Statistical Modeling of Storm Counts*

ANDREW R. SOLOW

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts

(Manuscript received 21 April 1988, in final form 13 September 1988)

ABSTRACT

A statistical model is presented of a recently compiled record of monthly extratropical storm counts for the mid-Atlantic coast of the United States for the period 1942–83. The counts are modeled as a Poisson process with nonstationary mean function. The mean function is decomposed into a secular component and a seasonal cycle. Because the form of the secular component is unknown, a nonparametric regression approach suitable for Poisson data is used to estimate it. The estimated secular component is generally constant through the 1950s, then declines through the 1970s. The estimate is found to be statistically significant. A Fourier series involving two harmonics is fit to the seasonal cycle. A preliminary check indicates that the seasonal cycle remains stable through time. Some diagnostics based on suitably defined residuals are presented that generally confirm the goodness-of-fit and distributional assumptions underlying the model.

1. Introduction

This paper presents a statistical model of a record of the number of extratropical storms that impinged upon part of the mid-Atlantic coast of the United States during the period 1942–83. These storm counts were compiled on a monthly basis by Dolan et al. (1987). The corresponding annual counts are plotted in Fig. 1a. The monthly counts exhibit a strong seasonal cycle. This is illustrated in Fig. 1b, where the monthly counts—with the estimated secular component removed—are plotted by month. Because the dataset begins in July 1942 and ends in June 1984, the convention is adopted that the storm year runs from July to June.

There is considerable practical and scientific interest in coastal storm climate. This interest stems from the enormous impact that coastal storms can have on shoreline processes (e.g., extreme sea levels) and natural and man-made structures, and from the realization that changes in the level and geographic pattern of atmospheric temperatures may lead to changes in the frequency, intensity, and trajectory of coastal storms (e.g., Wendland 1977). An important initial step towards understanding coastal storm climate (and other processes related to coastal storm climate) is an understanding of its historic behavior. This paper aims at describing the historic behavior of storm frequency as represented in the data of Dolan et al. (1987). Specifically, models of the secular and seasonal components of these data are developed. These models differ from standard regression-type models in two respects. First, the data are modeled as arising from a nonstationary Poisson process. The Poisson model is the natural statistical model for data in the form of counts, particularly when counts are small (as they are in the case of monthly data). Second, because the form of any secular trend in the data is unknown, a nonparametric approach is taken. Under this approach, the form of the secular trend is determined by the data themselves, and not imposed a priori.

Other studies of storm frequency include Mather et al. (1964), Reitan (1974, 1979), and Hayden (1981). These studies consider different datasets, and do not attempt any statistical analysis. Mooley (1981) and Thompson and Guttorp (1986) consider Poisson and related models for a record of severe cyclonic storms in the Bay of Bengal, although the nature of their data and their methods of analysis differ from those presented here.

The remainder of this paper is organized as follows. In section 2, the basic model is presented and an algorithm is given for fitting the model. In section 3, the secular component is estimated using local likelihood estimation (Hastie and Tibshirani 1987). Local likelihood estimation is a nonparametric regression method suitable for Poisson data. The seasonal component is estimated in section 4. Some diagnostics based on residuals are presented in section 5. Section 6 contains some brief concluding remarks.

* Woods Hole Oceanographic Institution Contribution No. 6869.

Corresponding author address: Dr. Andrew R. Solow, Woods Hole Oceanographic Institution, Woods Hole, MA 02543.

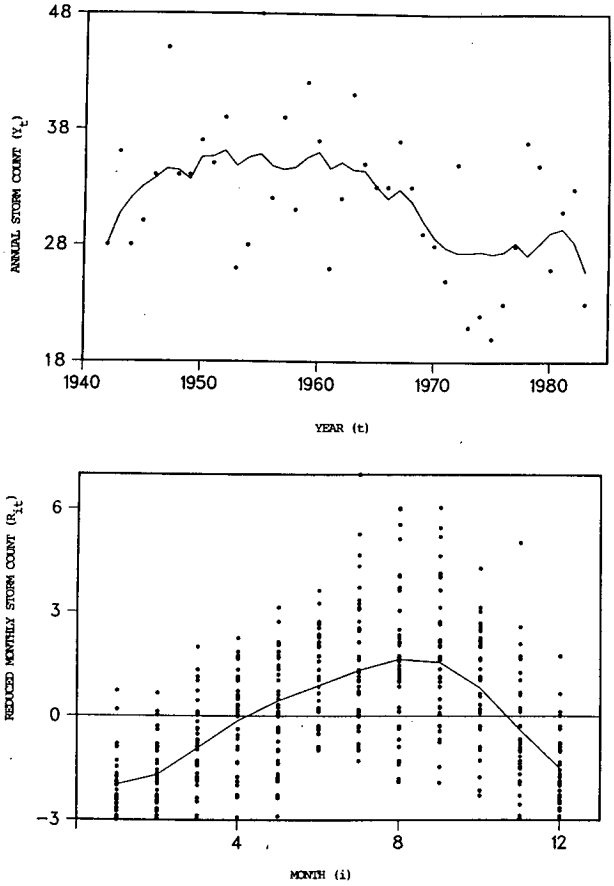

FIG. 1. (a) Annual counts of extratropical storms for the United States mid-Atlantic coast (Y_t), 1942–83 (Dolan et al. 1987). Also shown is the local likelihood estimate of the secular component ($\hat{G}_t$). (b) Reduced monthly storm counts for the United States mid-Atlantic coast (R_{it}) grouped by month, running from July ($i = 1$) through June ($i = 12$). Also shown is the estimate of the seasonal component ($\hat{f}_i$).

2. Approach

Let X_{it} be the storm count for month i in year t. The basic model is that X_{it} ($i = 1, \cdots, 12$, $t = 1, \cdots, n$) is a sequence of independent Poisson random variables with

$$E(X_{it}) = \text{var}(X_{it}) = m_{it}$$

where E denotes expectation and var denotes variance. The goal of the analysis is to fit a model of the form

$$m_{it} = f_i + g_t \qquad (1)$$

where f_i ($i = 1, \cdots, 12$) is a seasonal cycle that does not depend on t, and g_t ($t = 1, \cdots, n$) is a smooth, secular component that does not depend on i. To ensure estimability, the seasonal cycle is required to satisfy

$$\Sigma_i f_i = 0.$$

The assumption that the seasonal component is stable over t will be checked later. The assumption that the secular component is constant within a year is an approximation. An alternative would be to assume that the secular component is linear during the course of a year. Because g_t changes slowly, the results of the analysis are insensitive to this choice, and so the simplest alternative is adopted.

Let Y_t ($t = 1, \cdots, n$) be the storm count for year t. Because

$$Y_t = \Sigma_i X_{it},$$

the annual counts are also independent Poisson random variables with

$$\begin{aligned} E(Y_t) = \text{var}(Y_t) &= \Sigma_i m_{it} \\ &= 12 g_t \\ &= G_t. \end{aligned}$$

The following algorithm will be used to fit (1):

(i) estimate G_t ($t = 1, \cdots, n$) from the annual counts

(ii) form the reduced monthly counts:

$$R_{it} = X_{it} - \hat{g}_t \quad (i = 1, \cdots, 12, t = 1, \cdots, n)$$

where $\hat{g}_t = \hat{G}_t/12$ and $\hat{G}_t$ is the estimate of G_t found in (i)

(iii) estimate f_i ($i = 1, \cdots, 12$) from the reduced monthly counts

The reason that f_i and g_t are not estimated simultaneously from the monthly counts is that, because the form of g_t is unknown, the two components are confounded. This would not be the case if, for example, g_t is assumed to be linear.

3. Estimating the secular component

In this section, G_t ($t = 1, \cdots, n$) is estimated from the annual counts using local likelihood estimation (Hastie and Tibshirani 1987). Local likelihood estimation is a nonparametric regression method that is suitable for Poisson (and other nonnormal) data. The method assumes that the local behavior of the secular component can be approximated by a polynomial in t of specified, low order. In this paper, the secular component is assumed to be locally linear. The parameters of the local model are estimated by the method of maximum likelihood. If the data are independent and normally distributed with constant variance, the maximum likelihood estimates are given by ordinary least squares regression. For a wide class of models called generalized linear models (McCullagh and Nelder 1983), the maximum likelihood estimates can be found by iterative reweighted least squares regression.

For fixed t, the locally linear model assumes that

$$G_{t'} \approx b_{0t} + b_{1t} t', \quad t' \in N(t)$$

where $N(t)$ is a neighborhood of t. That is, the secular component is assumed to be approximately linear within $N(t)$, with local regression parameters b_{0t} and b_{1t} that depend on t (but not on t'). For fixed t, the estimates of b_{0t} and b_{1t} are found by iterative reweighted least squares regression applied to the data in $N(t)$. This procedure starts with initial estimates of b_{0t} and b_{1t} (e.g., the ordinary least squares estimates) and uses these estimates to find estimates of $\mathrm{Var}(Y_{t'})$, $t' \in N(t)$. These estimated variances are used in weighted least squares regression to find new estimates of b_{0t} and b_{1t}, and the procedure iterates to convergence. Once convergence is reached, the fitted value at t is taken as the estimate of Y_t. The procedure is repeated for each t in the record.

In applying local likelihood estimation, it is necessary to choose the neighborhood in which the local fitting is performed. Only symmetric nearest-neighbor neighborhoods will be considered here. In this case, the choice of neighborhood is reduced to the choice of span. Define the half-span, h, as the number of nearest-neighbor observations on each side of t used in the local fitting at t. The total number of observations used in local fitting at t (including the observation at t itself) is $2h + 1$. Note that the neighborhood is truncated near $t = 1$ and $t = n$. The choice of h entails a trade-off between variability and bias. If h is too small, the estimate of the secular component will be too variable (i.e., the fit will be dominated by randomness in the observations). If h is too large, the estimate will be biased (i.e., the fit will fail to capture local features).

There are a number of automatic methods for choosing h, most involving the optimization of some cross-validated measure of goodness-of-fit. Minimization of the cross-validated deviance will be used in this paper to choose h. The deviance is a measure of the discrepancy between observations and fitted values. For a single observation, the deviance is defined as

$$\mathrm{dev}(y, \hat{y}) = 2[l(y) - l(\hat{y})]$$

where y and $\hat{y}$ are the observed and fitted values of Y, respectively, and l is the log-likelihood (McCullagh and Nelder 1983). For a normally distributed observation, the deviance is the squared error. For an observation with a Poisson distribution the deviance is given by

$$\mathrm{dev}(y, \hat{y}) = 2[y \log(y/\hat{y}) - (y - \hat{y})]. \qquad (2)$$

The deviance is additive for independent observations. The total deviance serves as a generalization of the usual residual sum of squares for nonnormal data. The cross-validated deviance is the deviance when y_t is not used in forming $\hat{y}_t$. Cross-validation is used in the choice of h to avoid overfitting (Allen 1974).

Local likelihood estimation was applied to the observations shown in Fig. 1a. The half-span chosen by minimizing the cross-validated deviance was 5 years. The estimate of G_t ($t = 1, \cdots, n$) is also shown in Fig. 1a. This estimate indicates the presence of structure that is not entirely apparent from the observations alone. Qualitatively, the overall fit shown in Fig. 1a appears to capture the main features of the data. Note that the close fit to the first observation is not due to any constraint in the fitting. The decline in storm frequency for the period 1973–76 is not reproduced altogether adequately. This period of decline is short relative to h and is bracketed by high counts in 1972 and 1978. For this period, the assumption of local linearity is not very good. An alternative would be to assume locally quadratic behavior. Of course, it is possible to attribute this period of decline to random fluctuation, in which case any effort to fit it better would constitute overfitting. Note that the final observation appears to have a strong influence on the fit in the last few years. While the same might appear to be true for the first observation, this is not the case.

The estimate of G_t ($t = 1, \cdots, n$) shown in Fig. 1a ranges in value from 36.1 storms per year in 1952 to 25.8 storms per year in 1983. This represents a decline of around 30%. A test of the statistical significance of the fit shown in Fig. 1a can be based on the total deviance. For this fit, the deviance was 34.0. The deviance for fitting the null model:

$$G_t = G_0 \quad t = 1, \cdots, n$$

(i.e., constant mean annual storm frequency) was 52.2. The estimate of G_0 under this model is 32.1, the average annual storm count. The reduction in deviance achieved by using the local likelihood fit instead of the null fit is 18.2. To assess the statistical significance of this reduction, it is necessary to know its distribution under the null hypothesis. This distribution is approximately chi-squared, with degrees of freedom given by the number of parameters used in the local likelihood fit minus 1 (the number of parameters used in the null fit).

Hastie and Tibshirani (1987) give an approximation to the number of parameters used in local likelihood estimation, based on the discussion of Cleveland (1979). The approximation is given by the trace of the hat matrix (i.e., the n-by-n matrix that maps the n-vector of observations, y, into the n-vector of estimates, $\hat{y}$). The derivation of this approximation is rather technical, and the interested reader is referred to the original references. In the case of the storm count data, the value of this approximation is 5.35. Comparing the reduction in deviance to the chi-squared distribution with 4.35 degrees of freedom, the local likelihood fit is found to be significantly better than the null fit ($p = 0.002$). The local likelihood fit can also be compared to the global linear model:

$$G_t = b_0 + b_1 t \quad t = 1, \cdots, n$$

in the same way. The deviance for the fitted global linear model is 44.1. The estimates of b_0 and b_1 under this model are 36.5 and -0.202, respectively. The re-

duction in deviance achieved by using the local likelihood fit instead of the global linear fit is 10.1. Comparing this reduction to the chi-squared distribution with 3.35 (i.e., 5.35 − 2) degrees of freedom, the local likelihood fit is found to be significantly better than the global linear fit ($p = 0.024$).

It is worth pointing out that if the global linear model is compared to the global constant model, the reduction in deviance—8.1 for one degree of freedom—is highly significant ($p = 0.004$). Thus, even a simple parametric analysis would indicate the presence of trend in these data, although the estimated trend would be significantly worse than that estimated nonparametrically.

4. Estimating the seasonal cycle

Once G_t ($t = 1, \cdots, n$) was estimated, the reduced monthly counts:

$$R_{it} = X_{it} - \hat{g}_t, \quad i = 1, \cdots, 12, \quad t = 1, \cdots, n$$

were formed, where $\hat{g}_t = \hat{G}_t/12$. According to the basic model

$$E(R_{it}) \approx f_i$$

$$\text{var}(R_{it}) \approx f_i + g_t$$

A Fourier series model,

$$E(R_{it}) = f_i = \sum_{j=1}^{k} a_j \sin ij\pi/6 + b_j \cos ij\pi/6, \quad (3)$$

was fit by iterative reweighted least squares, in which the current estimate of f_i was added to $\hat{g}_t$ to form an estimate of $\text{var}(R_{it})$. The number of harmonics included in this model (i.e., the value of k) was chosen in a forward stepwise fashion. That is, an additional harmonic was included if the reduction in deviance achieved by its inclusion was statistically significant. Under the null hypothesis that the added harmonic is not needed (i.e., that its coefficients are both zero), this reduction has an approximate chi-squared distribution with two degrees of freedom (since the additional harmonic uses two coefficients).

The total deviance for the model including a single harmonic was 495.3. The inclusion of a second harmonic reduces the total deviance to 479.8. This reduction is highly significant ($p = 0.001$). The inclusion of a third harmonic reduces the total deviance to 478.7. This reduction is not statistically significant ($p = 0.58$), so the final model includes two harmonics. The estimated coefficients in this model are

$$\hat{a}_1 = -1.23 \quad \hat{b}_1 = -1.19$$

$$\hat{a}_2 = -0.21 \quad \hat{b}_2 = -0.31$$

and the estimate of f_i ($i = 1, \cdots, 12$) is shown in Fig. 1b. The estimate reaches its maximum in February and its minimum in July, and has a range of 3.61 storms. Incidentally, the total deviance for a model including two harmonics but no secular component is 496.1. The reduction in deviance achieved by including the secular component—16.3 for 5.35 degrees of freedom—is highly significant ($p = 0.008$).

The basic model (1) assumes that the seasonal cycle is stable through time. An alternative to this model is one in which the counts for different months have different trends through time. For example, the secular behavior shown in Fig. 1a could be due to changes in winter storm frequency alone, with storm frequency in other seasons remaining constant through time. Behavior of this type would appear as an unstable seasonal cycle in the reduced monthly counts. As a rough check of the stability of the seasonal cycle, the following exercise was performed. Let

$$I_{it} = \begin{cases} 1, & \text{if} \quad it \leqslant 252 \\ 0, & \text{otherwise} \end{cases}$$

That is, I_{it} is an indicator that takes the value 1 if X_{it} is in the first half of the record and 0 otherwise. A model of the form

$$E(R_{it}) = I_{it}(\Sigma_j a_j \sin ij\pi/6 + b_j \cos ij\pi/6) + (1 - I_{it})(\Sigma_j c_j \sin ij\pi/6 + d_j \cos ij\pi/6)$$

was fit to the reduced monthly counts. This model allows the coefficients of the seasonal cycle to be different in the two halves of the record. The model (with two harmonics) was fitted by iterative reweighted least squares. The estimated coefficients are

$$\hat{a}_1 = -1.37, \quad \hat{b}_1 = -1.11$$

$$\hat{a}_2 = -0.20, \quad \hat{b}_2 = -0.52$$

$$\hat{a}_3 = -1.09, \quad \hat{b}_3 = -1.28$$

$$\hat{a}_4 = -0.22, \quad \hat{b}_4 = -0.12$$

The total deviance for this model is 474.3. The reduction in deviance achieved by splitting the record in this way—5.5 for four degrees of freedom—is not statistically significant ($p = 0.24$). This result supports the assumption of a stable seasonal cycle. However, the roughness of this analysis should be stressed, and a more systematic investigation of the stability of the seasonal cycle is currently underway.

5. Model validation

In this section some diagnostics based on residuals from the fitted model are presented. The use of residuals in generalized linear models is discussed by Pregibon (1979) and McCullagh and Nelder (1983). The aim of these diagnostics is to validate the fitted model. This model consists of a systematic component and a random component. The systematic component refers to the estimates of f_i ($i = 1, \cdots, 12$) and g_t ($t = 1, \cdots, n$). The random component refers to the as-

sumption that the monthly counts are independent Poisson variates.

A natural choice of residual in a generalized linear or locally linear model is the deviance residual. For Poisson data, the deviance residual is

$$d_t = \text{sgn}(y_t - \hat{y}_t)\{2[y_t \log(y_t/\hat{y}_t) - (y_t - \hat{y}_t)]\}^{1/2}$$

Note that the sum of squared deviance residuals is equal to the total deviance, and in this sense the deviance residual is the natural analogue to the ordinary residual for normal data.

In Fig. 2, a plot of the annual deviance residuals, d_t, against the fitted values, $\hat{y}_t$, is shown. This plot is useful for evaluating the overall model fit to the annual counts and for checking the assumed relationship between the mean and variance of Y_t. Departures from the fitted model appear as trend or curvature in this plot, while departures from the assumed mean–variance relationship appear as trend in the spread of points. It is important to bear in mind that even if the assumed mean–variance relationship is correct, intervals on the $\hat{y}_t$ scale with a high density of points will produce a wider range of residuals than intervals with a low density of points.

Turning to Fig. 2, note that there is no evidence of trend or curvature, so the overall fit to the annual counts is judged to be satisfactory. The spread of the residuals appears to be large for both low and high values of $\hat{y}_t$ and small for moderate values of y_t. Note, however, that there are 15 points with $\hat{y}_t$ between 27.1 and 30.0, 6 points with $\hat{y}_t$ between 30.1 and 33.0, and 18 points with $\hat{y}_t$ between 33.1 and 36.0. This suggests that the behavior of the spread is due to differences in the density of points and not to any misspecification of the mean-variance relationship.

To evaluate the overall fit to the monthly counts, in Fig. 3 the sample autocorrelation function of the monthly deviance residuals is shown. This plot is useful for detecting residual trend or periodicity that is not captured in the fit. Figure 3 indicates no gross structure

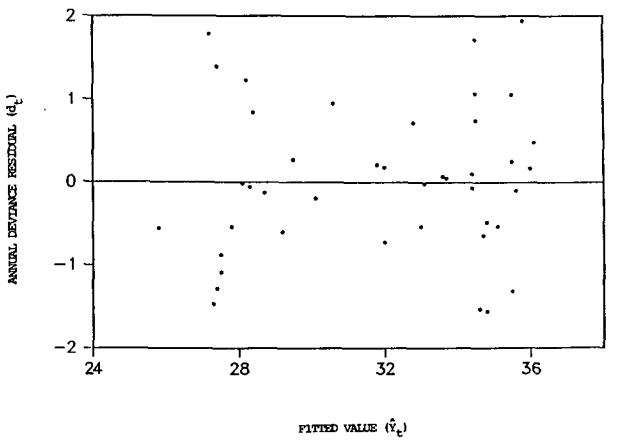

FIG. 2. Annual deviance residuals (d_t) vs annual fitted values ($\hat{Y}_t$).

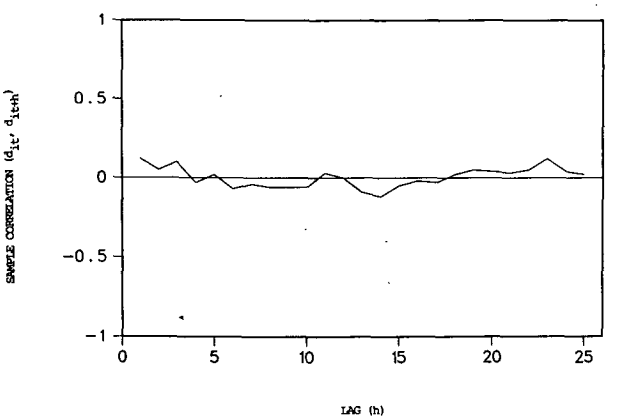

FIG. 3. Sample autocorrelation function of the monthly deviance residuals. Lag (h) is measured in months.

in the monthly residuals—periodic or otherwise—and generally confirms the goodness of fit to the monthly counts.

The monthly deviance residuals are also useful for detecting departures from the assumptions about the random component of the data. These assumptions are that the monthly counts are independent Poisson variates. Provided these assumptions are correct (and provided the systematic component is adequately captured), the monthly deviance residuals are approximately independent unit normal (Pregibon 1979). Figure 3 indicates the possible presence of weak lag-1 serial correlation in the monthly deviance residuals. For the purposes of model-fitting, the magnitude of the estimated correlation—0.12—is small enough to be negligible. The existence of serial correlation in monthly storm counts is an important and interesting matter, and deserves further study. To check unit normality, the normal probability plot for the monthly deviance residuals is shown in Fig. 4. This is a plot of the ordered deviance residuals against the corresponding expected order statistics for the unit normal distribution (see, for example, David 1981). This plot nearly has unit slope and shows no important departures from linearity. On the basis of these checks, the assumption that the monthly counts are independent Poisson variates is generally confirmed.

6. Discussion

The purpose of this paper has been to present a statistical model of the storm count data summarized in Fig. 1. The modeling approach represents a compromise between a fully parametric analysis, which allows formal inference within a restricted class of models, and a fully nonparametric analysis, which considers a richer class of models at the expense of formal inference. Although the estimation of the secular component is nonparametric (in the sense that no parametric form is specified in advance), sufficient structure is

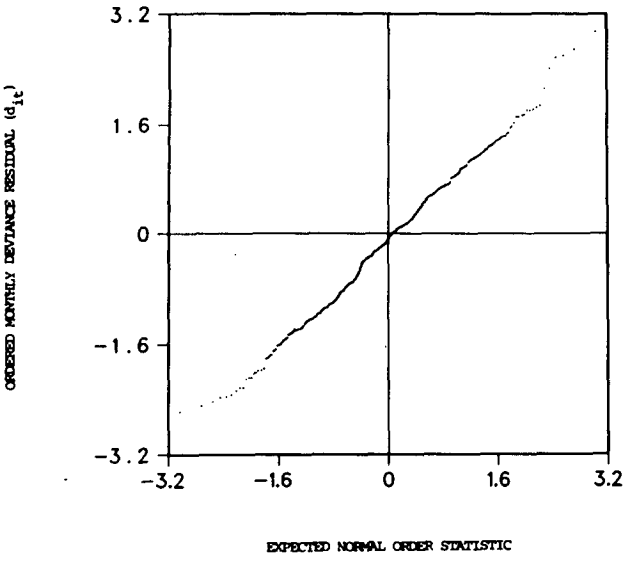

FIG. 4. Normal probability plot for the monthly deviance residuals.

imposed upon the data by the model to allow formal inference. In addition, the diagnostic procedures applied to the final model fit allow the goodness of fit and the validity of the model assumptions to be checked.

Turning to the results of the analysis, the local likelihood fit shown in Fig. 1a indicates that storm frequency broadly decreased from the 1950s through the 1970s. There is some evidence of an increase at both ends of the record, but it is important not to overinterpret the behavior at the ends where the observations have high leverage. The overall fit is significantly better than both the global constant fit and the global linear fit. From a practical point of view, the 30% decline in storm frequency from the early 1950s to the end of the record is also significant. The residual analysis indicates that the full model captures the behavior of the monthly counts fairly well. In particular, Fig. 3 shows no important residual structure, although there is some suggestion of weak lag-1 serial correlation. Finally, preliminary results do not show any evidence of instability in the seasonal cycle through time.

Because the model of storm counts uses only time as a regressor variable, it is purely descriptive. A descriptive model can be extremely useful. For example, the results presented in this paper indicate that a successful physical model of coastal storm climate should be able to explain the broad decline in storm frequency over the period of record. In addition, these results may prove useful in explaining the historic behavior of other coastal processes such as erosion rates or extreme sea levels. Nevertheless, an important extension of this analysis is the construction of explanatory statistical models relating storm frequency to other climatological factors (e.g., sea surface temperature). Because the relationships between storm frequency and other factors are unknown (and likely to be complex), the multivariate version of local likelihood estimation would be useful for exploring these relationships.

Acknowledgments. This work was supported by the J. N. Pew Charitable Trusts and by the Marine Policy Center of the Woods Hole Oceanographic Institution. The author wishes to thank D. G. Aubrey, two anonymous reviewers, and especially R. Rosen for their useful comments.

REFERENCES

Allen, D. M., 1974: The relationship between variable selection and data augmentation and a method for prediction. *Technomet.*, **16,** 125–127.

Cleveland, W. S., 1979: Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74,** 829–835.

David, H. A., 1981: *Order Statistics.* Wiley, 360 pp.

Dolan, R., B. Hayden, K. Bosserman and L. Lisle, 1987: Frequency and magnitude data on coastal storms. *J. Coastal Res.*, **3,** 245–247.

Hastie, T., and R. Tibshirani, 1987: Local likelihood estimation. *J. Am. Stat. Assoc.*, **82,** 559–567.

Hayden, B., 1981: Secular variation in Atlantic coast extratropical cyclones. *Mon. Wea. Rev.*, **109,** 159–167.

Mather, J., H. Adams and G. Yoshioka, 1964: Coastal Storms of the eastern United States. *J. Appl. Meteor.*, **3,** 693–706.

McCullagh, P., and J. A. Nelder, 1983: *Generalized Linear Models.* Chapman and Hall, 280 pp.

Mooley, D. A., 1981: Applicability of the Poisson probability model to the severe cyclonic storms striking the coast around the Bay of Bengal. *Sankhya,* **43B,** 187–197.

Pregibon, D., 1979: Data analytic methods for generalized linear models. Ph.D. thesis, University of Toronto, 164 pp.

Reitan, C. H., 1974: Frequencies of cyclones and cyclogenesis for North America: 1951–1970. *Mon. Wea. Rev.*, **102,** 861–868.

——, 1979: Trends in the frequencies of cyclone activity over North America. *Mon. Wea. Rev.*, **107,** 1684–1688.

Thompson, M. L., and P. Guttorp, 1986: A probability model for severe cyclonic storms striking the coast around the Bay of Bengal. *Mon. Wea. Rev.*, **114,** 2267–2271.

Wendland, W. M., 1977: Tropical storm frequencies related to sea surface temperatures. *J. Appl. Meteor.*, **16,** 477–481.