Statistical Science
2000, Vol. 15, No. 2, 153–167

Nonparametric Analysis of Temporal Trend When Fitting Parametric Models to Extreme-Value Data

Peter Hall and Nader Tajvidi

Abstract. A topic of major current interest in extreme-value analysis is the investigation of temporal trends. For example, the potential influence of "greenhouse" effects may result in severe storms becoming gradually more frequent, or in maximum temperatures gradually increasing, with time. One approach to evaluating these possibilities is to fit, to data, a parametric model for temporal parameter variation, as well as a model describing the marginal distribution of data at any given point in time. However, structural trend models can be difficult to formulate in many circumstances, owing to the complex way in which different factors combine to influence data in the form of extremes. Moreover, it is not advisable to fit trend models without empirical evidence of their suitability. In this paper, motivated by datasets on windstorm severity and maximum temperature, we suggest a nonparametric approach to estimating temporal trends when fitting parametric models to extreme values from a weakly dependent time series. We illustrate the method through applications to time series where the marginal distributions are approximately Pareto, generalized-Pareto, extreme-value or Gaussian. We introduce time-varying probability plots to assess goodness of fit, we discuss local-likelihood approaches to fitting the marginal model within a window and we propose temporal cross-validation for selecting window width. In cases where both location and scale are estimated together, the Gaussian distribution is shown to have special features that permit it to play a universal role as a "nominal" model for the marginal distribution.

Key words and phrases: Bandwidth, cross-validation, extreme-value distribution, kernel, location estimate, nonparametric regression, Pareto distribution, probability plot, scale estimate.

1. INTRODUCTION

In applications of extreme-value methods to meteorological or environmental data, a major topic of current interest is assessment of temporal trends in measurements such as temperatures or storm intensities. A trend might be present in terms of scale, for example, when fitting a generalized Pareto distribution or it might be in location, in particular when fitting an extreme-value distribution. There can also be trends in "shape," described, for example through the exponent of a fitted Pareto distribution. In that case, evidence that the exponent was decreasing over time would imply that the distribution was becoming more heavy tailed, corresponding to generally more severe events and greater variability in severity. When an extreme-value distribution is used to model data, temporal trends in location, scale and shape are all potentially of interest.

One approach to assessing trend is to test for a linear or log-linear change, as for example in the work of Smith (1989) on ground-level ozone concentrations and of Rootzén and Tajvidi (1997) on damage by windstorms. However, regardless of the conclusions of such analyses, they almost invariably raise further questions about the nature of the

Peter Hall is Professor, Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia. Nader Tajvidi is Associate Professor, Department of Mathematics, Linköping University, 581 83 Linköping, Sweden

trend. If a test fails to reject the null hypothesis of no trend, against the alternative of linearly increasing trend, is it because there is an increasing trend but it is not statistically significant or is the trend quite different from increasing? If the null hypothesis is rejected, is the trend really linear or is it more complex? In some instances it is of interest to take an alternative route, exploring the nature of any trend prior to constructing a formal test for it.

In the present paper we suggest an adaptive, nonparametric approach to solving these problems. We fit models that are structural at any given point in time, but vary with time in a nonstructural way. Two particular datasets motivate our methodology: the windstorm data of Rootzén and Tajvidi (1997) and data on temporal change of annual maximum temperatures in Australia. In each case, fitting a linear trend to parameters of Pareto-type or extreme-value distributions suggests a tendency for the response variables to slowly increase with time. In the case of the Australian temperature data, this would accord with concerns about global warming.

We argue, however, that in both cases the trend is actually more complex than such a simple parametric analysis would allow. We show that if nonparametric methods are used to fit time-varying parameters then, for both datasets, it appears that “average” values of the response variable at first decrease and then increase with time.

Although the temperature data studied here are confined to the southeast Australian state of Victoria, the same broadly “convex” trend for variation of maximum temperature with time is apparent throughout much of the eastern half of Australia. Moreover, it is reflected not just in maximum temperatures but also in rainfall; the eastern half of the Australian continent appears to have been both colder and wetter in the middle of the twentieth century than it was in the 50 preceding or 50 succeeding years.

Once such trends have been revealed they can be incorporated into a new, nonlinear parametric model that may be fitted to data by relatively conventional means. Thus, our adaptive, nonparametric techniques might be viewed as exploratory tools, rather than as methods for final analysis. Either way, they offer an adaptive approach that might have been helpful in earlier analyses. In one such case, involving the study of ground-level ozone concentration, Smith (1989) fitted a linear trend to the mean of an extreme-value distribution. In discussion of Smith’s paper, Raftery (1989) argued that a linear trend might not adequately represent the manner in which ozone concentrations changed with time and suggested that a change-point model might be more appropriate. The methods proposed for addressing that suggestion were parametric, however; an adaptive, nonparametric approach would have been beneficial as an exploratory tool. The ozone data are unfortunately no longer available for analysis, but if they were, analyses of this type would likely help resolve questions raised in the discussion of Smith (1989).

We focus particularly on fitting Pareto, generalized-Pareto, extreme-value and Gaussian distributions to data. It is shown that the Gaussian model has special properties which make it attractive as a universal approach to simultaneous local estimation of location and scale, valid even when the model is incorrect. That method avoids the need to first compute an undersmoothed estimator of location, and then calculate residuals, in order to compute a function-valued estimator of scale. We develop local methods, based on a continuum of probability plots, for assessing goodness of fit when function-valued parameters are involved. Our procedure for actually fitting a model is based on a kernel approach to “local likelihood,” and is directly applicable to general weakly dependent time-series data, not just those connected with extreme events.

In the context of more conventional sampling problems, rather than just temporal sequences of events, our local-likelihood approach is closely related to “local estimating equations” techniques suggested by Carroll, Ruppert and Welsh (1998), differing for example in the method used to select bandwidth and in our notion of a nominal Gaussian model for location and scale. These authors note particular applications to nutritional epidemiology, to analysis of lung cancer mortality rates and to modelling overdispersion in count and assay data.

Additionally, a great many local methods for curve fitting can be interpreted as local-likelihood based. They include, for example, local polynomial fitting (e.g., Fan and Gijbels, 1996) and locally weighted regression more generally (e.g., Cleveland and Devlin, 1988), if we fit a local model in which the errors in response variables are Normally distributed. Of these, the LOESS method in SPlus (Chambers and Hastie, 1992) is arguably the best known. At a higher level of methodological sophistication, local-likelihood techniques include methods where the mean and variance of the response are related through a link function; see, for example, Weisberg and Welsh (1994), who illustrated applications using an example from quality management in an industrial setting.

In closely related work, Fan, Heckman and Wand (1995) provided a treatment of kernel-weighted general linear models and quasi-likelihood.

Mixed parametric and nonparametric approaches to inference have also been discussed by Staniswallis (1989), who noted particular applications to the Cox proportional hazards model to survival analysis; Gu, Bates, Chen and Wahba (1989) and Gu (1990), who used spline rather than kernel smoothing to weight likelihoods and addressed applications to survival analysis and Severini and Staniswallis (1994), who considered quasi-likelihood applications to evaporation rates in engineering contexts. Spatial applications are illustrated in examples treated by Gu (1990) and Carroll, Ruppert and Welsh (1998). Independently of the work in the present paper, Davison and Ramesh (2000) have developed local likelihood-based methods for smoothing sample extremes.

Extreme-value theory for dependent sequences has been studied in depth by Leadbetter, Lindgren and Rootzén (1983, Part II) and reviewed by Leadbetter and Rootzén (1988).

Our motivating datasets are introduced and discussed in Section 2. Local-constant and local-linear versions of our methods are introduced in Section 3, and applied to the data in Section 4. Section 5 outlines the estimators' theoretical properties. Higher-order local polynomial techniques may be treated similarly. However, we found in our numerical work that even when estimating just location and scale, where local-linear methods require computation of four function-valued parameter estimators (rather than two), local-linear estimation could become heavily saturated, particularly when design points were sparse.

2. EXAMPLES AND PARAMETRIC METHODS FOR ANALYSIS

Our first example is of the intensities of windstorms, measured in millions of Swedish kroners (MSEK) of damage, experienced by the Swedish insurance group Länsförsäkringar during the 12-year period 1982–1993. Sample size is $N = 45$. The data are depicted in Figure 1(a), and are discussed in more detail by Rootzén and Tajvidi (1997). A storm was defined to occur if damage exceeded 0.9 MSEK and if certain meteorological conditions were met. We standardized the time interval $[1982, 1993]$ by transforming it linearly to $\mathscr{I} = [0, 1]$.

Of particular interest, especially to the company that collected the data, are potential trends in storm strength, for example, any tendency of strength to increase with time. Indeed, fitting a generalized Pareto distribution (GPD) with log-linearly varying scale shows a gradual tendency for storm strength to increase.

A likelihood ratio test for a log-linear increase may be conducted by fitting a GPD and assuming that storm strengths, conditional on the times of storms, are independent random variables. The test results in statistical significance only at the 0.32 level of probability, however. The data suggest visually that an increase might be present only in the last 80% of the period 1982–1993, although a test applied to that region results in significance only at the 0.09 level, and this level would only be increased if we took into account the way in which the choice of interval depended on the data.

The inconclusive nature of these results indicates the need for a more exploratory, less structured ap-

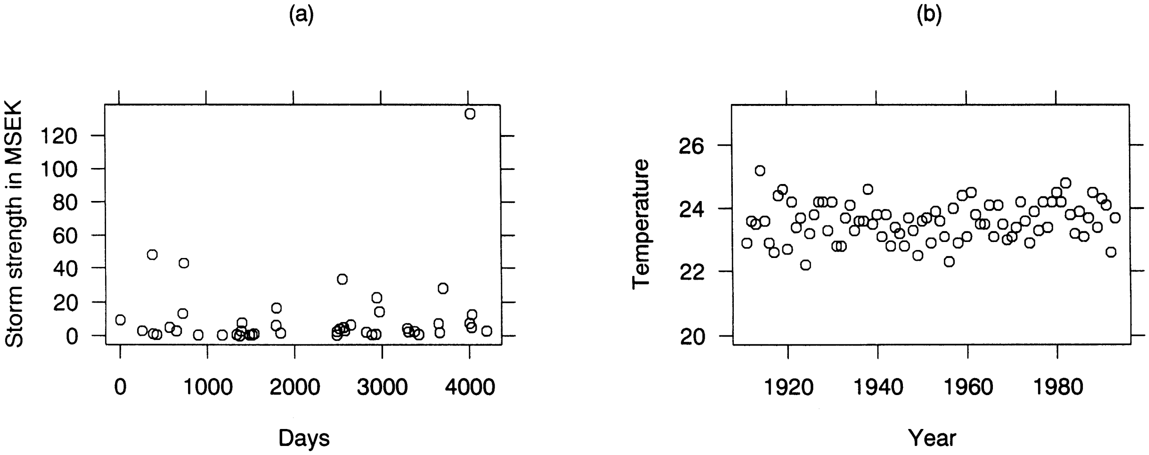

FIG. 1. *Data on windstorms and temperatures in Victoria, Australia. Panel* (a) *graphs storm strengths* (*i.e., insurance losses*) *exceeding* 0.9 MSEK, *against day of occurrence* (*measured from 1 January 1982*) *in the period 1982–1993. Panel* (b) *shows annual maximum temperatures, in degrees Celsius, taken over* 34 *weather stations in Victoria, Australia, from 1910 to 1993.*

proach to analysing trend — one that is more sensitive to fluctuations in the data and less constrained by assumptions made outside the dataset, although admittedly less explicit in terms of strict assessment of goodness of fit. The adaptive, nonparametric methods suggested in Section 3 offer such an alternative approach. We would stress that those techniques might not necessarily be viewed as an end in themselves; rather, their purpose could be to suggest appropriate parametric models that might be fitted.

Our second example is of Australian temperature extrema; see Jones (1994) and Torok and Nicholls (1996) for discussion of a very large dataset from which our data were excerpted. Figure1(b) depicts the maximum value, over all 34 weather stations that were operating in the state of Victoria from 1910 to 1993, of annual temperatures (in degrees Celsius) during this period. Fitting a generalized extreme-value distribution with linearly varying location shows a slight tendency for temperatures to increase over time, although not to an extent that would lend strong support to concerns about global warming, for example.

It is not clear from this analysis whether there really is a generally increasing trend or whether the time trend is more complex than linear. Using our more adaptive methods we shall argue in Section 4 that the trend is likely nonlinear in the case of the Victorian temperature data and that in fact there is evidence of a minimum value of maximum temperature occurring in about 1950 in the state of Victoria. Analysis of a larger data set shows that this "convex" trend was exhibited in eastern Australia more generally during the twentieth century. Rainfall data lend additional support to this claim.

3. NONPARAMETRIC ESTIMATION OF TREND

3.1 General Methodology

Suppose data are gathered in the form $\{(X_1, T_1), \dots, (X_N, T_N)\}$, where T_i denotes the time at which an event of strength X_i is observed. We might think of the (X_i, T_i) sequence as a marked point process, with X_i being the mark on the point represented by T_i. Usually N is random, representing the number of events observed in a given time interval $\mathscr{I}$. It is assumed that the distribution of X_i, given $T_i = t$, has density $f(\cdot|\theta)$, where $\theta = \theta(t)$ is a vector of dimension d (a d-vector) and is a smooth function of t. Likelihood-based methods will be formulated under the assumption that the pairs (X_i, T_i) are independent. However, depending on the nature of the dependence, first-order properties of bias and variance of estimators of θ can be unchanged if the X_i's form a weakly dependent time series, conditional on the T_i's. See Section 5 for discussion.

Assume that, conditional on T_i, X_i has density $f\{\cdot|\theta(T_i)\}$. Put $g(x|\theta) = \log f(x|\theta)$ and, given a bandwidth $h > 0$ and a kernel K, define

$$K_i(t) = K\Big(\frac{t - T_i}{h}\Big).$$

Let v_0, v_1 denote d-vectors, being candidates for $\theta(t)$ and $\dot{\theta}(t) = d\theta(t)/dt$, respectively; put $\omega_i = \omega_i(v_0, v_1) = v_0 + (T_i - t)\, v_1$ and define

$$(3.1) \qquad \ell(v_0, v_1|t) = -\sum_{i=1}^{N} g\{X_i|\omega_i(v_0, v_1)\}\, K_i(t).$$

Two estimators of $\theta(t)$ are, respectively, the "local-constant" estimator, $\hat{\theta}(t) = \hat{v}_0$, where $\hat{v}_0$ minimizes $\ell(v_0, 0|t)$ with respect to v_0, and the "local-linear" estimator, $\tilde{\theta}(t) = \tilde{v}_0$, where $(\tilde{v}_0, \tilde{v}_1)$ minimizes $\ell(v_0, v_1|t)$ with respect to (v_0, v_1). The case where the T_i's are nonrandom, for example, regularly spaced, may be treated identically.

A minor modification allows us to estimate some of the parameters globally and others locally. For simplicity of notation, let us write $f(x|\eta, \psi)$ instead of $f(x|\theta)$, where $\theta = (\eta, \psi)$; we wish to estimate η locally and ψ globally. Put $g(x|\eta, \psi) = \log f(x|\eta, \psi)$, define $\omega_i(v_0, v_1)$ as before and replace $g\{X_i|\omega_i(v_0, v_1)\}$ by $g\{X_i|\omega_i(v_0, v_1), \psi\}$ in (3.1). Holding ψ fixed, compute the estimator $\tilde{\eta}_\psi(t) = \tilde{v}_0$, in the local linear case by minimizing the right-hand side of the new version of (3.1) with respect to (v_0, v_1). Now select $\hat{\psi}$, a global estimator of ψ, by minimizing

$$(3.2) \qquad -\sum_{i=1}^{N} g\{X_i|\tilde{\eta}_\psi(T_i), \psi\}$$

with respect to ψ. The final estimator of η is $\tilde{\eta}_{\hat{\psi}}$. There is an obvious local-constant version of this procedure, but there we should confine summation at (3.2) to indices i such that T_i is not close to the edges of $\mathscr{I}$, so as to avoid boundary problems.

Often $\mathscr{I}$ is a compact interval, and although θ would typically be continuous on this interval, it would generally have jump discontinuities at its ends. In theory the local-constant estimator is more seriously affected by such edge effects than the local-linear estimator, although in practice we found that both often performed similarly for small-to-moderate sample sizes. We also noticed that local-constant estimators suffered less from design sparseness.

When constructing the local-linear estimator it is often necessary to reparametrize, so as to ensure that $v_0 + (T_i - t)\, v_1$ always lies in the parameter space. For example, if the jth component $\theta^{(j)}$ of θ

must be positive in order for $f(\cdot|\theta)$ to be well defined, then it may be appropriate to fit the linear model $v_0^{(j)} + (T_i - t)\, v_1^{(j)}$ to $\log \theta^{(j)}(T_i)$ rather than to $\theta^{(j)}(T_i)$.

3.2 Fitting a Nominal Gaussian Location-and-Scale Model

As with conventional likelihood-based methods, the functional estimators $\hat{\theta}$ and $\tilde{\theta}$ can be inconsistent for θ if the model $f(\cdot|\theta)$ is incorrect. In particular, this is typically true when we estimate location and scale by taking $\theta = (\mu, \sigma)$ and $f(x|\theta) = \sigma^{-1}\phi\{(x-\mu)/\sigma\}$, where the distribution corresponding to the probability density ϕ has zero mean and unit variance and μ, σ are nondegenerate functions (of t).

The case where ϕ is the standard Gaussian density is an exception, however. There, the local-constant estimator $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$, and local-linear estimator $\tilde{\theta} = (\tilde{\mu}, \tilde{\sigma})$, are consistent for the (location, scale) vector under appropriate regularity conditions, that do not include correctness of the Gaussian model. In the case where μ and σ are both nondegenerate functions, this result is apparently new; it is discussed in theoretical detail in Section 5.2. The resulting estimators of μ and σ are first-order equivalent to their local-constant and local-linear counterparts in the case of conventional least-squares regression, and so performance is not sacrificed by estimating them together. (The same is true if we parametrize scale in terms of σ^2 rather than σ.) Usually, however, functional estimation of σ would proceed by first calculating an undersmoothed estimator of μ, then computing residuals, then centering and finally, passing a local-linear smoother through squared and centered residuals. Therefore, our nominal Gaussian model approach saves computational effort. Theoretical details will be given in Section 5.

3.3 Assessing Goodness of Fit

Formal testing of goodness of fit of models for extreme-value data is not often a practical proposition, in particular because alternative hypotheses that might be employed in a likelihood-ratio test (for example) can be complex to conduct inference for. These problems are exacerbated when a bandwidth, as well as more conventional parameters, are estimated from data using kernel weights. An assessment in terms of probability plots is often more appropriate.

In keeping with our local approach to parameter fitting, we suggest computing probability plots locally. Thus, our plots are defined in the continuum, there being a different plot for each $t \in \mathscr{I}$. To construct a plot within a window of width $2h$ centered on t, let us write $(X_1', T_1'), \ldots, (X_M', T_M')$ for those values of (X_i, T_i) for which $T_i \in [t-h, t+h]$. Let $F(x|\theta)$ denote the distribution function corresponding to the density $f(x|\theta)$. If the model is correct then the values of $Z_i = F\{X_i'|\theta(T_i')\}$ are uniformly distributed on $[0, 1]$, and so a plot of expectations of ranked values of Z_i, against the respective ranks, would produce exactly a straight line. This suggests taking the probability plot to be a graph of values of $\widehat{Z}_i = F\{X_i'|\bar{\theta}(T_i')\}$ against their respective ranks, where $\bar{\theta}$ denotes either $\hat{\theta}$ or $\tilde{\theta}$.

An alternative is to plot ranked values of $\widehat{Z}_i' = F\{X_i'|\bar{\theta}(t)\}$ against their ranks. However, $\bar{\theta}(t) - \theta(T_i')$ is of order only h for $T_i' \in [t-h, t+h]$, whereas $\bar{\theta}(T_i') - \theta(T_i')$ is of order h^2, assuming that h is chosen to give an optimal rate of convergence. (These results follow from theoretical properties outlined in Section 5.) Consequently, $\widehat{Z}_i' - Z_i = O_p(h)$, whereas $\widehat{Z}_i - Z_i = O_p(h^2)$, and so the method based on $\widehat{Z}_i$, rather than $\widehat{Z}_i'$, is preferable. Even so, it can be advantageous to choose h a little smaller than is optimal for estimating θ, so as to reduce the effects of systematic error. This also reflects the fact that, in view of the shape of the kernel used to compute $\bar{\theta}(\cdot)$, data corresponding to T_i's that are further from t should receive less weight than those near to t.

3.4 Selecting Bandwidth

There are several potential approaches to bandwidth choice. They include plug-in methods, based on formula for asymptotic mean squared error that we shall give in Section 5; techniques based on approximate log-likelihood ratios; a cross-validation algorithm that we shall discuss below and a modified form of Ruppert's (1995) empirical bias technique, used by Carroll, Ruppert and Welsh (1998) in a setting related to ours. In a multivariate problem such as that of estimating a vector function θ, conventional plug-in methods are not so attractive because they involve selection of a range of pilot bandwidths.

Our approach is related to cross-validation for nonparametric density estimation, but differs in that it cross-validates over the times, T_i, as well as the strengths, X_i, and of course, it employs the parametric model for f. In asymptotic terms it chooses the bandwidth to minimize

$$\begin{aligned} D_1 &= D_1(h) \\ &\equiv \int_{\mathscr{I}} E\{(\bar{\theta}-\theta)^T \Sigma(t)\, (\bar{\theta}-\theta)\}\, \lambda(t)\, dt\,, \end{aligned} \tag{3.3}$$

where $\bar{\theta}$ denotes the estimator of θ (either the local-constant or the local-linear estimator), $\Sigma(t)$ is the $d \times d$ positive-definite matrix of which the (j_1, j_2)th element is

$$E\big[f_{j_1}\{X|\theta(t)\}\, f_{j_2}\{X|\theta(t)\} \,\big|\, T = t\big],$$

(X, T) is a generic pair (X_i, T_i), $f_j(x|\theta) = (\partial/\partial\theta^{(j)})\, f(x|\theta)$, λ is proportional to the intensity of the point process that generated the T_i's, and $\mathscr{J}$ is the time interval over which we wish to optimize performance of $\bar{\theta}$.

As we shall show, minimizing D_1 is asymptotically equivalent to minimizing

$$\text{(3.4)}\qquad \begin{aligned} D_2 \equiv \int_{\mathscr{J}} E\Big(\big[f\{X|\bar{\theta}(t)\} \\ - f\{X|\theta(t)\}\big]^2 \,\Big|\, T = t\Big)\, \lambda(t)\, dt, \end{aligned}$$

where we take (X, T) to be independent of the dataset from which we computed $\bar{\theta}$. The integral in (3.4) may be written as

$$\text{(3.5)}\qquad \begin{aligned} E\bigg[\int dx \int_{\mathscr{J}} f\{x|\bar{\theta}(t)\}^2\, \lambda(t)\, dt \\ -2 \int dx \int_{\mathscr{J}} f\{x|\theta(t)\}\, f\{x|\bar{\theta}(t)\}\, \lambda(t)\, dt \bigg], \end{aligned}$$

plus a term that does not depend on $\bar{\theta}$ and so is immaterial to minimization. Let $\bar{\theta}_{-i}$ denote the version of $\bar{\theta}$ computed while omitting the data pair (X_i, T_i). Then, up to a constant of proportionality and treating the data pairs as independent, an empirical approximation to (3.5) is given by

$$\text{(3.6)}\qquad \begin{aligned} CV_1(h) \equiv \sum_{i=1}^{N} I(T_i \in \mathscr{J}) \int f\{x|\bar{\theta}_{-i}(T_i)\}^2\, dx \\ -2 \sum_{i=1}^{N} I(T_i \in \mathscr{J})\, f\{X_i|\bar{\theta}_{-i}(T_i)\}. \end{aligned}$$

We suggest choosing h to minimize $CV_1(h)$. Owing to problems that the local-constant estimator has with edge effects, $\mathscr{J}$ should, in the local-constant case, be chosen to be contained properly within $\mathscr{I}$.

A minor modification of this method enables us to select bandwidth when estimating some parameters locally and others globally. Adopting the notation $f(x|\eta, \psi)$ suggested in Section 3.1, with η and ψ denoting local and global parameters respectively, the analogue of the criterion at (3.6) is

$$\begin{aligned} CV_2(h) \equiv \sum_{i=1}^{N} I(T_i \in \mathscr{J}) \int f\{x|\bar{\eta}_{-i}(T_i), \hat{\psi}_{-i}\}^2\, dx \\ -2 \sum_{i=1}^{N} I(T_i \in \mathscr{J})\, f\{X_i|\bar{\eta}_{-i}(T_i), \hat{\psi}_{-i}\}, \end{aligned}$$

where the subscript "$-i$" denotes that version of an estimator has been computed with datum (X_i, T_i) omitted from the sample. We select h to minimize $CV_2(h)$.

To appreciate why D_2 is asymptotically equivalent to D_1, write $\nabla f(x|\theta)$ for the d-vector of which the jth element is $(\partial/\partial\theta^{(j)})\, f(x|\theta)$, and note that by Taylor expansion of $f(x|\theta)$ with respect to θ,

$$\begin{aligned} D_2 &\sim \int_{\mathscr{J}} E\Big[\Big(\big[\nabla f\{X|\bar{\theta}(t)\}\big]^T (\bar{\theta} - \theta)\Big)^2 \,\Big|\, T = t\Big]\, \lambda(t)\, dt \\ &\sim \int_{\mathscr{J}} E\Big[\Big(\big[\nabla f\{X|\theta(t)\}\big]^T (\bar{\theta} - \theta)\Big)^2 \,\Big|\, T = t\Big]\, \lambda(t)\, dt \\ &= D_1. \end{aligned}$$

4. APPLICATIONS

First we apply the methods suggested in Section 3 to the Swedish windstorm data introduced in Section 2. In our initial analysis we fitted the Pareto model,

$$\text{(4.1)}\qquad \begin{aligned} F(x|\beta, c) &\equiv P(X \le x) \\ &= 1 - c\, x^{-\beta}, \qquad x > c^{1/\beta}, \end{aligned}$$

where β and c are positive constants. By conditioning on the smallest observed data value we were able to take $c = 1$, so that the only unknown parameter, $\theta = \beta$, represented distribution shape. Statements about the size of bandwidth should be interpreted on the scale of the interval $\mathscr{I} = [0, 1]$, to which we transformed the time period [1982, 1993]. Throughout our analysis of both the windstorm and the temperature data we used the biweight kernel, $K(x) = (15/16)(1 - x^2)^2$ for $|x| \le 1$. The cross-validation argument suggested in Section 3.4 produced bandwidths $h = 0.38$ and $h = 0.68$ in local-constant and local log-linear cases, respectively.

Figure 2 shows a plot of the local-linear estimate of β with $h = 0.68$. It might perhaps be argued that the most extreme storm, at time $t = 0.96$, represents an outlier from a contaminant distribution. Its influence on our analysis is only minor, however, and in fact if this storm is removed from the data then the plot is virtually unchanged, the main effect being that the curve decreases a little less steeply on the right-hand side of the mode at $t \approx 0.3$. A local-constant plot is also similar, although there the curve increases less steeply on the left-hand side of the mode.

Plots with smaller bandwidths also have the same features, except that (i) there is a tendency for a shoulder to appear on the right-hand side of the peak at about $t = 0.7$ (approximately the year 1990), and (ii) the plots exhibit cusps and other fluctuations due to problems with design sparsity and with the smallest data value (on which we conditioned) disappearing from the local dataset as t is moved along the axis. The latter problem is always

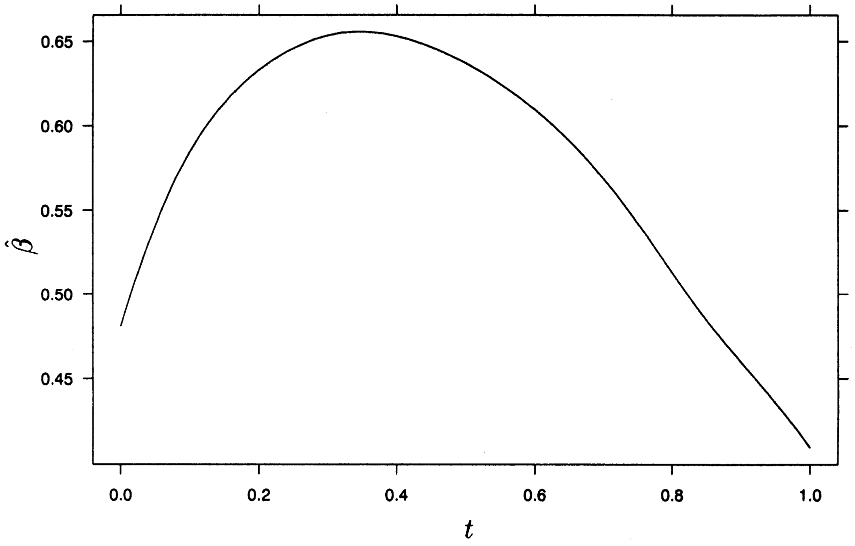

FIG. 2. *Pareto model fitted locally to windstorm data. The figure depicts the local-linear estimate of the shape parameter, β, in the case of a fitted Pareto model* [*see* (4.1)] *and with bandwidth $h = 0.68$ chosen by cross-validation.*

potentially troublesome for distributions, such as those connected with extremes, where the support depends on unknown parameter values. Although it did arise occasionally in our analysis (see for example the cusp at $t = 0.25$ in Figure 6), it was not as prevalent or as serious as we had anticipated.

Figure 3 gives local probability plots for the Pareto fit, using $h = 0.68$. A more conventional global probability plot, with β fitted globally, demonstrates an extremely poor fit.

Figure 2 suggests a general tendency for the severity of storms to at first decrease up until about the end of 1986 and then start to increase. This general behavior is also borne out by our second analysis, which, moreover, lends support to the suggestion that a period of very slow change in average storm intensity occurred during the late 1980s and early 1990s; see point (i) two paragraphs above. In the second analysis we fitted the generalized Pareto distribution (GPD),

$$F(x|\gamma, \sigma) = 1 - \{1 + (\gamma/\sigma)\, x\}_{+}^{-1/\gamma}, \tag{4.2}$$

where $x_{+} = \max(x, 0)$, $\sigma > 0$ denotes scale, γ represents shape and the support of the distribution is the positive half-line for $\gamma > 0$ and the interval $0 < x < -\sigma/\gamma$ for $\gamma < 0$. When $\gamma = 0$ we interpreted F as the exponential distribution, $1 - \exp(-x/\sigma)$. Cross-validation suggested the bandwidth $h = 0.17$ when using the local-constant method and fitting γ and σ together [see Figure 4(a)]. In the local-linear case, direct computation of the cross-validation criterion and of the estimators themselves was seriously hindered by problems with sparse design; bear in mind that in the local-linear case we are in effect fitting four continuous functions simultaneously. For simplicity and brevity, rather than employ remedial methods to overcome these problems, we shall confine attention to local-constant fitting in the GPD context.

Figure 4(b) depicts a plot of the local-constant function estimate $\hat{\sigma}$ when γ is also estimated locally, using bandwidth $h = 0.20$. (We chose h a little larger than the bandwidth recommended by cross-validation, so as to reduce "wiggliness" of the curve estimate on the plateau.) Estimated scale at first decreases to a minimum in mid-1985 and then virtually increases monotonically for the rest of the period, except for a plateau between 1988 and 1991. The same shape and almost identical locations of the trough and of endpoints of the plateau are observed if σ is fitted locally and γ fitted globally and also if the storm at $t = 0.96$ is removed, the main difference in the latter case being that the trough is deeper. Figure 5 is a local probability-plot for the local-constant fit with $h = 0.20$, when both γ and σ vary locally.

Because increasing scale corresponds to greater storm severity, we conclude from both the Pareto and GPD analyses that storm severity at first decreased from 1982 to 1985 or 1986 and then generally increased until the end of the data set in 1993, with intensity varying relatively little for about three years from 1988. These results shed new light on the nonstatistically significant linear trend fitted by Rootzén and Tajvidi (1997). The results suggest that variations in storm intensity between 1982 and 1993 most likely were not monotone increasing over the full interval, although they

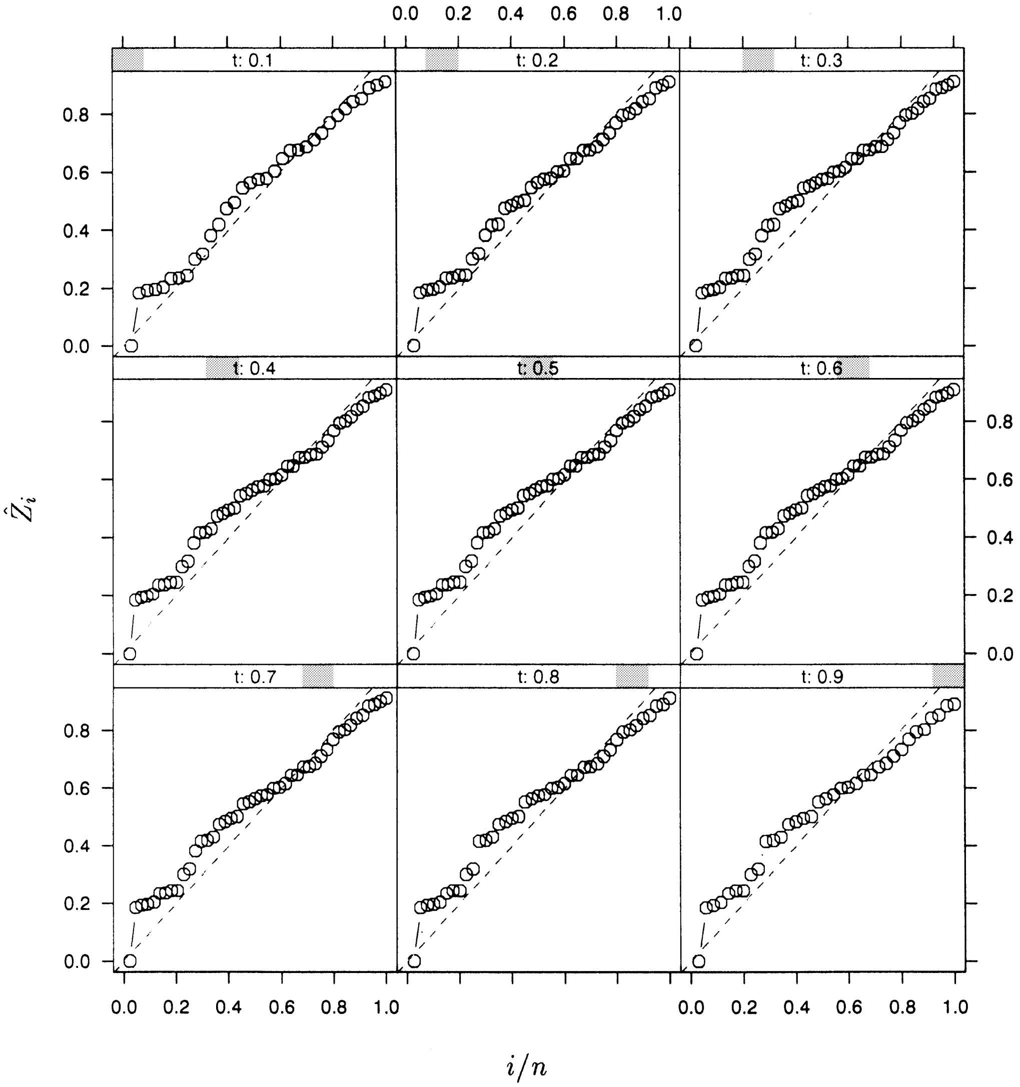

FIG. 3. *Local probability plots for Pareto fit. Working from left to right and top to bottom, the nine panels correspond to the nine values of* $t = 0.1(0.1)0.9$. *In each panel,* n *is the number of observations in* $[t - h, t + h]$ *and the dashed line corresponds to the equation* $y = x$.

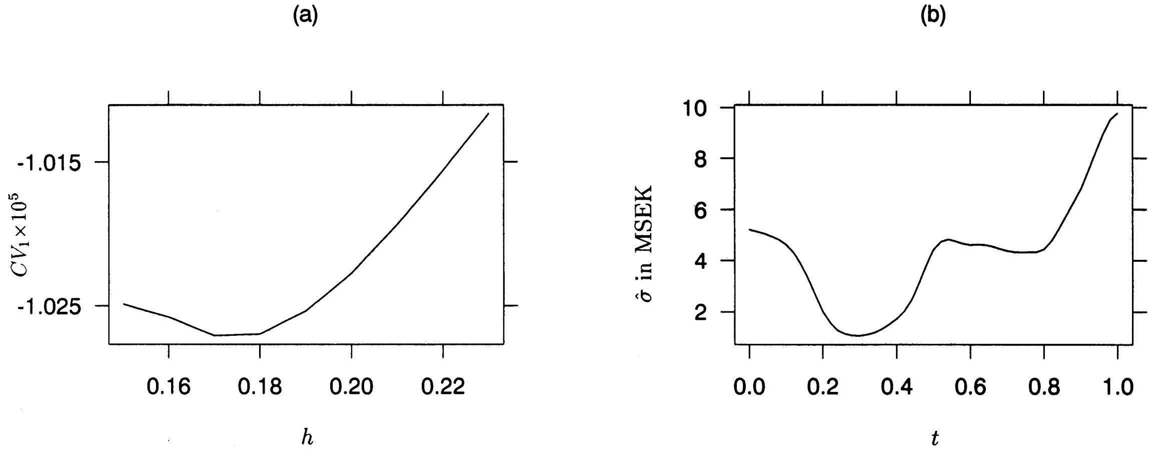

FIG. 4. *GPD model fitted locally to windstorm data. Panel* (a) *graphs the cross-validation criterion* $CV_1(h)$ *in the case of fitting a GPD model* [*see* (4.2)] *by local-constant smoothing, when both* γ *and* σ *are fitted locally. Panel* (b) *shows the corresponding local-constant estimate of* σ, *with bandwidth increased slightly to* $h = 0.20$ *relative to that suggested by panel* (a).

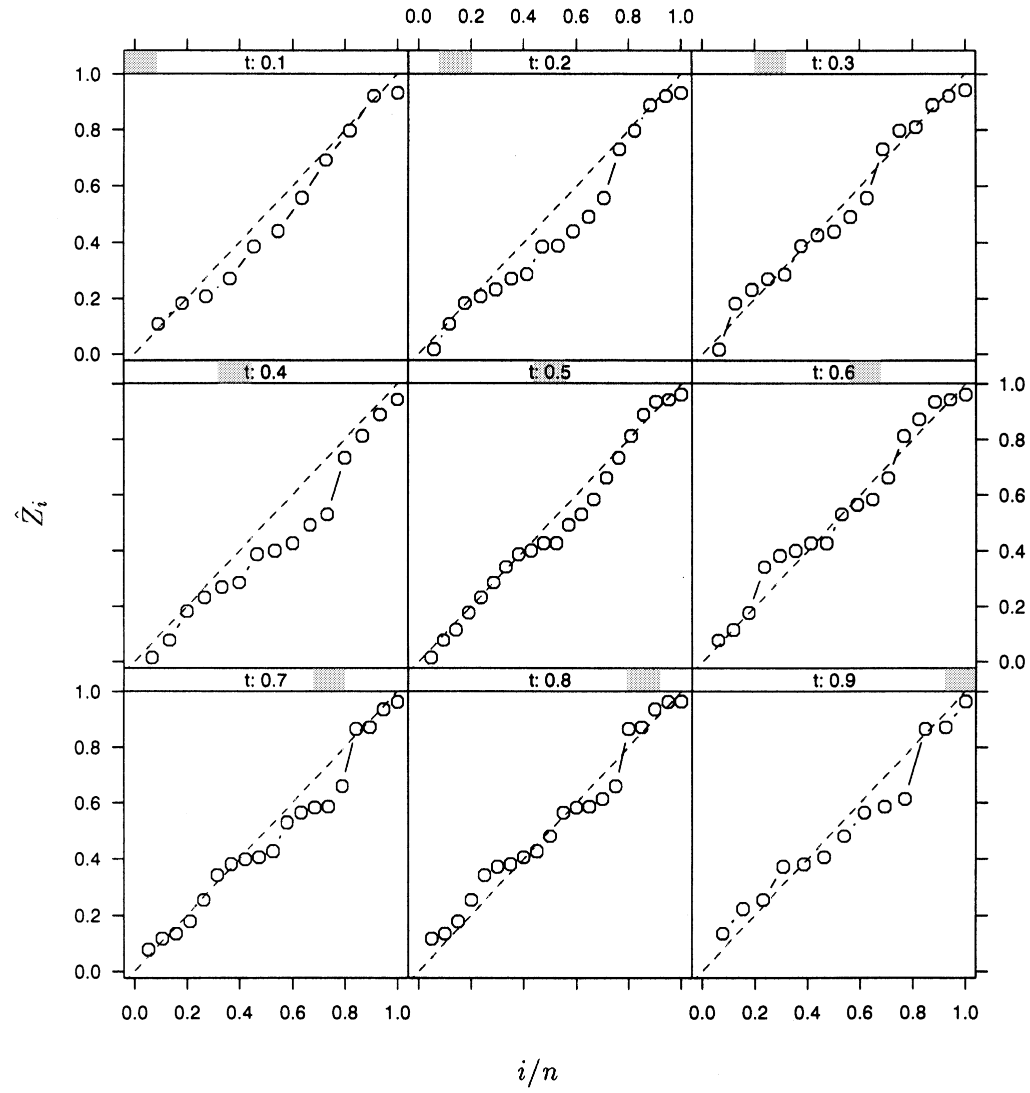

FIG. 5. *Local probability-plots for GPD fit. Panels are for* $t = 0.1(0.1)0.9$, *and illustrate goodness of fit of the two-parameter* GPD *model addressed in Figure* 4. *In each panel, n is the number of observations in* $[t-h, t+h]$ *and the dashed line corresponds to the equation* $y = x$.

appear to have been increasing during the last two-thirds of that period.

Next we address the temperature data for Victoria, Australia, introduced in Section 2. We standardized so that [1910, 1993] was transformed to the interval $\mathscr{I} = [0, 1]$. Bandwidth should be interpreted on this scale. Panels (a)–(c) of Figure 6 show plots of local-constant estimates of the components of $\theta = (\mu, \sigma, \gamma)^T$, representing location, scale and shape, respectively, derived from fitting a generalized extreme-value (GEV) distribution with distribution function

$$(4.3) \qquad F(x|\theta) = \exp\left[-\{1 + (\gamma/\sigma)(x-\mu)\}_+^{-1/\gamma}\right].$$

The support of the distribution is $x < \mu - (\sigma/\gamma)$ if $\gamma < 0$, and $x > \mu - (\sigma/\gamma)$ for $\gamma > 0$. When $\gamma = 0$ we interpreted $F(x|\theta)$ as the limit, $\exp[-\exp\{-(x-\mu)/\sigma\}]$. The quality of the fit is illustrated in Figure 7.

The bandwidth for panels (a)–(c) of Figure 6, $h = 0.21$, was chosen by cross-validation. It represents the second, and lowest, local minimum in a graph of $CV_1(h)$ for $0 < h \le 1$. That function does assume lesser values for larger values of h, but none of them is a turning point; the function is monotone decreasing there. This type of behavior is well known in nonparametric density estimation (see, e.g., Hall and Marron, 1991), and our choice of the second minimum would be standard in that setting. The cusp problems at about $t = 0.25$ in Figure 6 are caused by the data sparseness difficulty noted in the case of windstorm data and diminish if a larger bandwidth is used.

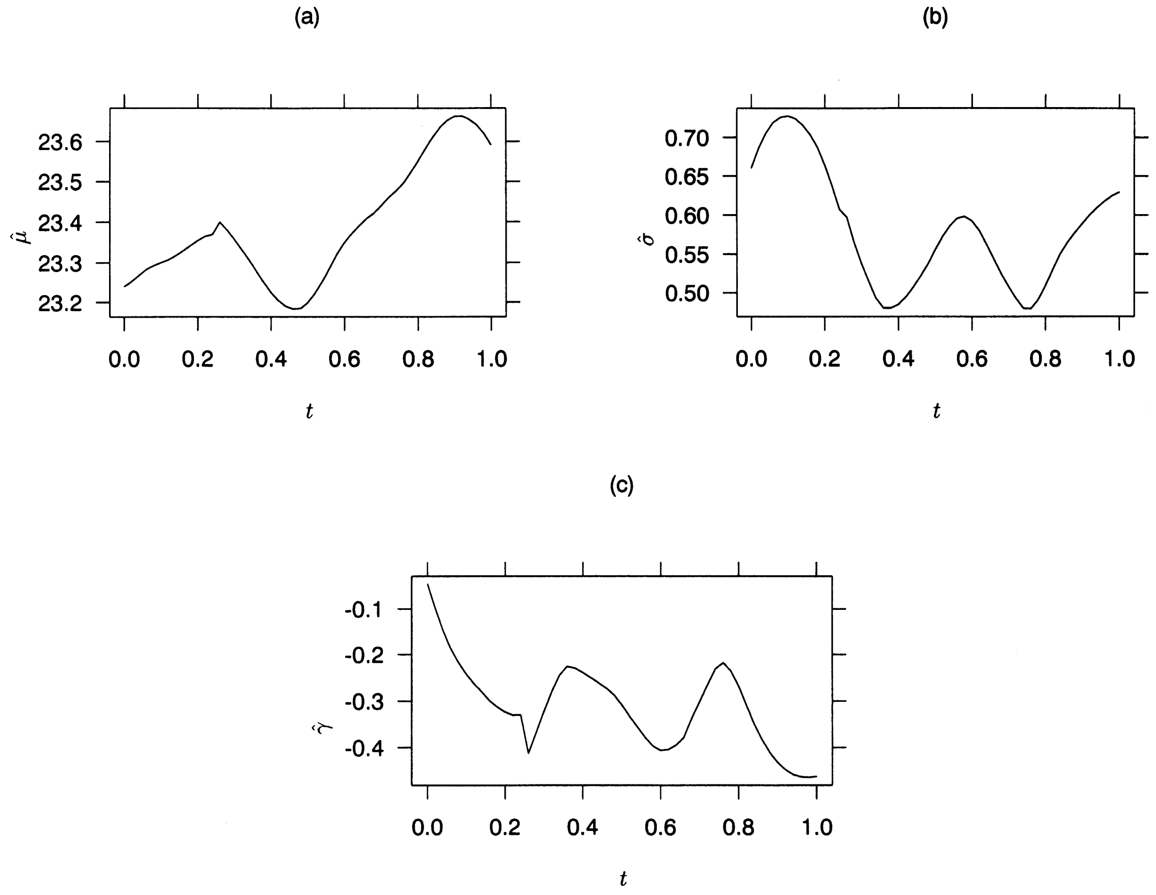

FIG. 6. *Parameter plots for temperature-data example. Panels* (a)–(c) *show local-constant estimates of μ, σ and γ, respectively, derived by fitting the GEV distribution in* (4.3).

Fitting all three components of θ by local-linear methods requires fitting six continuous functions simultaneously and not surprisingly suffers from more serious data sparseness problems. However, local-linear methods work well if only two parameters—μ and σ, say—are fitted within any one time window. The associated global value of γ is -0.251, and the cusp problems noted above do not arise.

Graphs of estimates of μ and σ are similar for both fits. In particular, graphs of estimates of μ are multimodal, with a minimum at a point corresponding to about 1950, maxima at about 1933 and 1983 and the latter maximum higher than the former. Note that in the three-parameter fit, $\hat{\gamma}$ tends to change in the opposite direction to $\hat{\sigma}$, thereby accommodating some of the temporal changes in scale. As a result, the amplitude of fluctuations of estimates of σ is greater for the two-parameter fit than for the three-parameter fit. However, the pattern of peaks and troughs in estimates of σ is the same, and the places where they occur are almost identical.

Broadly similar estimates of μ and σ are also obtained by fitting a Gaussian model, although the data are somewhat skewed in the right tail. A significant advantage of fitting the GEV distribution is that it provides information about the upper endpoint of the distribution of admissible temperatures.

These results provide substantially more information than is available from merely fitting a linear trend to the data. Indeed, the marked nonmonotonicity of the adaptively fitted trend would indicate that monotone trend models, suggested by the hypotheses that a steadily warming trend in temperatures was evident early this century, are not appropriate for Victorian data during the period 1910–1993. Further support for this view may be obtained from analysis of maximum temperature data for the eastern half of Australia, which also exhibited a minimum at about midcentury; see Torok and Nicholls (1996). This was associated with a steady increase in rainfall, to a maximum in about 1950 (Nicholls and Lavery, 1992). The multimodal character that we have observed for maximum Victorian temperatures is not

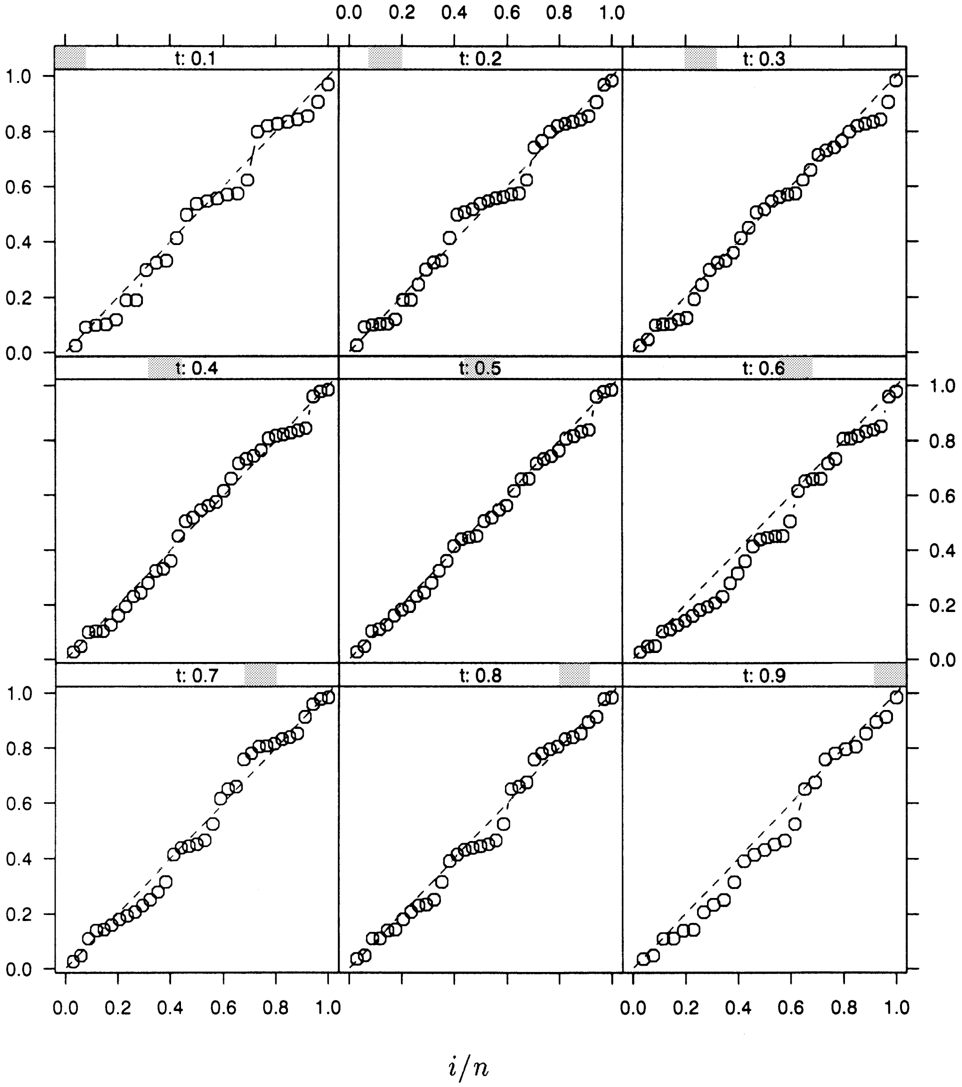

FIG. 7. *Local probability-plots for GEV fit. Panels are for* $t = 0.1(0.1)0.9$, *and describe goodness of fit of the three-parameter GEV distribution fit illustrated in Figure* 6. *In each panel, n is the number of observations in* $[t-h, t+h]$ *and the dashed line corresponds to the equation* $y = x$.

clearly evident across eastern Australia as a whole, however.

On the other hand, trends in maximum temperatures recorded at weather stations in the western half of Australia are commonly upwards through most of the twentieth century, with the result that the average annual maximum Australian temperature, discussed by Torok and Nicholls (1996), evidences behavior quite different from that observed in the eastern half. The average annual maximum is derived by taking the mean of maximum annual temperature readings at 224 weather stations across Australia. These averages are plotted in Figure 8(a) for data through the period 1890–1993. (For later plots we have transformed the period to the interval [0, 1].) By virtue of the central limit theorem, average maxima would be expected to fit a Gaussian distribution well. The local probability plots that result after local-constant and local-linear fits of both location and scale under a Gaussian model strongly support this claim.

The mean curve is generally increasing, although with a slight decrease at the end of the nineteenth century and a plateau or slight dip in the middle of the twentieth century. The scale curve is bimodal, with its trough at about 1940 and its peaks at about 1910 and 1960, the former being more pronounced. See panels (b) and (c) of Figure 8, which depict local-linear fits. Cross-validation suggested the bandwidth $h = 0.15$; see panel (d) of the figure. (There is also a local minimum of CV_1 at the

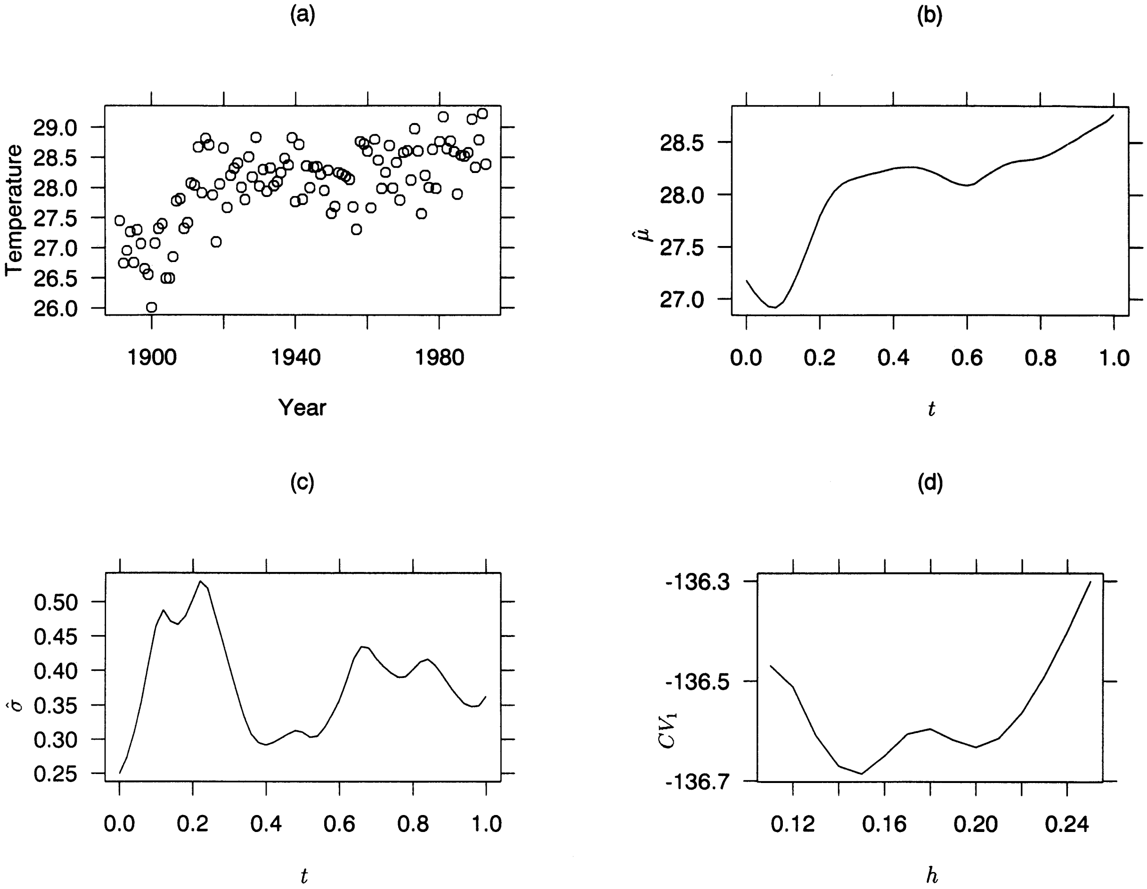

FIG. 8. *Location and scale estimates for average Australian maximum temperature. "Average annual maximum temperature" is defined in Section* 4. *Panel* (a) *shows the raw data, and panels* (b) *and* (c) *plot the mean* (μ) *and standard deviation* (σ), *respectively, for local-linear fits to a Gaussian model. Bandwidth* $h = 0.15$ *was suggested by the cross-validation curve plotted in panel* (d).

inordinately small value $h = 0.04$, which, again following Hall and Marron, 1991, we ignored.)

5. THEORETICAL PROPERTIES

5.1 Properties of Estimators When the Model Is Correct

We begin by describing a time-series model for the data (X_i, T_i). Let $\{(Y_i, S_i),\ i \geq 1\}$ be a stationary time series, and put

$$\pi(j) \equiv \sup_{-\infty<i<\infty} E\left\{ \sup_{A \in \mathscr{F}_{i+j}^{\infty}} \left|P(A|\mathscr{F}_1^i) - P(A)\right| \right\},$$

where $\mathscr{F}_i^j$ denotes the σ-field generated by $\{(Y_k, S_k),\ i \leq k \leq j\}$. Given an integer $\nu \geq 1$, and a compact interval $\mathscr{I}$ which we take without loss of generality to have unit length, let $\{(X_i, T_i),\ 1 \leq i \leq N\}$ denote those values of (Y_j, S_j) for which $1 \leq j \leq \nu$ and $S_j \in \mathscr{I}$.

Next we introduce notation. Let (X, T) and (Y, S) be generic values of (X_i, T_i) and (Y_i, S_i), respectively, and put $g(\cdot|\theta) = \log f(\cdot|\theta)$, where $f\{\cdot|\theta(t)\}$ is the density of Y conditional on $S = t \in \mathscr{I}$ (equivalently, of X conditional on $T = t$). Define $\dot{\theta}^{(j)}(t) = d\theta^{(j)}(t)/dt$, $\ddot{\theta}^{(j)} = d^2\theta^{(j)}(t)/dt^2$,

$$g_j(\cdot|\theta) = \left(\partial/\partial\theta^{(j)}\right) g(\cdot|\theta),$$
$$g_{j_1, j_2}(\cdot|\theta) = \left(\partial^2/\partial\theta^{(j_1)}\partial\theta^{(j_2)}\right) g(\cdot|\theta).$$

Let $V = V(t)$ denote the inverse of the $d \times d$ matrix of which the (j_1, j_2)th element is

$$E\left[g_{j_1}\{X|\theta(T)\}\, g_{j_2}\{X|\theta(T)\} \,\middle|\, T = t\right]$$
$$= -E\left[g_{j_1, j_2}\{X|\theta(T)\} \,\middle|\, T = t\right].$$

Put $\kappa_2 = \int u^2 K(u)\, du$, $\kappa_k^+ = \int_{u>0} u^k K(u)\, du$, $\kappa = \int K^2$ and

$$\kappa^+ = 4 \int_0^{\infty} \left\{1 - \left(\kappa_1^+/\kappa_2^+\right) u\right\}^2 K(u)^2\, du.$$

Assume that (a) $\sum_{j\geq 1} j^2\, \pi(j)^{\varepsilon} < \infty$ for some $\varepsilon \in [0, 1)$ (we interpret 0^{ε} as 0 when $\varepsilon = 0$, so that the case $\varepsilon = 0$ corresponds to m-dependence); (b) $p \equiv P(S_i \in \mathscr{I}) > 0$, the distribution of S_i, conditional on $S_i \in \mathscr{I}$, is absolutely continuous with density

ξ, say; (c) ξ has a continuous derivative in a neighbourhood of t and satisfies $\xi(t) > 0$; (d) for all values of θ in a neighbourhood of $\theta(t)$, $f(\cdot|\theta)$ satisfies the regularity conditions of Lehmann (1983, pages 329f), which are sufficient for the Cramér–Rao lower bound to be attained in the same neighborhood; (e) $\theta(\cdot)$ has two continuous derivatives in a neighborhood of t, and $V(\cdot)$ is nonsingular and continuous in that neighborhood; (f) K is a symmetric, compactly supported probability density and (g) $h = h(\nu) \to 0$ and $\nu \to \infty$ in such a manner that $\nu h \to \infty$. Then $\nu\lambda$, where $\lambda \equiv p\,\xi$, equals the intensity of the point process $\{T_1, \dots, T_N\}$ on $\mathscr{I}$. In formulating the results below we suppress the argument, t, of λ, θ, $\dot{\theta}$, $\ddot{\theta}$ and V.

THEOREM 5.1. *Assume conditions* (a)–(g).

(i) *Local-constant estimator. If t is an interior point of $\mathscr{I}$ then $\hat{\theta} = \hat{v}_0$ satisfies*

$$\begin{aligned}\hat{\theta} &= \theta + \kappa_2\, h^2 \left\{\tfrac{1}{2}\,\ddot{\theta} + (\lambda'/\lambda)\,\dot{\theta}\right\} \\ &\quad + (\nu\lambda h)^{-1/2}\, Z + o_p\left\{h^2 + (\nu h)^{-1/2}\right\},\end{aligned} \tag{5.1}$$

where Z is asymptotically Normal $N(0, \kappa V)$.

(ii) *Local-linear estimator. If t is an interior point of $\mathscr{I}$ then $\tilde{\theta} = \tilde{v}_0$ satisfies*

$$\begin{aligned}\tilde{\theta} &= \theta + \tfrac{1}{2}\,\kappa_2\, h^2\, \ddot{\theta} \\ &\quad + (\nu\lambda h)^{-1/2}\, Z + o_p\left\{h^2 + (\nu h)^{-1/2}\right\},\end{aligned} \tag{5.2}$$

where again Z is asymptotically Normal $N(0, \kappa V)$. If t is an endpoint of $\mathscr{I}$, say the lower endpoint where $\lambda(t+) > 0$ and $\lambda(t-) = 0$, then

$$\begin{aligned}\tilde{\theta} &= \theta + \tfrac{1}{2}\left\{(\kappa_2^+)^2 - \kappa_1^+\kappa_3^+\right\}\left(\kappa_2^+\right)^{-1} h^2\, \ddot{\theta} \\ &\quad + (\nu\lambda h)^{-1/2}\, Z + o_p\left\{h^2 + (\nu h)^{-1/2}\right\},\end{aligned} \tag{5.3}$$

where now Z is Normal $N(0, \kappa^+\, V)$.

REMARK 5.1 (Other approaches to modelling the data). Our model for the way in which the data (X_i, T_i) might be generated by a stationary process (Y_i, S_i) is only one of several that might be considered. We could treat the X_i's as a time series conditional on the set $\{T_1, \dots, T_N\}$, and ask that the T_i's form a point process of a specific type, for example a Poisson cluster process or a sequence of points arrayed on a regular grid within the interval $\mathscr{I}$. The latter is the case for our Australian temperature data, and there λ should be treated as identically constant on $\mathscr{I}$.

In contexts of that type the nature of dependence may be such that asymptotic variance matrices differ from those given in Theorems 5.1 and 5.2. However, the bias terms would be the same to first order, since under condition (d) the estimators are asymptotically linear in the data, and expected values of the components of a time series do not depend on the nature or strength of dependence. Furthermore, under weak dependence the orders of magnitude of the stochastic terms would also be unchanged. This may be seen most easily in the case of m-dependent data, where, in view of the aforementioned asymptotic linearity, the limiting variance is inflated by a factor of at most $m + 1$. The fact that the gridpoints in the Australian temperature example represent years, rather than (for example) months, suggests that stochastic relationships among X_i's, even for nearby T_i's, should be small. Analysis of correlations supports this conjecture and so encourages use of the methods suggested in Section 3.

REMARK 5.2 (Terms corresponding to bias and variance). The terms of size h^2 on the right-hand sides of (5.1), (5.2) and (5.3) represent the dominant contributions to systematic error, or bias, in those respective contexts. The terms of size $(\nu h)^{-1/2}$ denote the dominant contributions to stochastic error, or error about the mean. Because the stochastic and systematic error terms are of identical orders when h is of size $\nu^{-1/5}$, then this is the optimal size of bandwidth. The variance matrix of Z is of course the maximum-information variance associated with the Cramér–Rao lower bound.

REMARK 5.3 (Edge effects). Result (5.1) fails if t is at either end of $\mathscr{I}$, because there the second term on the right-hand side of (5.1) is $O(h)$ rather than $O(h^2)$. Comparing (5.2) and (5.3) we see that, although t being at an end of $\mathscr{I}$ has had some impact on terms in the asymptotic expansion, it does not affect their orders of magnitude. This is in marked contrast to the local-constant case.

REMARK 5.4 (Globally estimated parameters). Suppose one or more parameters (components of ψ) are estimated globally, and the other parameters (components of θ) are estimated locally. Let ψ^0 denote the true value of ψ. Then,

$$\tilde{\theta}_{\hat{\psi}} = \tilde{\theta}_{\psi^0} + O(h^2) + o_p\left\{h^2 + (\nu h)^{-1/2}\right\},$$

where the "$O(h^2)$" term is purely deterministic. In particular, the asymptotic variance of $\tilde{\eta}_{\hat{\psi}}(t)$ equals that of the "ideal" form $\tilde{\eta}_{\psi^0}(t)$, although the bias components of these quantities differ in terms of order h^2. [This is also true for local-constant estimators, provided that terms that would cause boundary problems are dropped from the series at (3.2).] That this is correct even for time-series data, satisfying condition (a), is a consequence of the fact that the stochastic error of $\hat{\psi}$ is negligible relative to the stochastic error of estimators of θ.

REMARK 5.5 (Cross-validation). It may be proved that the empirical bandwidth chosen by cross-validation, suggested in Section 3.4, is of size $\nu^{-1/5}$. Indeed, if $\hat{h}$ denotes the bandwidth that minimizes CV_1, then $\nu^{1/5}\hat{h}$ converges in probability to a finite, nonzero constant. When all parameters are estimated locally, the constant (C, say) is such that $C\nu^{-1/5}$ produces asymptotic minimization of D_1, defined at (3.3); see also Remark 5.2. This result has a straightforward analogue in the case where some parameters are estimated locally and others globally. If the variables (Y_i, S_i) used to generate the (X_i, T_i)'s are sufficiently weakly dependent, if $\mathscr{J} = \mathscr{I}$ and if estimators are defined by local-linear means, then sufficient regularity conditions are (a)–(g) combined with Hölder continuity of K. Methods used to derive this result may be adapted from work of Hart and Vieu (1990).

5.2 Properties of Estimators under a Nominal Gaussian Model

Define $\theta = (\mu, \sigma)^T$ and fit the model $f(x|\theta) = \sigma^{-1}\phi\{(x-\mu)/\sigma\}$, where $\phi(z) = (2\pi)^{-1/2}\exp(-\frac{1}{2}z^2)$. However, in contradistinction to Section 5.1, we do not ask that the distribution of X, given $T = t$, actually have density $f\{\cdot|\theta(t)\}$.

Define κ, κ_2 as in Section 5.1, let (X, T) denote a generic value of (X_i, T_i), put $U = \{X - \mu(T)\}/\sigma(T)$, and let $W = W(t)$ be the variance matrix of $(\sigma U, \frac{1}{2}\sigma^2 U^2)^T$ conditional on $T = t$, where $\sigma = \sigma(t)$. To conditions (a)–(g) introduced in Section 5.1, adjoin the following assumptions: (h) the functions $\mu(u) \equiv E(X|T = u)$ and $\sigma(u)^2 \equiv \operatorname{var}(X|T = u)$ both have two continuous derivatives within a neighborhood of $u = t$, and (i) for some $\varepsilon > 0$, $E(|X|^{4+\varepsilon}|T = u)$ is bounded uniformly in u within some neighbourhood of t. Let $\hat{\theta} = (\hat{\mu}, \hat{\sigma})^T$ denote the local-constant estimator of θ, and let $\tilde{\theta} = (\tilde{\mu}, \tilde{\sigma})^T$ be the local-linear estimator.

THEOREM 5.2. *Assume conditions* (a), (b) *and* (e)–(i). (i) *Local-constant estimator. If t is an interior point of $\mathscr{I}$ then* (5.1) *holds, where now Z is Normal $N(0, \kappa W)$.* (ii) *Local-linear estimator. If t is an interior point of $\mathscr{I}$, then* (5.2) *holds for the same interpretation of Z as before. If t is an endpoint of $\mathscr{I}$, say the lower endpoint where $\lambda(t+) > 0$ and $\lambda(t-) = 0$, then* (5.3) *is true, where now Z is Normal $N(0, \kappa^+ W)$.*

The theorem fails to hold for a general standardised density ϕ.

5.3 Technical Arguments

We conclude with general remarks about proofs of Theorems 5.1 and 5.2. The arguments are relatively straightforward when the pairs (Y_i, S_i)—and hence also the data (X_i, T_i)—are independent. Modifications for the time-series case follow conventional lines. In brief, the method is based on Taylor expansion, in which we isolate stochastic and deterministic terms. For example, starting with (3.1) we write $\ell(v_0, v_1|t)$ as a Taylor expansion in powers of $\Delta_i = (\Delta_i^{(j)}) \equiv \omega_i(v_0, v_1) - \theta(T_i)$. The terms that arise have the form

$$(5.4) \qquad \sum_{i=1}^{N} g_{j_1,\ldots,j_r}(X_i|\theta_i)\,\Delta_i^{(j_1)}\cdots\Delta_i^{(j_r)}\,K_i(t),$$

where $g_{j_1,\ldots,j_r}(x|\theta) = (\partial^r/\partial\theta^{(j_1)}\cdots\partial\theta^{(j_r)})\,g(x|\theta)$ and $\theta_i = \theta(T_i)$. The expected value of the quantity at (5.4) is, trivially, the same in the cases of time-series and independent data. That the quantity satisfies a central limit theorem with identical asymptotic variances in these two cases may be seen by (i) appealing to results of Peligrad (1986, Theorem 1.7) to establish asymptotic Gaussianity, with asymptotic variance equal to variance of the series at (5.4); (ii) expressing the latter variance in the form $A + B$, where

$$A = \nu\lambda\operatorname{var}\{g_{j_1,\ldots,j_r}(X_i|\theta_i)\,\Delta_i^{(j_1)}\cdots\Delta_i^{(j_r)}\,K_i(t)\}$$

equals the variance that would arise in the case of independent data, and B represents cross-product terms arising on account of correlation; (iii) using methods of Yoshihara (1976) to bound B by a quantity that, in view of our condition (a), equals $o(\nu h)$ as $\nu \to \infty$ and (iv) using elementary calculus to prove that A is asymptotic to νh multiplied by a quantity that does not depend on ν or h. See also Fan, Yao and Tong (1996) and Hall, Wolff and Yao (1999).

6. CONCLUSIONS

We have shown that, while elementary extreme-value models often provide a good "local" fit to data distributions, simple parametric models for temporal trend can give a particularly misleading impression of the way in which distributions of meteorological or environmental data vary with time. For example, they can suggest a gradual monotone trend in storm intensity or annual maximum temperature, when it is more empirically plausible that these quantities at first decrease and then increase, relatively steeply, with time. We have suggested an alternative, nonparametric approach to estimating temporal trend and shown it to be particularly revealing when applied to two real data sets. Finally, we have outlined theoretical results which demonstrate that, under regularity conditions, our methods achieve high levels of performance in large samples.

ACKNOWLEDGMENTS

We are particularly grateful to Alex Kariko and Neville Nicholls, of the Bureau of Meteorology Research Centre in Melbourne, Australia, who provided both the temperature data and advice on its interpretation. Discussions with Alan Welsh and comments of the referee and Editor were also particularly helpful. We are grateful, too, to Anthony Davison for providing a copy of the paper Davison and Ramesh (2000). Nader Tajvidi's research was supported by a grant from the Swedish Foundation for International Cooperation in Research and Higher Education.

REFERENCES

Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998). Local estimating equations. *J. Amer. Statist. Assoc.* **93** 214–227.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*. Wadsworth, Pacific Grove, CA.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 597–610.

Davison, A. C. and Ramesh, N. I. (2000). Local likelihood smoothing of sample extremes. *J. Roy. Statist. Soc. Ser. B* **62** 191–208.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21** 196–216.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90** 141–150.

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83** 189–206.

Gu, C. (1990). Adaptive spline smoothing in non-Gaussian regression models. *J. Amer. Statist. Assoc.* **85** 801–807.

Gu, C., Bates, D. M., Chen, Z. and Wahba, G. (1989). The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal. Appl.* **10** 457–480.

Hall, P. and Marron, J. S. (1991). Local minima in cross-validation functions. *J. Roy. Statist. Soc. Ser. B* **53** 245–252.

Hall, P., Wolff, C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* **94** 154–163.

Hart, J. D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.* **18** 873–890.

Jones, P. D. (1994). Hemispheric surface air temperature variations: a reanalysis and an update to 1993. *J. Climate* **7** 1794–1802.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

Leadbetter, M. R. and Rootzén, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.* **16** 431–478.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

Nicholls, N. and Lavery, B. (1992). Australian rainfall trends during the twentieth century. *Internat. J. Climatology* **12** 153–163.

Peligrad, M. (1986). Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables. In *Dependence in Probability and Statistics* (E. Eberlein and M. S. Taqqu, eds) 193–223. Birkhäuser, Boston.

Raftery, A. E. (1989). Discussion of Smith (1989). *Statist. Sci.* **4** 378–381.

Rootzén, H. and Tajvidi, N. (1997). Extreme value statistics and windstorm losses: a case study. *Scand. Actuarial J.* 70–94.

Ruppert, D. (1995). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. Unpublished manuscript.

Severini, T. A. and Staniswallis, J. G. (1994). Quasilikelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501–511.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone (with discussion). *Statist. Sci.* **4** 367–393.

Staniswallis, J. G. (1989). The kernel estimates of a regression function in likelihood-based models. *J. Amer. Statist. Assoc.* **84** 276–283.

Torok, S. J. and Nicholls, N. (1996). A historical annual temperature dataset for Australia. *Austral. Met. Mag.* **45** 251–260.

Weisberg, S. and Welsh, A. H. (1994). Estimating the missing link function. *Ann. Statist.* **22** 1674–1700.

Yoshihara, K. (1976). Limiting behaviour of U-statistics for stationary absolutely regular processes. *Z. Wahrsch. Verw. Gebiete* **35** 237–252.