**ACKNOWLEDGMENTS**

This research was supported by ONR Contract N00014-91-J-1074. I am grateful to Chris Fraley for writing the Fracdiff software and for helpful comments, and to my long-time collaborator John Haslett for many useful discussions.

# Comment

**Richard L. Smith**

Jan Beran has written an excellent and timely review of a topic that is gaining increasing attention in a whole variety of fields. As his review makes clear, the origins of the subject go back a long way and were rooted in practical problems in several fields. However, it is only in recent years, stimulated by the development of the fractional ARIMA model, that the subject has started to receive widespread attention among statisticians. Beran does a superb job of bringing together the extensive results that now exist on the effects of long-range dependence on a whole range of statistical inferences. Nevertheless, I suspect it is in the identification and estimation of long-range models themselves that readers will take the greatest interest, and it is here that I concentrate my comments.

A common feature of long-range models is that the spectral density $f(x)$ satisfies the relation

$$(1) \qquad f(\omega) \sim b\omega^{1-2H}, \quad \omega \to 0.$$

Beran's equation (6) is a slight generalization of this, replacing the constant $b$ by a slowly varying function, but for most purposes (1) suffices. One feature of many of the results about the effect of long-range dependence, such as Beran's equation (8), is that they depend on the spectral density only through the constants $b$ and $H$. In fact, (8) itself depends only on $H$, but many related results depend also on the scaling constant $b$. For this reason, it is of interest to look for direct estimators of $b$ and $H$, rather than assume some parametric model such as fractional ARIMA. I have been particularly interested in estimators based on the periodogram, which are among those reviewed in Section 2.4. If $I_n(\omega)$ denotes the periodogram at frequency $\omega$ based on $n$ observations, then it is "well-known" that the sampling distribution of $I_n(\omega_j)$ at the Fourier frequencies $\omega_j = 2\pi j/n$ for $0 \le j < n/2$ is approximately that of independent exponential random variables with means $f(\omega_j)$. If we assume $f(\omega) = b\omega^{1-2H}$ then this suggests that $1 - 2H$ could be estimated as the slope of a linear regression of log $I_n(\omega_j)$ on log $\omega_j$. This idea has been suggested by a number of authors, in particular Geweke and Porter-Hudak (1983). Two refinements of Geweke and Porter-Hudak seem desirable:

a. Geweke and Porter-Hudak used least squares regression of log periodogram ordinate on log frequency. In contrast, since the asymptotic distribution of $I_n(\omega_j)$ is exponential, a regression of $\log I_n(\omega_j)$ based on errors from the Gumbel distribution function $1 - \exp(-e^x)$ would seem preferable. I call this the maximum likelihood (ML) approach, in contrast to Geweke and Porter-Hudak's least squares (LS) approach.
b. In addition, it is becoming increasingly clear that it is necessary to restrict the range of frequencies used in the regression, say to $n_0 \le j \le n_1$ where $1 < n_0 < n_1 << n/2$. At the lower end, the difficulty is that the above-mentioned "well-known" properties of the periodogram apparently break down for very low frequencies in the case of a long-range model (see, e.g., Künsch, 1987; Haslett and Raftery, 1989). At the upper end, the problem arises from the fact that (1) is only an asymptotic relation, not an identity, so attention must be restricted to small $\omega$. A more formal argument along these lines was presented in my discussion of Haslett and Raftery (1989).

It seems to me that Graf's HUB00 and HUBINC estimators deal with problem (a), albeit in a quite different way from the ML approach being suggested here, but do not contain anything that corresponds directly to the selection of $n_0$ and $n_1$. In view of this, I am somewhat doubtful about the theoretical justification of these estimators.

The rest of this discussion concerns three examples, two of them taken from Beran's paper, which illustrate the importance of appropriate selection of $n_0$ and $n_1$ in this approach.

The first of these is the Nile data. Beran's Figure 3 plots the periodogram in log-log coordinates. It can be seen that the plot is decreasing at an approximately

*Richard L. Smith is Professor, Department of Statistics, University of North Carolina, Chapel Hill, North Carolina 27599-3260.*

linear rate across its entire range, but that the first and third points on the plot are also the highest ones, and since these are also high-leverage points, they might cause trouble if we do not make $n_0$ at least 4. In fact, with $n_1$ selected somewhat arbitrarily to be 100, we find estimates of $H$ from a LS regression to be 0.87 (standard error 0.07) with $n_0 = 1$, 0.84 (0.08) with $n_0 = 2$, and so on as 0.85 (0.08), 0.80 (0.08), 0.82 (0.09) with $n_0 = 3, 4, 5$. Using ML regression in place of LS regression, we have estimates 0.88 (standard error 0.05), 0.86 (0.05), 0.87 (0.06), 0.78 (0.07) and 0.79 (0.07) again under $n_0 = 1, \dots, 5$. There does indeed seem to be a slight jump in the estimates after $n_0 = 3$. Fixing $n_0 = 5$ and varying $n_1$, we have estimates 0.97 (0.16), 0.86 (0.08), 0.86 (0.06), 0.87 (0.05) for $n_1 = 50, 150, 200, 250$ under LS regression, and 0.86 (0.13), 0.85 (0.05), 0.82 (0.04), 0.84 (0.04) under ML regression. Apart from the case $n_1 = 50$, which has high standard error, these results seem to be quite stable with respect to $n_1$, with the ML estimates of $H$ slightly lower than the LS estimates.

As a further illustration, Figure 1 shows standardized residuals from the ML regression of log periodogram on $\log \omega_j$, with $n_0 = 5$ and $n_1 = 150$. Part (a) shows a plot against $j$, and part (b) is a probability plot. The main feature of these plots is a group of 6 or 7 outliers corresponding to very low values of the periodogram. If LS regression is used and a normal probability plot constructed, then the effect of these outliers is even more pronounced. However, there is no other obvious departure from the model, and given that the outliers seem evenly distributed amongst the frequencies (Figure 1a), it seems unlikely that they are excessively influential. It may well be that Graf's HUBINC estimator is dealing effectively with this feature and I would be interested in Beran's comment on this point.

The overall conclusion seems to point to an estimator of $H$ in the range 0.8–0.85 with a standard error between 0.05 and 0.08. In contrast, the confidence intervals quoted by Beran for the HUB00 and HUBINC estimators correspond to standard errors around 0.02–0.03. Considering the variation of the regression-based estimators with $n_0$ and $n_1$, and the contrast between the LS and ML estimators, I think Graf's procedures are underestimating the standard errors, but the two sets of results are still broadly consistent, and in particular provide clear evidence of long-range dependence.

The situation with the NBS data seems to me much less clear cut. For one thing, when I computed the periodogram I found an extreme low outlier at $j = 43$; this does not appear on Beran's Figure 4 and distorts the estimates whether the ML or LS method is used. There are, of course, other ways of dealing with extremely low values of the periodogram. For example, this one is much less noticable if the data are tapered before calculating the periodogram. Therefore, I do not believe this feature is too important in itself, but it does have an effect on regression estimates and this needs to be taken into account. What is important is that in this case the visual evidence, if the plot is inspected carefully enough, suggests to me that the linear decrease does not persist over the whole plot, but only over about the first 40 or so periodogram coordinates. For this reason, I have taken $n_1 = 40$ and have obtained LS estimates for $H$ of 0.69 (standard error 0.13), 0.56 (0.17), 0.52 (0.21) under $n_0 = 1, 3, 5$, ML estimates of 0.68 (0.09), 0.65 (0.13), 0.64 (0.17). In this case the ML estimates are more stable than the LS estimates, but all the estimates have rather large standard errors. If we remove the outlier at $j = 43$ and extend the range of $n_1$, keeping $n_0 = 3$, we obtain LS estimates 0.59 (0.09), 0.49 (0.07) and ML estimates 0.71 (0.07), 0.60 (0.05) at $n_1 = 80, 120$. Gumbel plots

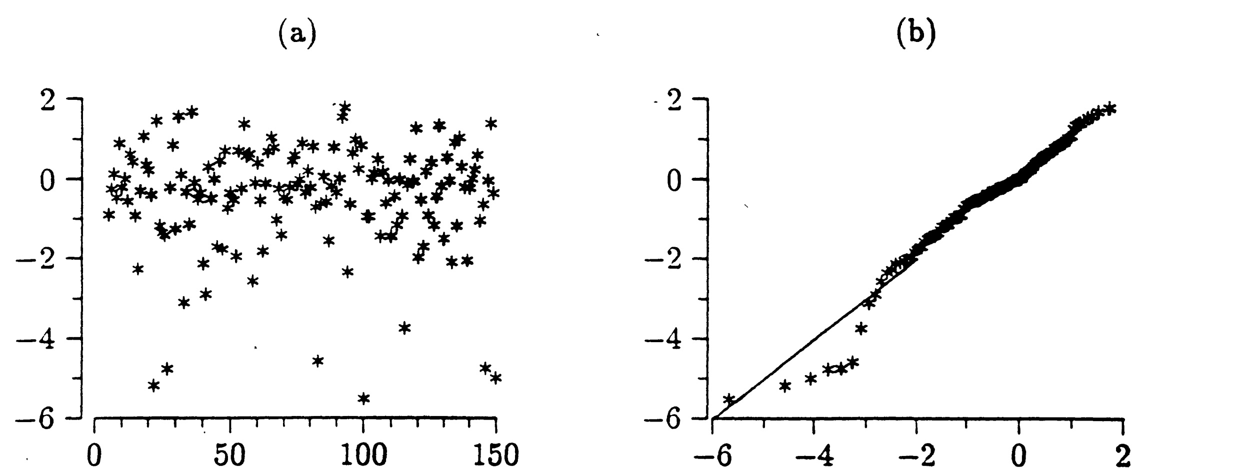


FIG. 1. *Plot of Gumbel residuals from model fitted to Nile data with $n_0 = 3$, $n_1 = 150$. (a) Plot of jth residual against j. (b) Probability plot of ordered residuals.*

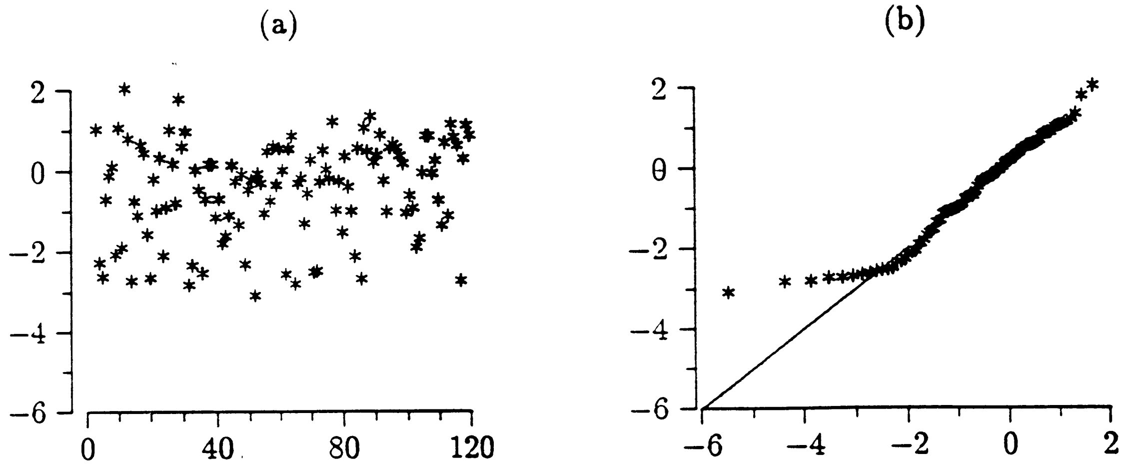


FIG. 2. *Plot of Gumbel residuals from model fitted to NBS data with* $n_0 = 3$, $n_1 = 120$, *the* 43*rd ordinate having been removed. (a) Plot of jth residual against j. (b) Probability plot of ordered residuals.*

for the case $n_0 = 3$ and $n_1 = 120$ are shown in Figure 2; my feeling about Figure 2a is that it seems to show a decrease over the first third of the plot followed by an increase, which is consistent with the fact that both the LS and ML estimates of $H$ take their lowest value when $n_1 = 120$. Figure 2b shows a clear discrepancy from the Gumbel distribution in the lower tail. In fact, in this case the normal distribution of log periodogram ordinates shows a much better fit than the Gumbel!

I can see no clear-cut interpretation of these results. There seems to be some evidence for long-range dependence, but it is not decisive, and I would not feel comfortable myself with Beran's claim that it is significant at 1%.

The third example is based on Smith (1993). Recently, there has been some interest in the use of long-range dependence models for the analysis of global warming phenomena. In particular, Bloomfield (1992) considered both conventional and fractional ARIMA models for global temperature series, concluding that there was significant evidence for long-range dependence in the context of fractional ARIMA models. Bloomfield and Nychka (1992) extended this analysis by comparing a number of approaches for deriving standard errors of estimated trend parameters based on the spectrum of the time series. In Smith (1993), I considered specifically the case of regression on a linear trend when the spectral density of the errors satisfies (1). If we regress observations $y_n$ on regressor $x_n = n$, $1 \le n \le N$, and if $\hat{\beta}$ is the least squares estimator of the slope, then under (1) we have

$$\text{(2)} \qquad \operatorname{Var} \hat{\beta} \sim \frac{36 b \pi (1-H)}{H(1+H)\Gamma(2H)\sin(\pi - \pi H)} N^{2H-4},$$

a special case of Yajima's (1988) results mentioned in Section 2.3. If we can estimate $b$ and $H$, then equation

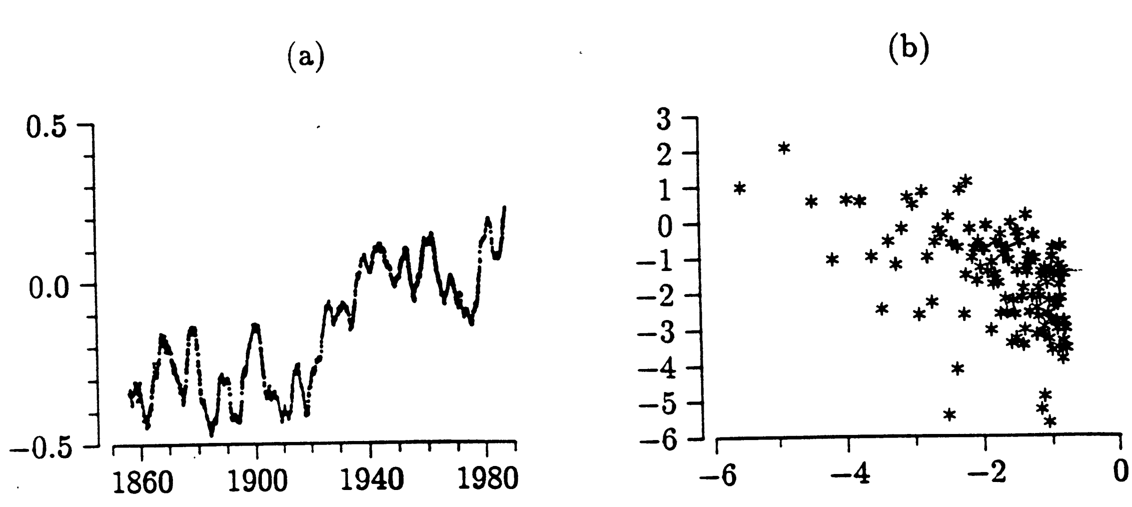


FIG. 3. *IPCC data. (a) Smoothed plot of raw data, monthly averages for* 1854–1989 *(b) The first* 120 *coordinates of the periodogram on log-log scales, calculated from the residuals of a linear regression fitted to the unsmoothed raw data.*

(2) provides an approximate formula that can be used to estimate the standard error of $\hat{\beta}$. As an example, we consider a series of 1632 average monthly temperatures over the Northern Hemisphere (land and sea) used by the IPCC (Intergovernmental Panel on Climate Change) for its global warming analyses. Figure 3a presents the data, smoothed by applying a 49-month moving average and centered around the 1950–1979 mean value for each month. It shows a pattern of steady behavior until about 1920, followed by a sharp rise between 1920 and 1940, then a gradual decrease until about 1975, followed by the sharp rise that has triggered the present alarm about global warming. Over the whole series, there is a clear rise in temperature, but whether it is due to the greenhouse effect is a matter of intense debate among climate scientists.

A linear trend was fitted to this data series (unsmoothed) and resulted in a estimated trend of 0.40°C per century, a figure consistent with several other estimates of global warming over the last century and a half. The first 120 periodogram ordinates of the residuals are plotted on log-log scales in Figure 3b. The pattern is quite similar to the two series quoted by Beran, and again seems to show evidence of long-range dependence. This is confirmed by the estimates $H = 0.90$ with standard error 0.05, based on $n_0 = 1$, $n_1 = 120$; also $\hat{b} = 0.0033$. When these figures are inserted into (2) (adjusted for the unit of trend) the standard error of the estimated trend is around 0.1, which is again consistent with earlier estimates of standard error including those quoted by Bloomfield (1992). My main doubt about this conclusion is whether the series can really be assumed stationary, given the obvious inconsistencies in methods of measurement over the last century and a half, but this would take us into other aspects beyond the scope of the present discussion.

I believe the message of all three examples is that the concept of long-range dependence must be taken seriously. At the same time, exactly how these examples are to be interpreted could be a matter of considerable debate. Jan Beran is to be congratulated on his very clear and comprehensive review, and I hope it will act as a springboard for much further research in this area.

### ACKNOWLEDGMENTS

Example 3 is based on data provided by Dr. P. Jones of the Climate Research Unit, University of East Anglia. I thank Jan Beran for providing me with the Nile and NBS data, and Peter Bloomfield for copies of his references. This work was supported in part by NSF Contract DMS-91-15750.

# Rejoinder

**Jan Beran**

I would like to thank the discussants for their stimulating comments and valuable suggestions. Their comments emphasize once more that long memory is an important issue to anybody who uses statistical inference, since it occurs rather frequently in real data and strongly influences the validity (and power) of standard tests and confidence intervals. Particularly interesting are the data examples analyzed by Smith (global warming–climatological data), Haslett and Raftery (wind speed–meteorological data) and Dempster and Hwang (employment series–economic data), since these are examples that concern everyone (and not just a selected group of scientists). Parzen summarizes the main message of the paper very clearly by saying that in data analysis, we always have to decide whether the data (either the original measurements or residuals, e.g., after subtracting a regression function) are white noise, a short-memory process or a long-memory process. The same view is expressed in a more general context by Mosteller and Tukey (1977, p. 119 ff): "even in dealing with so simple a statistic as the arithmetic mean, it is often vital to use as direct an assessment of its internal uncertainty as possible. Obtaining a valid measure of uncertainty is not just a matter of looking up a formula." In other words, no formula should be applied without checking its approximate validity. Naturally, this does not only refer to "classical" formulas, such as $\mathrm{var}(\overline{X}) = \sigma^2 n^{-1}$, but also to the "new" formulas, such as $\mathrm{var}(\overline{X}) = L(n)n^{2H-2}$ $(0 < H < 1)$, given in the present review paper.

One major reason why the question of long memory is usually not dealt with in daily statistical practice is the lack of statistical software packages. Haslett and Raftery's program (and its implementation in the next release of SPLUS) is therefore a welcome contribution. As already mentioned briefly after formula (12) and