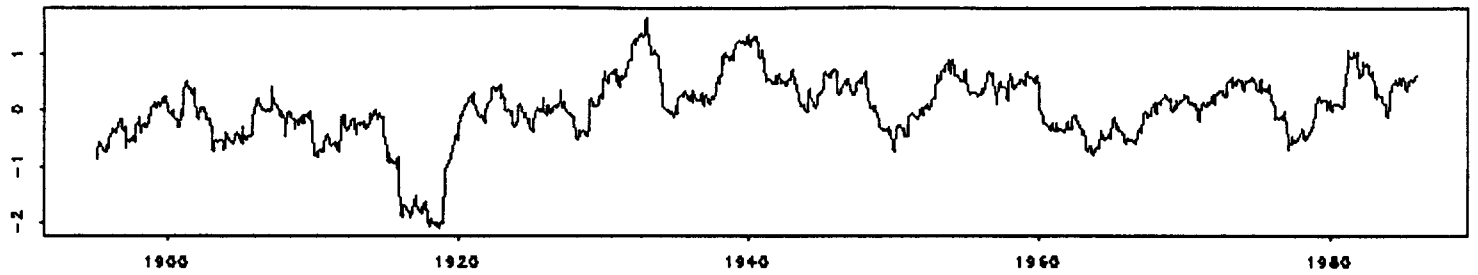


Figure 6: Global Data

