NISS

Modeling High Threshold Exceedances of Urban Ozone

Richard L. Smith and Li-Shan Huang

Technical Report Number 6
December, 1993

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

MODELING HIGH THRESHOLD EXCEEDANCES OF URBAN OZONE

by

Richard L. Smith and Li-Shan Huang

National Institute of Statistical Sciences
and
University of North Carolina

First version: December 28 1993. Research supported by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0. Address for correspondence: Richard L. Smith, Department of Statistics, University of North Carolina, Chapel Hill, N.C. 27599-3260; email RS@STAT.UNC.EDU.

Abstract. Urban ozone arises as a consequence of the emissions of nitrous oxides and hydrocarbons into the atmosphere, but it is also very strongly affected by meteorological conditions. In analyzing ozone data, it is hard to separate genuine trends, that might be explained by changes in the levels of emissions, from apparent ones determined by meteorology. Previous statistical studies of this problem have used classification and regression techniques. In particular, Bloomfield *et al.* (1993) have developed a detailed nonlinear regression model fitted to weighted averages of measured ozone concentrations at 45 monitoring stations in the Chicago area. Here, based on the same data set, we develop an alternative analysis aimed specifically at characterizing the probability of exceeding a high threshold.

Keywords: Ozone concentration, Meteorological adjustment, Regression, Extreme value analysis, Threshold exceedances.

1. Introduction

A major difficulty in the analysis of urban ozone data is how to separate genuine trends, that might be the consequence of changes in precursor emissions, from the effects of meteorological variability. This problem has been reviewed in Chapter 2 of the report of the National Research Council (1991), which discussed three broads approaches to it:

- *Measurement approaches:* Find some other ozone indicator (besides the present national ambient air quality standard, which is based on the number of daily ozone maxima exceeding a threshold of 120 ppb), which would be more "robust" against meteorological variability — for example, if might be beneficial to monitor the 80th or 95th percentile of the ozone distribution, or exceedances of lower thresholds;

- *Classification approaches:* Find a suitable classification of meteorological "types", and consider ozone exceedances separately within each type;

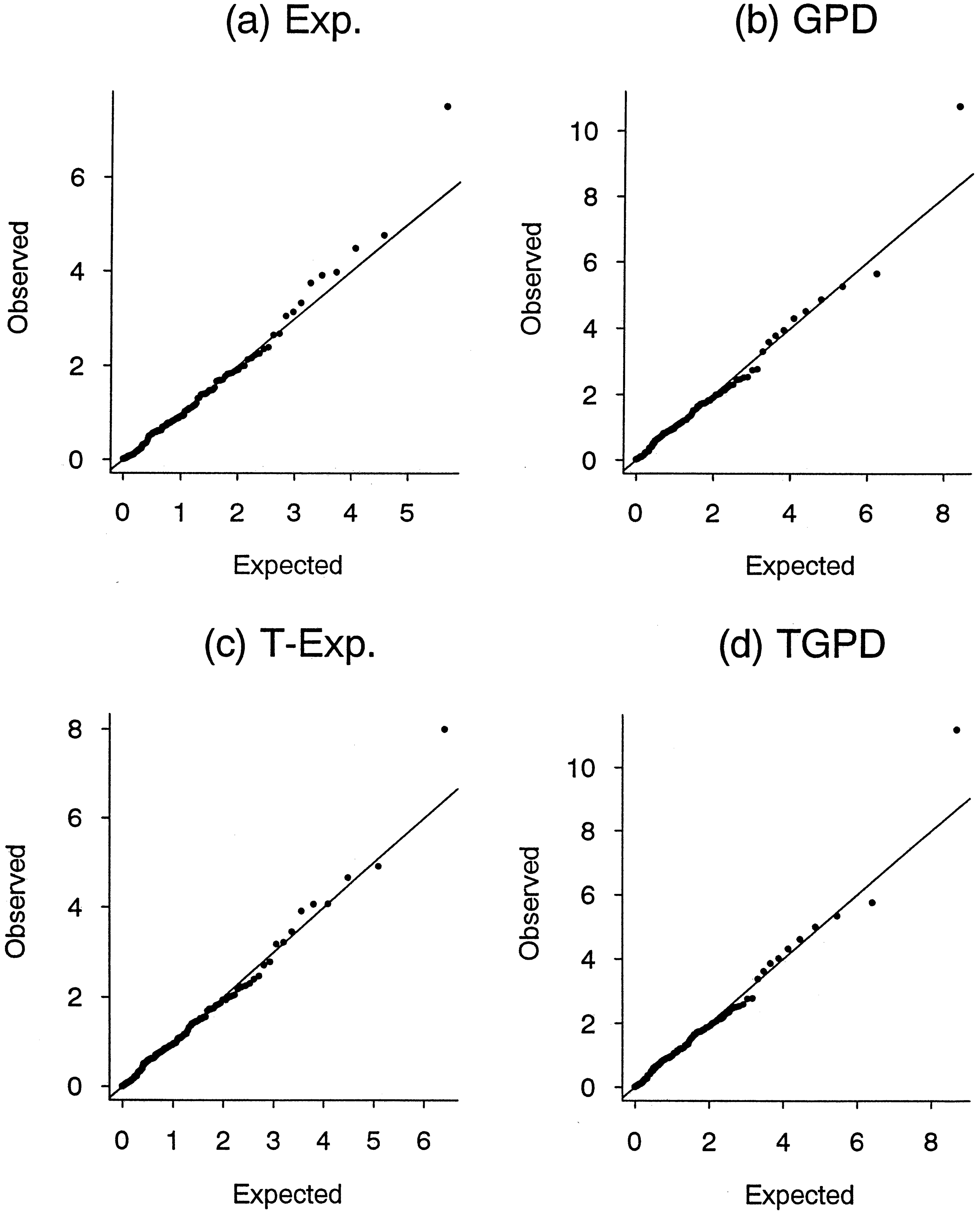
- *Regression approaches:* Construct a regression model linking ozone level to meteorological variables, and use that to identify days which have high ozone levels relative to the background meteorology.

Of the three approaches, the regression approach is the one most amenable to a detailed statistical analysis. In earlier work as part of the current project, Bloomfield, Royle & Yang (1993) used principal components analysis to determine suitable weighted averages of ozone levels at 45 urban stations in the Chicago area, and then constructed a nonlinear regression model to relate the resulting weighted average to measured meteorological variables. Their model will be described in more detail in Section 4.

One possible disadvantage of regression approaches is that they may fail to characterize well what is happening in the extremes of the distribution. Since the current ozone standard is defined in terms of extremes, this is a natural concern. The report of the National Research Council (1991) specifically mentioned this as a drawback of regression analysis, pointing out that high ozone levels have been underpredicted by such analyses, and suggesting that the explanation lies in the standard method of fitting a regression model, which is based on minimizing the sum of squares of residuals over the whole data set, not giving enough attention to the extremes of the data. Cox & Chu (1992) recognised this as a difficulty with standard regression analyses based on the normal distribution, and proposed an alternative model based on the Weibull distribution, using a maximum likelihood fit. The reports of Cox & Chu (1992) and Bloomfield *et al.* (1993) considered the power of their models to predict the 95th and 99th percentiles of the ozone distribution, with seemingly satisfactory results, but this leaves open the question of whether alternative analyses, focussed specifically on extreme quantiles or threshold crossings, could produce better results. The main purpose of the present paper is to consider alternative methods of analysis with just such an objective.

Extreme value theory is that branch of statistics concerned with distributions of extreme values in random samples. Classical extreme value theory is based on asymptotic

Fig. 8.4: Station R, Sums of excesses over 80



(a) Exp.

(b) GPD

(c) T-Exp.

(d) TGPD

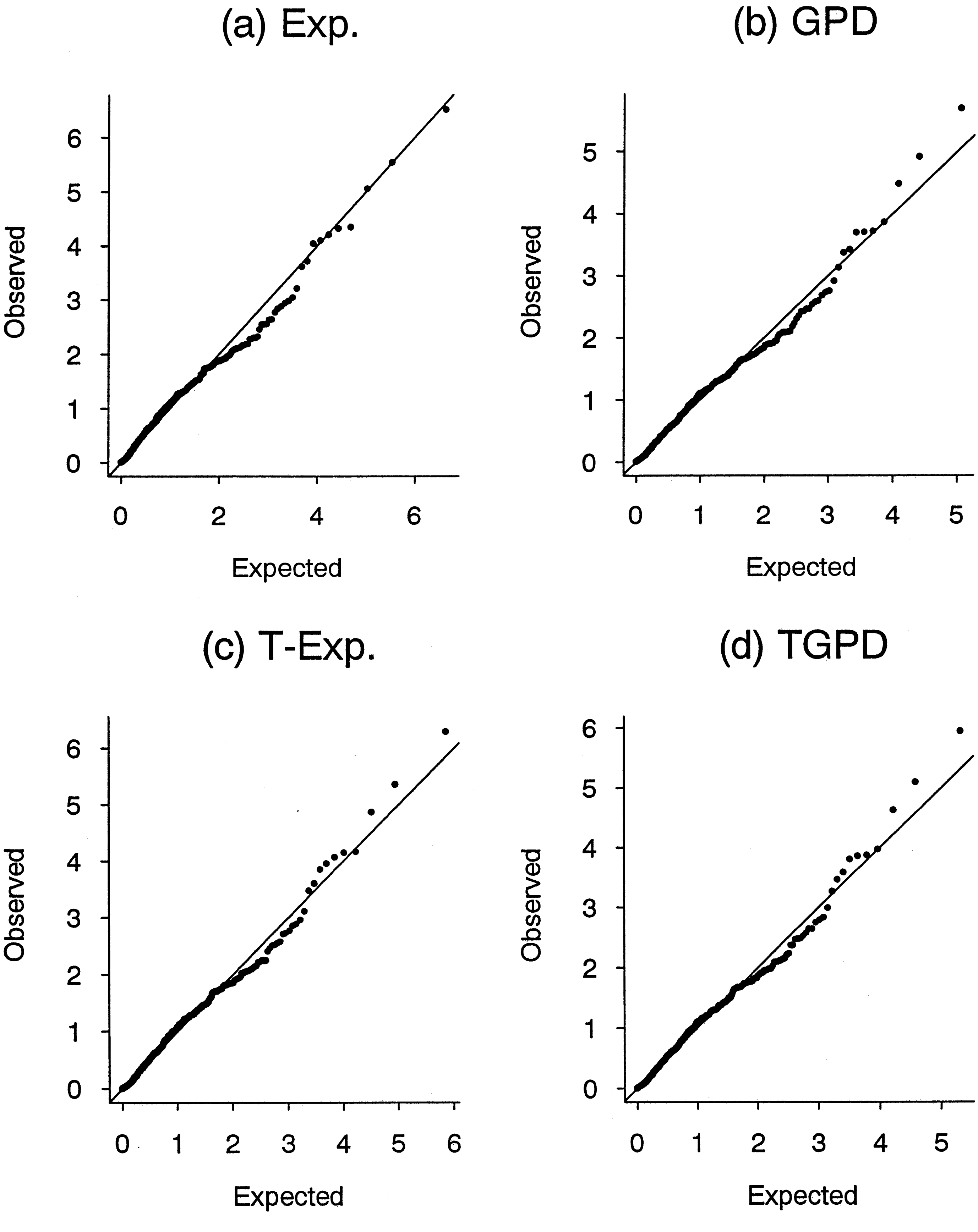
distributions for extremes in large samples (Gumbel 1958). Although there have been applications of classical extreme value techniques in air quality problems (Singpurwalla 1972, Roberts 1979), these are somewhat restricted in their applications because data series are generally too short for analyses based on annual maxima, which is the classical domain of the theory. A more specific objection to classical extreme value theory, in the context of the present study, is that it would be almost impossible to figure out how to incorporate the meteorological effects which are our main interest in the context of a classical analysis.

An alternative method of extreme value analysis which has been developed extensively in recent years, however, is that based on exceedances over high thresholds. Among the papers developing this approach are Smith (1989), Davison & Smith (1990) and a recent review by Smith (1993). The broad aim of these methods is to find suitable probability distributions for exceedance times and excess values, which can then be estimated by maximum likelihood. A specific selling point of these methods is that they can easily be extended to include covariates through a regression analysis, and indeed the paper of Smith (1989) showed how such an analysis could be used to study trends in the extreme values of urban ozone, though in that paper without making any attempt to account for meteorology.

Davison & Hemphill (1987) analyzed the crossings of a single threshold level, taking account of missing values, through a generalized linear model for binary data. In their analysis, temperature was admitted as a covariate. Shively (1991) adapted a Poisson process viewpoint proposed by Smith (1989) to incorporate a number of meteorological variables, but still only for crossings of a single threshold level. In other words, his analysis accounted for the probability of crossing a given threshold, but did not attempt to model the actual values or excesses of those ozone readings above the threshold. This would be adequate if it were possible to formulate the entire study in terms of exceedances of a single threshold level, but in most cases, to calculate adequate extremal properties, it is necessary to extrapolate to higher thresholds than the one on which the statistical analysis is based, and for that, we need to model the excesses as well.

In the present report, we apply and extend these techniques to a number of data sets extracted from the Chicago ozone study. Specifically, we pick out three stations at which ozone levels are high, and also consider daily maxima across the Chicago network. A detailed description of the data used is in Section 2. Section 3 presents an analysis of the exceedances of a single threshold. In Section 4, this is compared with the model of Bloomfield *et al.*; essentially, our conclusion is that there are difficulties in applying the Bloomfield model to determine probabilities of crossing a high threshold, which do appear to be resolved by the model we are proposing. Section 5 then extends the analysis to include excesses over the threshold. In Sections 6 and 7, we propose some interpretations of these results in the light of the broader scientific aims of the project. More specifically, Section 6 discusses ways of identifying extreme ozone days after making adjustments for meteorology. This information may be useful in establishing control strategies or in suggesting ways that the interpretation of the ozone standard might be modified to account for meteorology. Section 7 presents a methodology for establishing how extreme a particular year was, in

Fig. 8.3: Station R, Sums of excesses over 60



(a) Exp.

(b) GPD

(c) T-Exp.

(d) TGPD

the light of past patterns of meteorology — this is of particular interest with regard to the year 1988, which has seen the highest ozone levels of recent years. Section 8 presents an alternative analysis, based on sums of exceedances over a threshold rather than the daily maximum. Finally, Section 9 summarizes the results of the report.

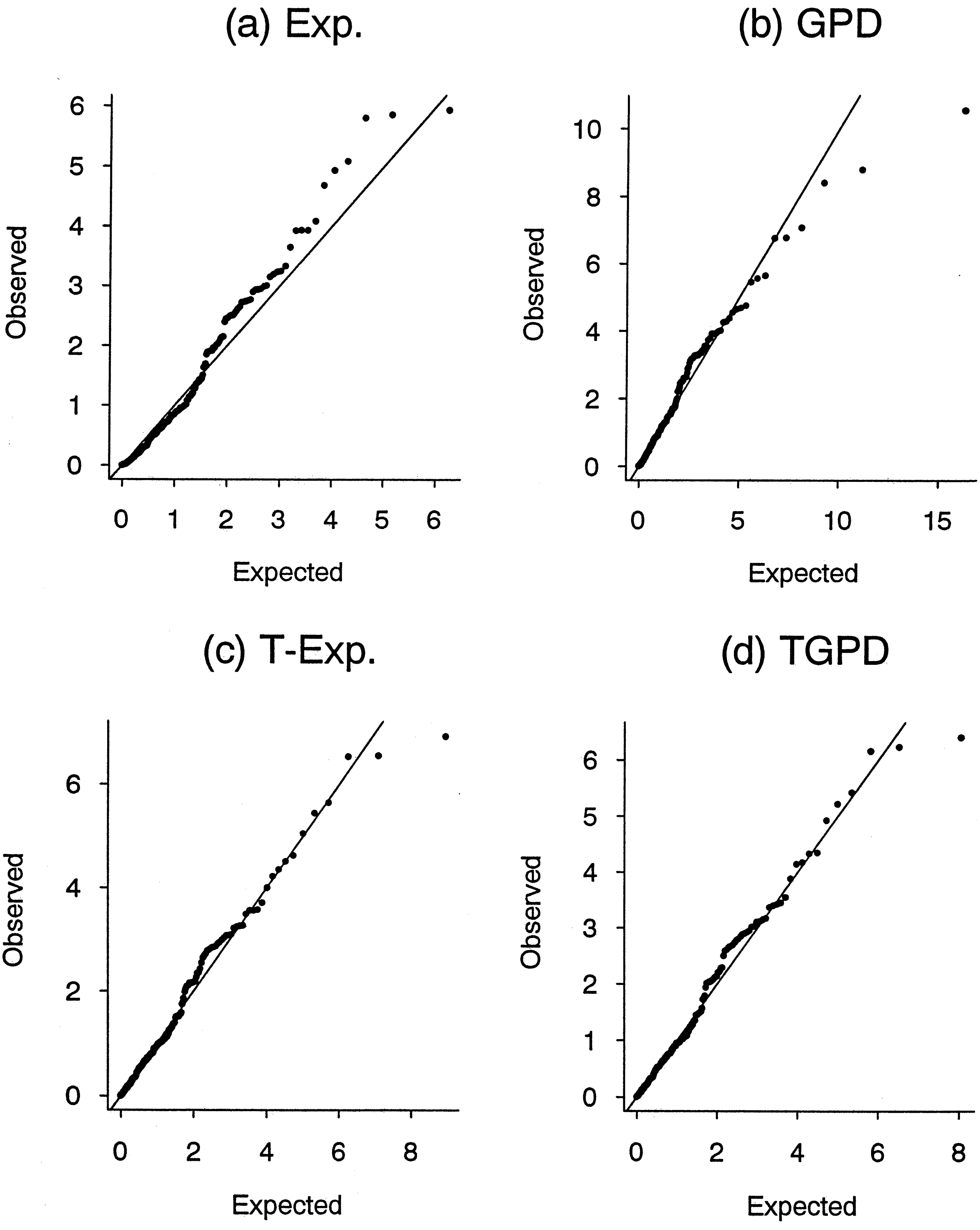
2. The data

The ozone data consist of hourly averages at 45 stations in the Chicago area. Detailed description of the stations and their locations is given by Bloomfield *et al.* (1993). For the present study, it is of particular interest to focus on stations with high ozone levels. Table 2.1 shows the sample size (i.e. total number of complete days available) and the number of exceedances of the daily maximum above the levels 120 and 150 ppb., for each of the 45 stations.

Table 2.1: Numbers of exceedances by station

Station ID	Sample size	Exc. of 120	Exc. of 150
170310001	655	4	0
170310003	446	1	1
170310009	330	1	0
170310032	970	13	2
170310037	1880	8	3
170310038	129	1	0
170310044	825	2	0
170310045	1222	2	0
170310050	2286	5	1
170310053	1062	4	2
170310062	49	0	0
170310063	130	0	0
170310064	467	2	0
170311002	1851	6	1
170311003	1515	15	1
170311601	1351	7	1
170312301	500	1	1
170313005	431	3	0
170314002	1595	12	1
170314003	1285	5	0
170315001	282	0	0
170316002	131	1	0
170317002	3448	41	11
170318003	164	0	0
170431002	765	1	1
170436001	2288	6	0
170890005	3284	5	0
170970001	3180	15	6
170971002	3269	38	9
170971003	275	0	0
170973001	3345	17	4
171110001	3374	6	2
171971007	1770	3	0
171971008	3192	7	1
180890011	190	0	0
180891016	1435	8	2
180892001	235	1	0
180892002	258	0	0
180892008	1056	11	2
181270020	686	7	2
181270021	596	7	1
181270024	1053	23	5
181270903	31	0	0
181271004	422	8	3
181271005	266	0	0

Fig. 8.2: Station P, Sums of excesses over 80



(a) Exp.

(b) GPD

(c) T-Exp.

(d) TGPD

Also computed were exceedances of the level 180 ppb., but only four stations had any exceedances here: 170971002, 180892008 and 181270024 had one each, and 170317002 had three. Based on these results, and taking sample sizes into account, three stations were selected for detailed analysis: 170317002, 170971002 and 181270024. In the rest of the report these are referred to as stations P, Q and R, respectively. Figure 2.1 (based on Figure 1 of Bloomfield *et. al*, 1993) shows these stations.

In addition to the daily maxima from these three stations, we have analyzed daily maxima for the whole network. Network maxima are not easy to define, because different stations have been in operation for different periods of time, and there are many missing values. However, Bloomfield *et. al* (1993, Section 4.2) applied median polish kriging to interpolate missing values within a subset of 16 stations and subsequently (Section 6.6) used the same data to construct network maxima based on those 16 stations. We use the same constructed data set here.

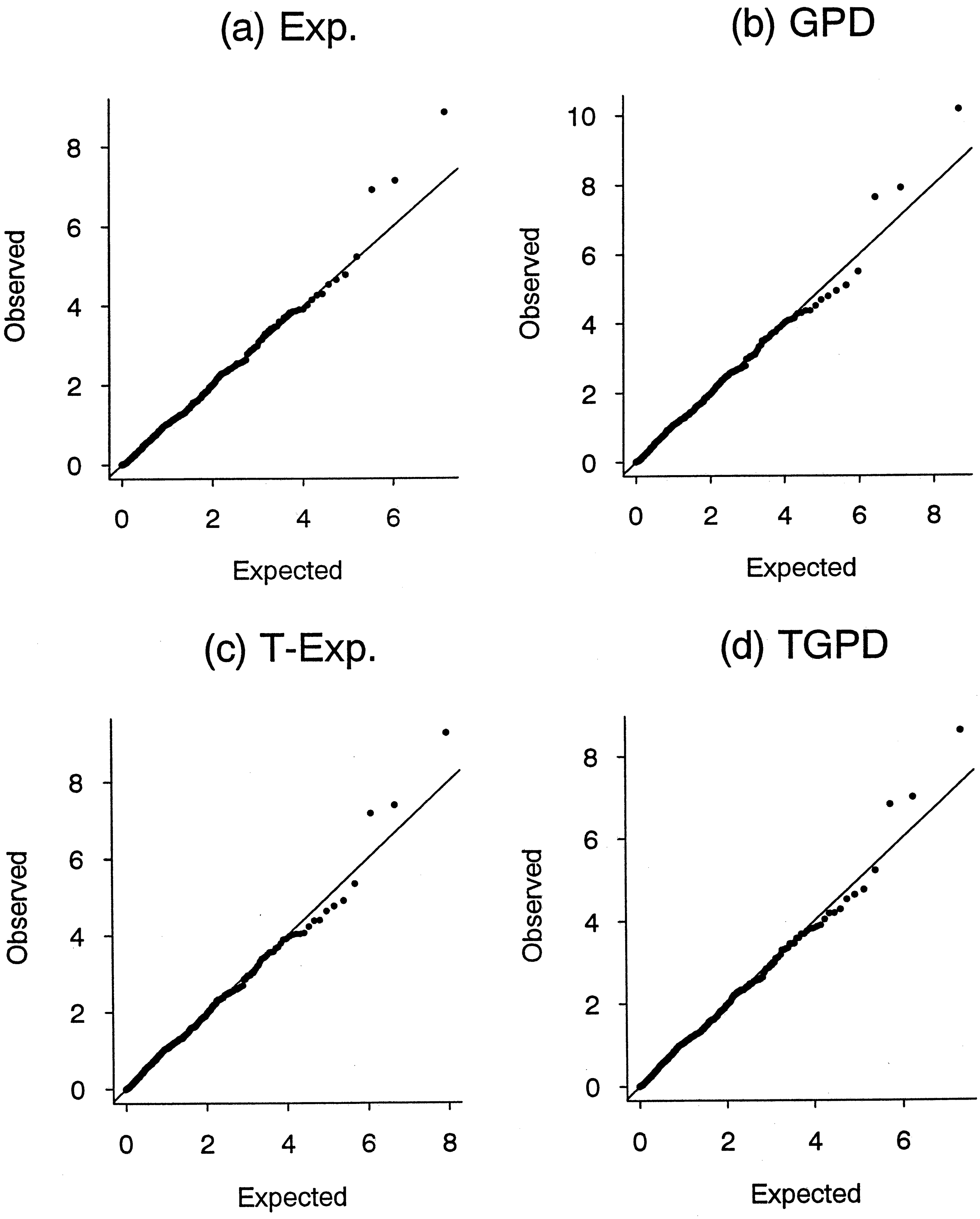
The meteorological data consist of surface weather data at a single station in Chicago. Data are available on an hourly basis, but for the purpose of the present study, only the noon values are considered. Bloomfield *et al.* (1993) consider extending the analysis to incorporate lagged data (i.e. using previous days' meteorology, as well as that of the current day, as covariates in predicting the ozone level of a specific day) and to use upper air data, but these have not been brought into the present analysis.

The measured meteorological variables are given in Table 2.2. Amongst these, TOTCOV and CHT were omitted from the analysis, TOTCOV because it is generally considered that OPCOV is a more relevant measure of cloud cover for the ozone problem, and CHT because it is difficult to interpret (a missing value or 0 actually meaning that the ceiling is too high to be measured). Both omissions are supported by statements in Bloomfield *et al.* (1993).

Table 2.2: Measured meteorological variables

Variable name	Description
TOTCOV	Total cloud cover (%)
OPCOV	Opaque cloud cover (%)
CHT	Ceiling height (m.)
PR	Barometric pressure (mb.)
T	Temperature (^{o}F)
TD	Dewpoint Temperature (^{o}F)
RH	Relative humidity (%)
Q	Specific humidity (g./kg.)
VIS	Visibility (km.)
WSPD	Wind speed (m./sec.)
WDIR	Wind direction (^{o} from North)

Fig. 8.1: Station P, Sums of excesses over 60



In addition, several additional variables were created from the above, and are described in Table 2.3. These are mainly motivated by the results in Bloomfield *et al.* (1993), who factored windspeed into directional components, and also used up to cubic terms in temperature (the divisors 10 and 1000 being introduced into T2 and T3 to improve numerical stability of the coefficients). Finally, the inclusion of the T.WSPD variable is motived by the conclusion of Bloomfield *et al.* (1993) that there is a strong temperature-windspeed interaction, though in their case the actual model is highly nonlinear (see Section 4) and it is open to question whether a single cross-product term, as here, will capture that effect.

Table 2.3: Additional variables created from the data

WIND.U	$-\text{WSPD}\times\sin(2\times\pi\times\text{WDIR}/360)$
WIND.V	$-\text{WSPD}\times\cos(2\times\pi\times\text{WDIR}/360)$
T2	$(\text{T-60})^2/10$
T3	$(\text{T-60})^3/1000$
T.WSPD	$\text{WSPD}\times(\text{T-60})$

In addition to the meteorological covariates, YEAR was used as a covariate (=1 for 1981, through to 11 for 1991), and two variables CDAY and SDAY, defined by

$$\text{CDAY} = \cos(2\times\pi\times\text{DAY}/365.25), \quad \text{SDAY} = \sin(2\times\pi\times\text{DAY}/365.25),$$

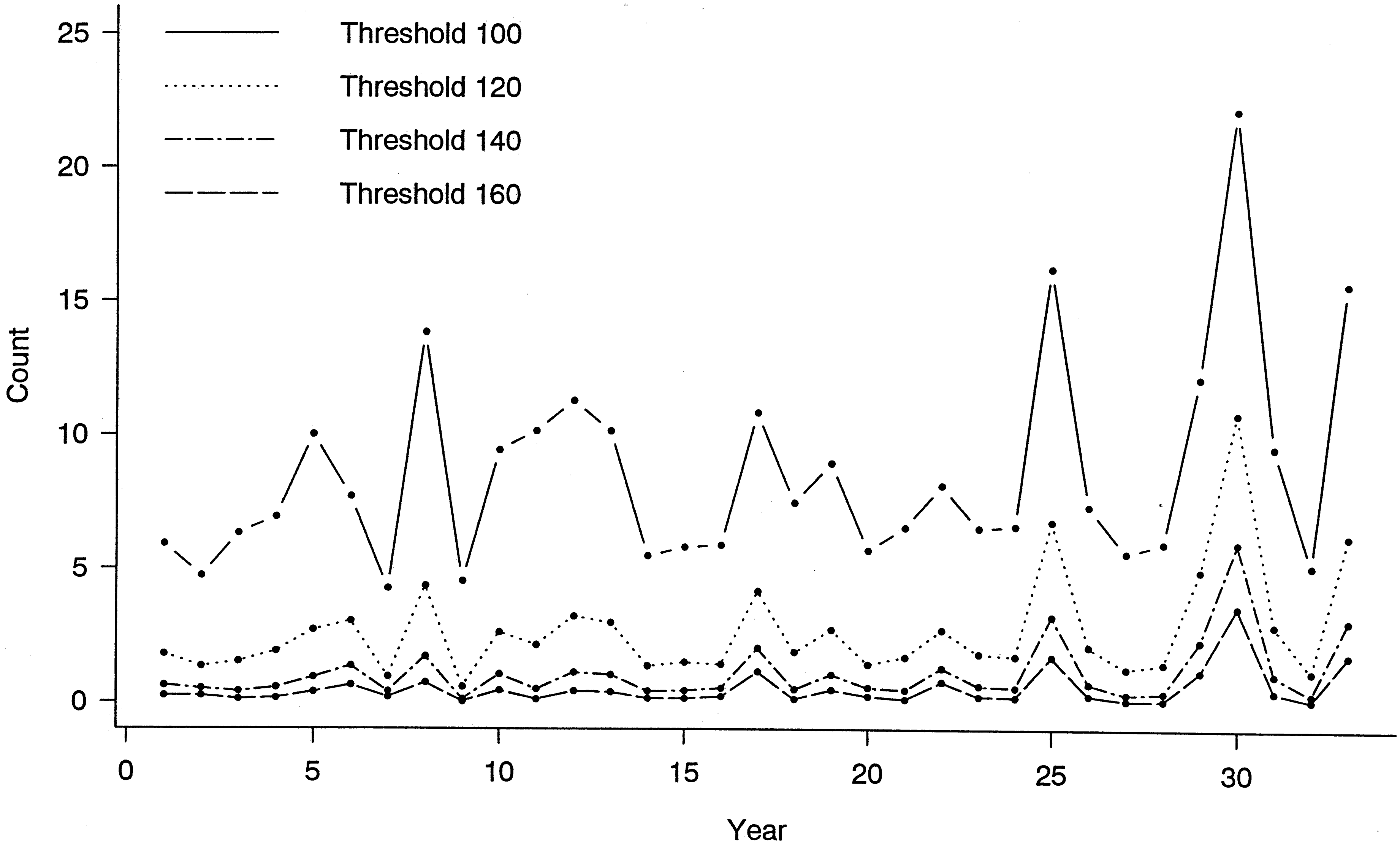
where DAY represents the day within the year (=1 for January 1, etc.). The inclusion of CDAY and SDAY is intended to reflect the fact, also discussed by Bloomfield *et al.* (1993), that there remains a residual seasonal effect even when the direct influence of season on meteorology has been taken into account.

3. Exceedances of a single threshold

Suppose now we are concerned to model the probability, as a function of the covariates, that the ozone on a specified day exceeds a specified threshold. For most of the discussion that follows, the threshold will be taken as the ozone standard 120 ppb, though the methodology applies to any high threshold and, especially in conjunction with the analysis of exceesses over the threshold (Section 5), there can be advantages in adopting a lower threshold so as to capture more data.

For the initial analysis we shall assume that separate days are independent. Such an assumption is not unreasonable, because it is widely believed that persistence of high-ozone conditions has more to do with persistence of meteorology than with the ozone itself. Nevertheless, the independence assumption is not exactly satisfied, and further on we shall describe ways to get around it.

Fig. 7.2: Expected exceedances 1959-1991
Model fitted to Station P



Under this independence assumption, the likelihood for the data may be defined by

$$L = \prod_i p_i^{\delta_i}(1-p_i)^{1-\delta_i}, \tag{3.1}$$

where p_i denotes the probability that the threshold is exceeded on day i, and δ_i is an indicator of whether the threshold is in fact exceeded on day i ($\delta_i = 1$ if the threshold is exceeded, 0 otherwise).

As a model for p_i, we assume the familiar logit model:

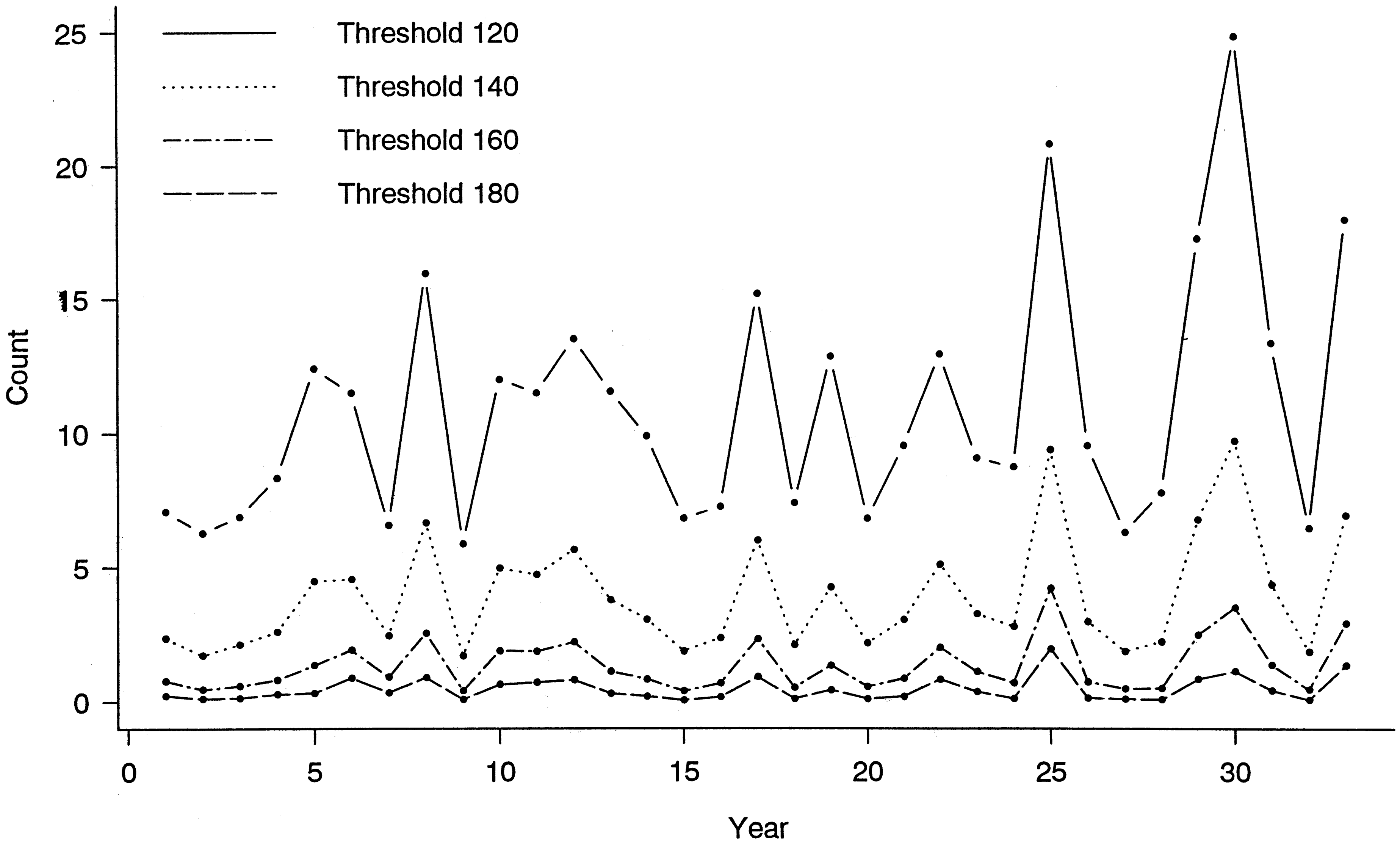
$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_j x_{ij}\beta_j, \tag{3.2}$$

where x_{ij} is the value of the j'th covariate on day i and β_j is the corresponding coefficient. We always assume $x_{i1} = 1$, i.e. there is always a constant term in the model, but the other covariates are selected from those described in Section 2.

For a given set of covariates, estimation proceeds by maximum likelihood, i.e. choose $\beta_1, \beta_2, \ldots$ to maximize L as defined by (3.1) and (3.2). In most of the examples in this paper, this has been achieved using Fortran programs employing the DFPMIN (Davidon-Fletcher-Powell variable metric function minimization) algorithm of Press *et al.* (1986), Chapter 10, to minimize the objective function $-\log L$. The algorithm was converted to double precision with EPS reset to 10^{-14}, and for most runs the convergence parameter FTOL was taken as 10^{-8} (the algorithm is taken to converge as soon as two consecutive function evaluations are within FTOL). The algorithm requires first-order derivatives of the objective function, as well as the function itself, but these have been adequately approximated using simple first-order differences ($\partial f/\partial x \approx \{f(x+\delta) - f(x)\}/\delta$, where we have taken $\delta = 10^{-6}$). The other modification of the published DFPMIN algorithm was to retain the HESSIN matrix as an argument of the function. This matrix contains an approximation to the inverse of the Hessian matrix of $-\log L$. In maximum likelihood theory this is the observed information matrix, used as an approximation to the variance-covariance matrix of the parameter estimates $\{\widehat{\beta}_j,\ j = 1, 2, \ldots\}$. In particular, the square roots of its diagonal entries define approximate standard errors for the parameters. However, it should be remembered that these standard errors are only an approximation and that alternative methods of obtaining standard errors, for instance via jackknife or bootstrap procedures, may well give better results. The procedure requires specification of starting values, but it does not appear to be sensitive to these and even starting from $\beta_j = 0$ for all j produced convergence in all cases tried, though sometimes rather slowly (50+ iterations). The only other programming point that requires care is to guard against exponential overflow; a trap was built into the function minimization routine to test for this, the objective function being set equal to a very large number (10^{10}) if the test failed.

For this specific model, specialized programs are not in fact required as it is a well-known model for which a number of packaged routines exist, for instance the `glm` command

Fig. 7.1: Expected exceedances 1959-1991
Model fitted to network maxima



to fit generalized linear models in Splus, or the LOGISTIC procedure of SAS. Shreffler (1993) successfully used the latter to obtain a very similar analysis of data from St. Louis, Missouri. However, the more complicated models to be discussed in Section 5 do not fit within the generalized linear models framework, and so do require specialized programming.

Stepwise selection of variables

Although the maximization of the likelihood function is mechanical, given the covariates, the process of deciding which covariates to include is not. The standard method of doing this is a stepwise selection, in which the variables are introduced one at a time, the minimized values of the negative log likelihood $-\log L$, hereafter referred to as NLLH, being used to decide which variables to introduce and when to stop. This is sometimes combined with, or replaced by, a backward selection procedure, in which initially too many variables are included, and the least significant ones are then dropped until all the remaining variables are significant. As a test of significance, one widely used measure is the t ratio defined as the value of the parameter divided by its standard error. As a rough guide, any variable with a t ratio greater than 2 in absolute value is considered significant. However, as mentioned above, the standard errors obtained from the function minimization routine are only approximate, and so should not be used as the sole criterion for including or excluding any variable. As an alternative to t ratios, NLLH may be used directly. Here, a suitable rule of thumb is that a variable is significant if its inclusion in the model reduces the NLLH by more than 2. This is based on the approximate χ_1^2 distribution for the deviance statistic (twice the difference of NLLHs for the two models being compared). If more than one variable is introduced at a time, then the χ_1^2 distribution is replaced by χ_q^2, where q is the number of new variables introduced.

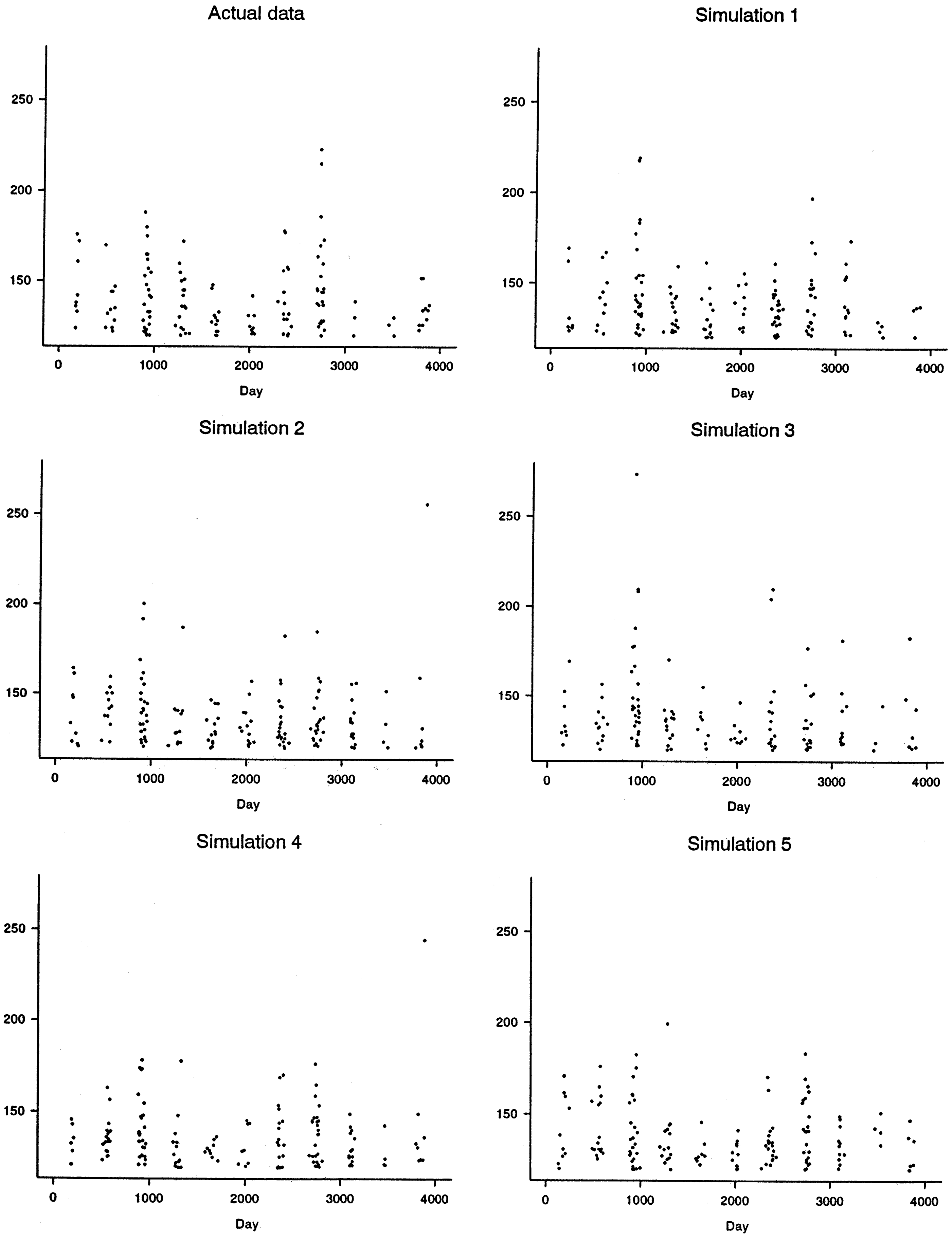
As an indication of how these ideas are applied, we now go in detail through the selection of variables for Station P, threshold 119.9 (the threshold was set slightly below 120 to ensure that a daily maximum of exactly 120 would be counted as an exceedance). In other cases considered in this paper, a similar procedure was followed, though in most cases, not in so much detail.

One other point to be noted is that there are certain combinations of variables which do not make sense. For instance, it would not make sense to include either of CDAY or SDAY without the other, since the relation between them depends on the quite arbitrary decision to define DAY with respect to January 1, rather than any other fixed date of the year. Similarly, WIND.U and WIND.V go together, but we do not include both them and WSPD, since the latter is determined completely (albeit nonlinearly) by the former. Also, we do not include the T.WSPD interaction term unless one of WSPD or the WIND.U/WIND.V pair is present.

With these considerations in mind, the detailed analysis is as follows.

Step 1. Since YEAR can be considered as a trend variable, an initial analysis is made using just YEAR as a covariate. This results in NLLH=221.740.

Fig. 5.12: Simulation comparisons for network maxima



Actual data

Simulation 1

Simulation 2

Simulation 3

Simulation 4

Simulation 5

Step 2. Various other variables were added to the model, in each case calculating NLLH. For example, when CDAY and SDAY were added together, NLLH dropped to 168.680. Other tries were OPCOV (NLLH=210.587), PR (221.722), T (136.369), TD (186.833), RH (202.383), Q (191.617), WSPD (206.610), WIND.U+WIND.V (220.636), VIS (221.481). T is clearly the most significant, so it is added to the model.

The model therefore now contains variables YEAR, T and has NLLH=136.369.

Step 3. New variables added to model: CDAY+SDAY (133.455), OPCOV(135.040), PR (134.204), TD (132.040), RH (132.045), Q (130.936), WIND.U+WIND.V+T.WSPD (123.821), WSPD+T.WSPD (123.369), VIS (136.250). The combination of variables WSPD+T.WSPD gives the biggest reduction in NLLH and so is selected. The model now contains variables YEAR, T, WSPD, T.WSPD and has NLLH=123.369.

Step 4. New variables: CDAY+SDAY (121.061), PR (122.925), OPCOV (122.076), TD (118.505), RH (118.438), Q(117.019), VIS (122.861). Q selected. The model now contains variables YEAR, T, Q, WSPD, T.WSPD and has NLLH=117.019.

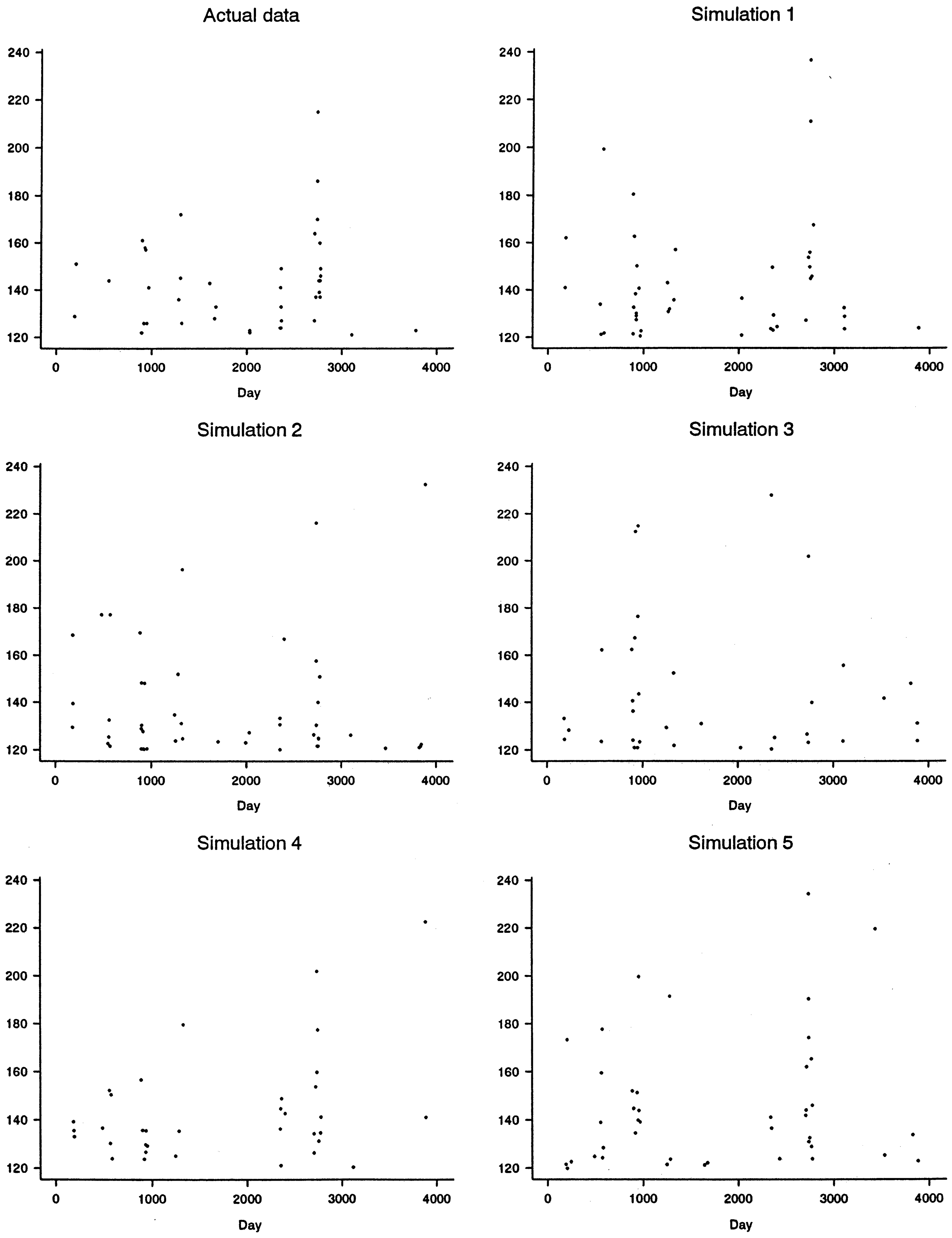
Step 5. New variables: CDAY+SDAY (114.553), PR (117.018), OPCOV (116.936), TD (115.536), RH (115.514), VIS (116.972). CDAY+SDAY selected. The model now contains variables YEAR, CDAY, SDAY, T, Q, WSPD, T.WSPD and has NLLH=114.553.

Step 6. Once again PR (114.553), OPCOV (114.323), TD (113.382), RH (113.545) VIS (114.485) were added to the model but it is clear that none of these is significant, so they were not included.

Step 7. The model is now complete in terms of the original variables tried, but there are still some additional models to be considered. All the current variables appear to be significant — the only one with a t ratio less than 2 in magnitude was SDAY (−0.9), but as discussed above it would not make sense to drop SDAY and retain CDAY, while the latter has a t ratio of −2.04, which clearly indicates it should be retained. However, we still have not used T2 and T3. In fact, including T2 reduces the NLLH to 114.215 while adding T3 as well creates no further reduction. Thus we do not include T2 and T3.

Step 8. So far, we are still modeling the days as independent. One way to improve on that assumption is to include an additional variable PDAY, defined to be 1 if the previous day is an exceedance (of the same threshold) and 0 otherwise. If the previous day is missing, this variable is simply omitted. Although this can be incorporated with minimal change in the software used to fit the model, the actual effect is to turn the model from independent observations into a first-order Markov chain — of which, more will be said in Section 5. In this case, adding PDAY reduces NLLH to 112.729, indicating that this variable is borderline significant. However, adding another variable PDAY2 (=1 if the previous day but one was an exceedance, otherwise 0) does not have any further effect (NLLH=112.618), indicating that the first-order Markov chain is adequate.

Fig. 5.11: Simulation comparisons for Station P



Actual data

Simulation 1

Simulation 2

Simulation 3

Simulation 4

Simulation 5

Step 9. A final step, considering that a major purpose of this whole analysis is to determine the significance of the YEAR variable in the presence of the meteorological factors, is to repeat the most important analyses without YEAR. For the model of Step 7, if YEAR is omitted, NLLH rises from 114.553 to 117.503. For the model of Step 8, it goes from 112.729 to 115.451. Either way, YEAR is still significant.

The two principal models that have been identified (with and without PDAY) are summarized in Tables 3.1 and 3.2.

Table 3.1: Station P, threshold 119.9 (NLLH=114.553)

Variable	Estimate	Stand. error	*t* Ratio
CONST	−38.58	4.837	−7.977
YEAR	−.1574	.06184	−2.545
CDAY	−2.995	1.471	−2.036
SDAY	−.4992	.5331	−.9364
T	.4476	.05628	7.954
Q	−.2166	.06017	−3.599
WSPD	.4800	.1781	2.695
T.WSPD	−.03308	.006211	−5.326

Table 3.2: Station P including PDAY, threshold 119.9 (NLLH=112.729)

Variable	Estimate	Stand. error	*t* Ratio
CONST	−36.66	5.142	−7.13
YEAR	−.1548	.06739	−2.297
CDAY	−2.857	1.557	−1.835
SDAY	−.5030	.6109	−.8233
PDAY	1.044	.5261	1.984
T	.4248	.05992	7.089
Q	−.215	.06278	−3.424
WSPD	.4402	.1872	2.351
T.WSPD	−.03144	.006639	−4.735

Similar analysis is applied to Station Q and Station R. The resulting models are the following. Note that T2 is included among the variables for station Q. In neither of these cases was PDAY significant.

Fig. 5.10: Network maxima exceedances, dependent model

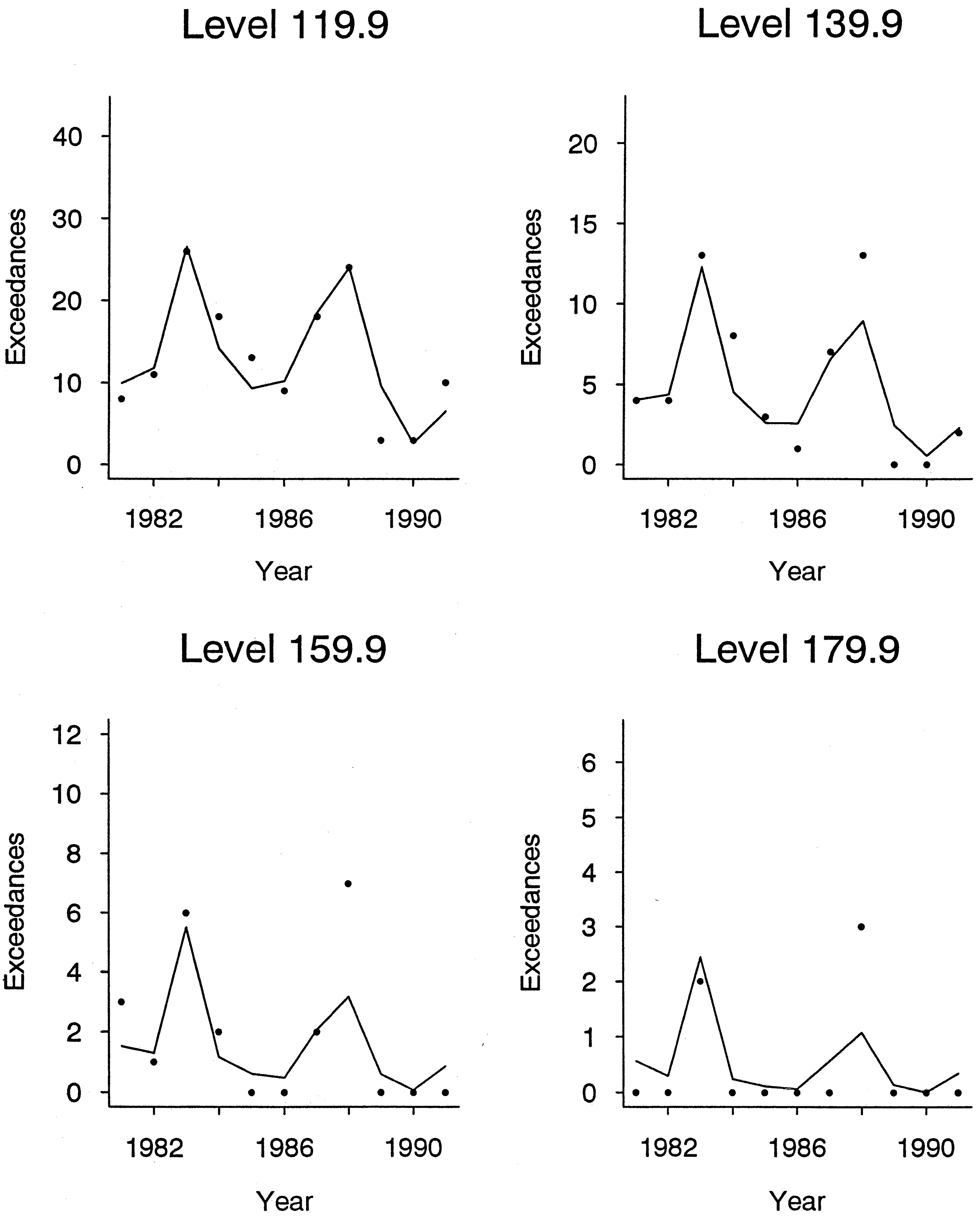


Table 3.3: Station Q, threshold 119.9 (NLLH=99.317)

Variable	Estimate	Stand. error	*t* Ratio
CONST	−171.2	46.05	−3.72
YEAR	−0.09339	0.0699	−1.34
T	2.256	0.6182	3.65
WIND.U	0.02198	0.09156	0.24
WIND.V	0.05041	0.09536	0.53
T2	−0.3335	0.1056	−3.16
T.WSPD	−0.02139	0.005378	−3.98

Table 3.4: Station R, threshold 119.9 (NLLH=67.524)

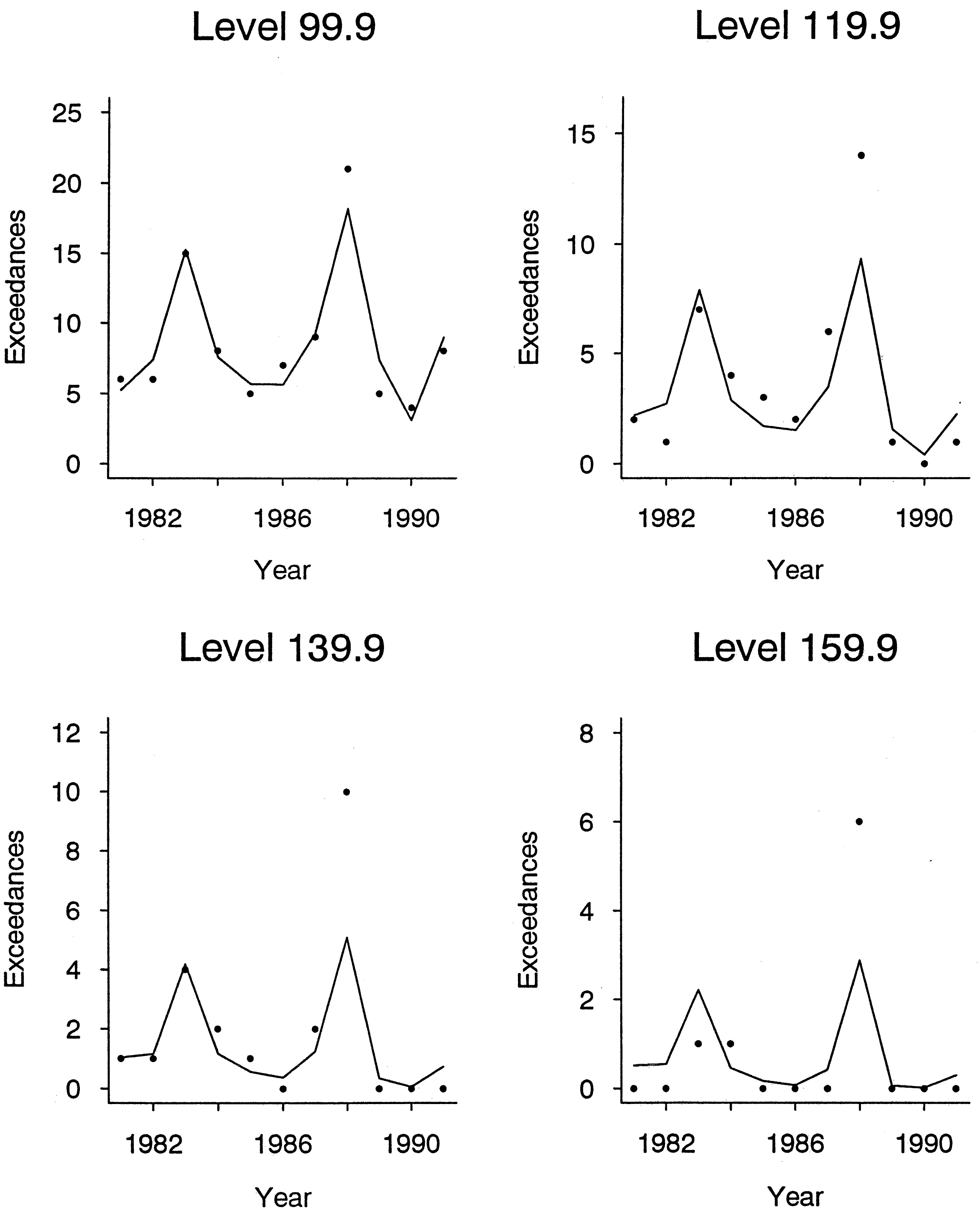
Variable	Estimate	Stand. error	*t* Ratio
CONST	−25.76	3.637	−7.083
YEAR	−.3851	.1614	−2.385
T	.2846	.04539	6.270
PR	.1660	.06312	2.623
WIND.U	.2024	.09925	2.040
WIND.V	−.1822	.08821	−2.066
T.WSPD	−.01382	.006699	−2.063

If the variable YEAR is omitted, then NLLH rises to 100.359 in the case of station Q and 70.166 for station R. This is still significant in the case of station R, but for station Q, whether judged from the *t* ratio or by the increase in NLLH, the trend does not appear to be significant.

One possible disadvantage of this analysis is that the actual number of exceedances (41, 38 and 23 respectively for stations P, Q and R), although in excess of the standard, is still rather small for making a definitive assessment of the trend in the light of so many covariates. We therefore consider also an analysis based on threshold 100, for which there are far more exceedances (94, 84 and 55). As we shall see in Section 5, it is still possible to construct estimates of the probability of exceeding the level 120 (or any higher level) by combining these results with models for the excesses over a threshold.

Tables 3.5, 3.6 and 3.7 show the results of the best model, after repeating the entire process of variable selection, for each of the three stations.

Fig. 5.9: Station P exceedances, dependent model



Level 99.9

Level 119.9

Level 139.9

Level 159.9

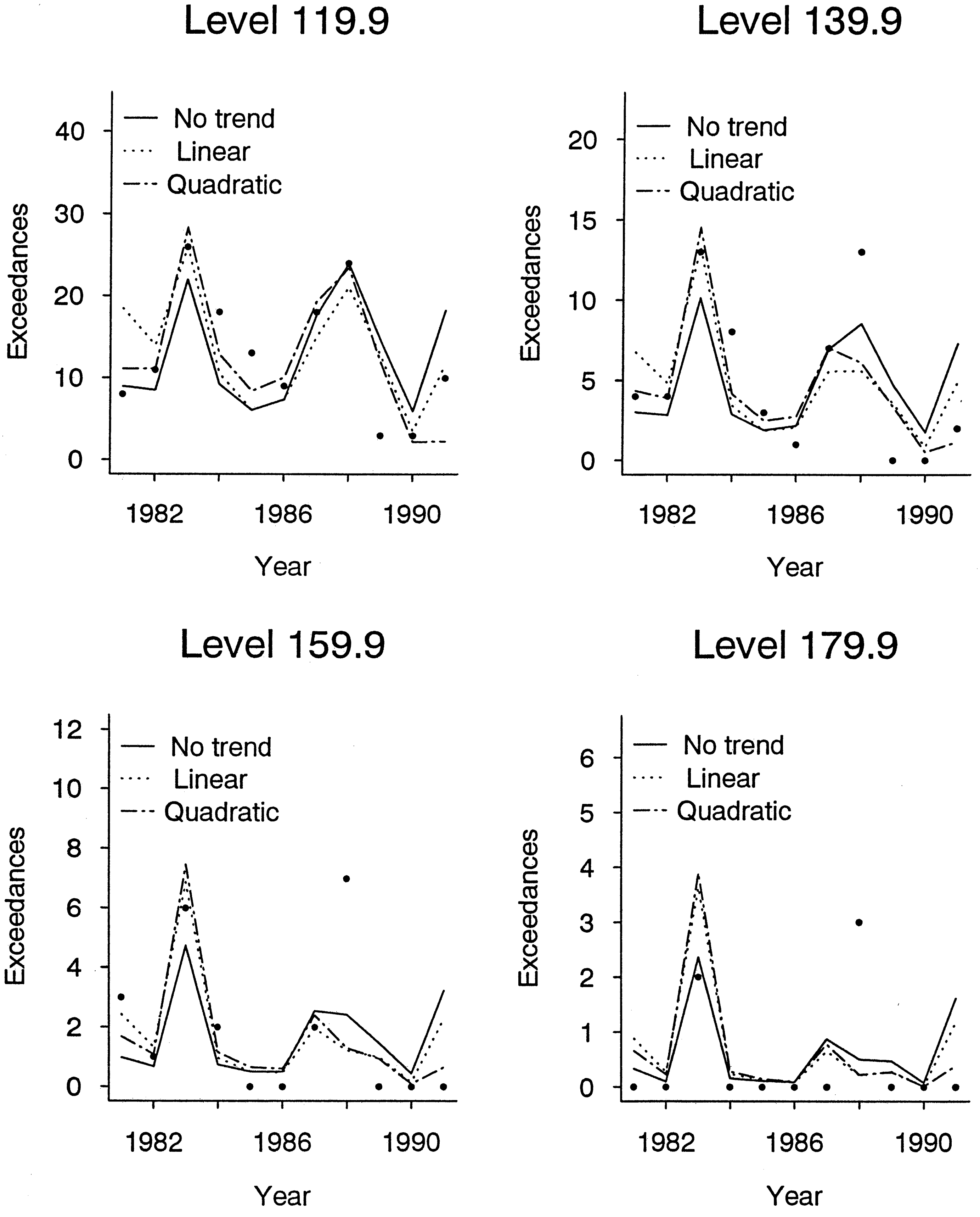
Table 3.5: Station P, threshold 99.9 (NLLH=220.315)

Variable	Estimate	Stand. error	*t* Ratio
CONST	−28.99	3.684	−7.87
YEAR	−0.08892	0.0432	−2.06
CDAY	−1.572	0.719	−2.19
SDAY	−0.2333	0.317	−0.74
PDAY	1.192	0.3201	3.72
T	0.3058	0.02899	10.55
TD	0.1805	0.09277	1.95
Q	−0.6315	0.2178	−2.90
WIND.U	0.06596	0.04931	1.34
WIND.V	0.07411	0.05648	1.31
VIS	−0.05836	0.02333	−2.50
T.WSPD	−0.02171	0.003469	−6.26

Table 3.6: Station Q, threshold 99.9 (NLLH=205.190)

Variable	Estimate	Stand. error	*t* Ratio
CONST	−34.84	4.917	−7.08
YEAR	−0.01963	0.04365	−0.45
CDAY	−1.717	0.8759	−1.96
SDAY	−0.4778	0.3835	−1.25
PDAY	1.178	0.3648	3.23
T	0.3056	0.0327	9.35
TD	0.2692	0.1179	2.28
Q	−0.7543	0.2675	−2.82
WIND.U	0.06855	0.05627	1.22
WIND.V	0.1563	0.063	2.48
T.WSPD	−0.02176	0.003745	−5.81

Fig. 5.8: Network maxima, exceedances over several levels



Level 119.9

Level 139.9

Level 159.9

Level 179.9

Table 3.7: Station R, threshold 99.9 (NLLH=106.066)

Variable	Estimate	Stand. error	*t* Ratio
CONST	−68.16	13.8	−4.94
YEAR	−0.184	0.1278	−1.44
T	0.8645	0.1639	5.27
RH	0.2951	0.07919	3.73
Q	−1.344	0.3067	−4.38
WIND.U	0.08165	0.07059	1.16
WIND.V	−0.1845	0.07138	−2.58
VIS	−0.09188	0.03728	−2.46
T.WSPD	−0.01191	0.00487	−2.45

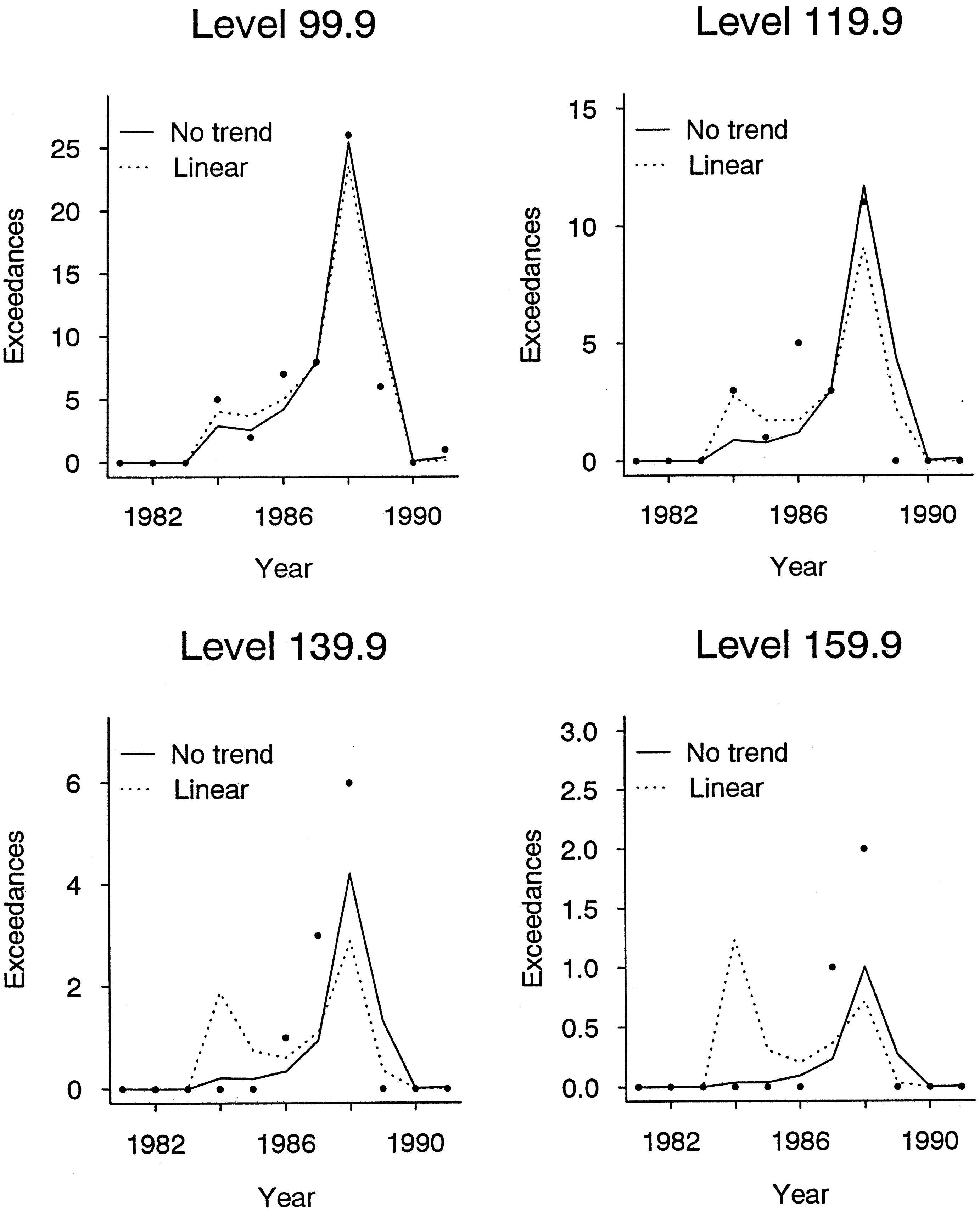
If the variable YEAR is dropped, then the NLLH rises to 222.289, 205.289 and 107.202 respectively for stations P, Q and R. This is still significant in the case of P (just!), but still not significant for Q and also no longer significant for R. There is thus some suggestion that it is harder to establish significant trends when the threshold is lowered. We shall return to this point later (Section 5).

One more issue is whether a linear trend is adequate — for instance, should we be using a quadratic trend? This can be tested by adding a new variable YEAR2, defined as the square of YEAR, and repeating the above analysis. In only one of the cases considered so far was this found to be significant, namely for station P at threshold 119.9, but in view of its importance for later discussion, we give the full result here:

Table 3.8: Station P including YEAR2, threshold 119.9 (NLLH=110.410)

Variable	Estimate	Stand. error	*t* Ratio
CONST	−41.45	5.274	−7.86
YEAR	.7496	.3505	2.14
YEAR2	−.0782	.02997	−2.61
CDAY	−3.07	1.576	−1.95
SDAY	−.4892	.6051	−.81
T	.461	.05928	7.78
Q	−.2318	.06393	−3.63
WSPD	.468	.1859	2.52
T.WSPD	−.03463	.006423	−5.39

Fig. 5.7: Station R exceedances over several levels



Comparison with a Poisson process formulation

Shively (1991) approached exactly the same problem, but using slightly different meteorological variables and data from Houston instead of Chicago, from a Poisson process viewpoint. This was apparently motivated by the procedure in Smith (1989), which itself was motivated by the fact that much of the modern probabilistic theory of extremes (Leadbetter *et al.* 1983, Resnick 1987) uses point process formulations as the most natural way of looking at the exceedances of a high level. Indeed, this approach unifies the different approaches towards the limit theorems of extreme value theory.

A *nonhomogeneous Poisson process* on $(0, \infty)$ with *intensity function* $\lambda(t)$, $t > 0$, is defined by the property that the probability of an "event" (in the present context, an exceedance of the threshold) in a short time interval $(t, t + \delta t)$ is of the form $\lambda(t)\delta t + o(\delta t)$, the probability of more than one such event being $o(\delta t)$, and such events are independent over different time intervals. For such a process, the number of events observed in any finite time interval, (t_1, t_2) say, has a Poisson distribution with mean

$$\int_{t_1}^{t_2} \lambda(t) dt.$$

The intuitive reason why we might expect such a process to be relevant to high-level exceedances of ozone stems from the fact that exceedances are "rare events" (at least, they are if the threshold is high enough), and if we also assume that the independence assumption is at least approximately satisfied, then a nonhomogenous Poisson process is the natural model which results.

Suppose now we have such a process with intensity function $\lambda(t; \beta)$ depending on an unknown parameter vector β. Suppose we observe the process on an interval $(0, T)$ and in that time observe N events (a random number) at times $T_1, ..., T_N$. The likelihood function to be maximized is given by

$$L = \prod_{i=1}^{N} \lambda(T_i; \beta) \cdot \exp\left\{ - \int_0^T \lambda(t; \beta) dt \right\} \tag{3.3}$$

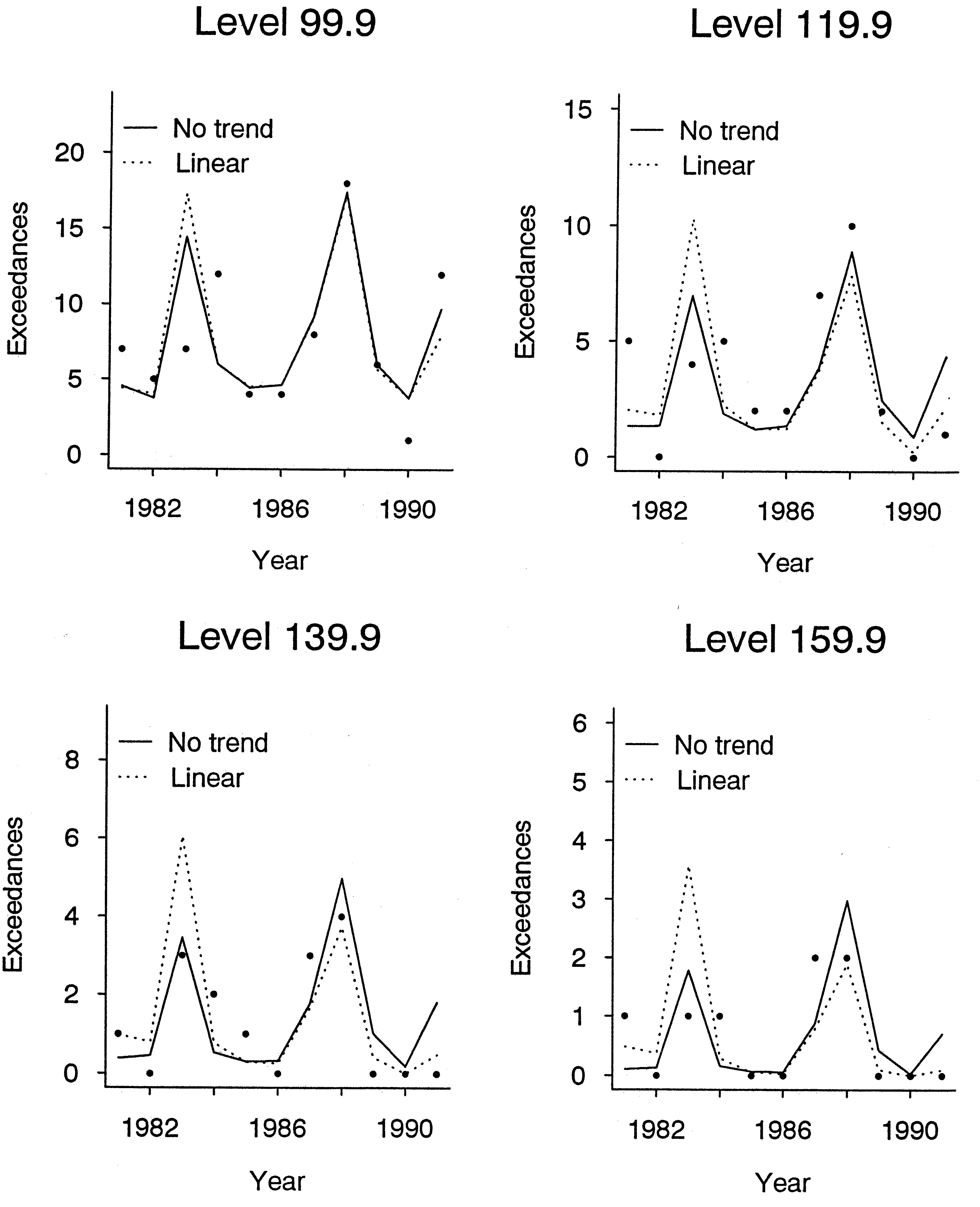
For a careful but non-measure-theoretic derivation of this, see Section 3.3 of Cox & Lewis (1966). A modern and mathematically rigorous treatment of inference from point process is the book by Karr (1986).

Shively (1991) modeled the dependence on the covariates through a logarithmic link function of the form

$$\log \lambda(t; \beta) = \sum_j x_{tj} \beta_j, \tag{3.4}$$

where x_{tj} is the value of the j'th covariate at time t. Using a discrete approximation to the integral in (3.3), he was then able to define a likelihood function for the parameters β_j, $j = 1, 2, ...$, and so to obtain maximum likelihood estimates.

Fig. 5.6: Station Q exceedances over several levels



Level 99.9

Level 119.9

Level 139.9

Level 159.9

An alternative viewpoint on this approach, however, is to note that the nonhomogeneous Poisson process, with each of the covariates being defined on a daily basis, is equivalent to assuming that the number of exceedance on day i, for each i within the observed period, has a Poisson distribution with mean λ_i, which we may write, imitating (3.4), in the form

$$\log \lambda_i = \sum_j x_{ij}\beta_j. \tag{3.5}$$

If δ_i denotes the observed number of exceedances on day i, then the likelihood is proportional to

$$L = \prod_i \lambda_i^{\delta_i} e^{-\lambda_i}. \tag{3.6}$$

Of course, in reality δ_i will only be 0 or 1, since we are applying these ideas to the daily maxima.

Comparing the likelihood defined by equations (3.1) and (3.2) with that defined by (3.5) and (3.6), the difference is seen to be the following: defining $A_i = \sum_j x_{ij}\beta_j$, the component of the likelihood arising from day i in the binomial model is

$$\frac{e^{A_i\delta_i}}{1+e^{A_i}},$$

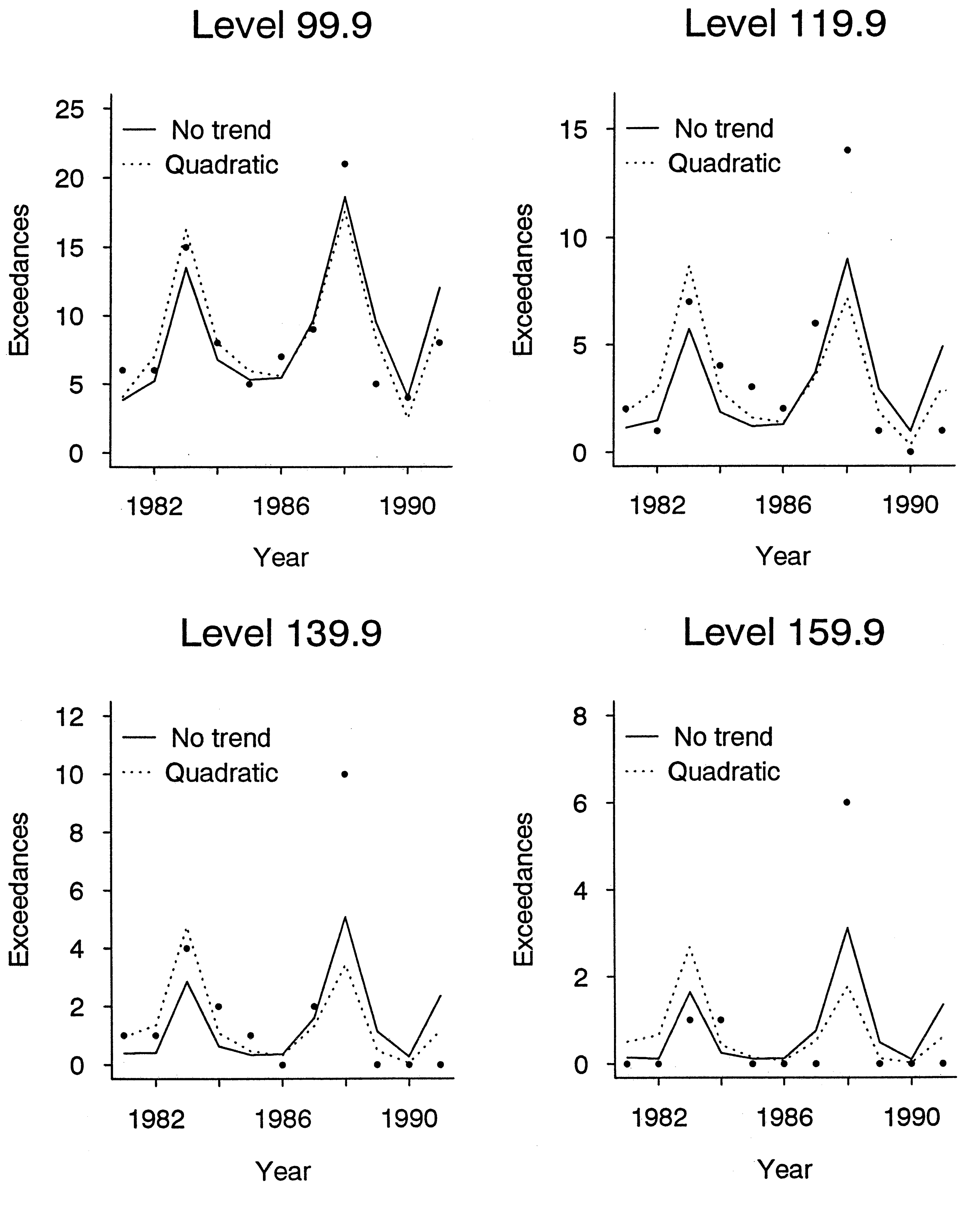
whereas that arising in the Poisson model is

$$\exp(A_i\delta_i - e^{A_i}).$$

Thus, going from the binomial model to the Poisson model amounts to replacing $1+e^{A_i}$ by $\exp(e^{A_i})$. This will make little difference if p_i or λ_i are uniformly small, i.e. if we are genuinely in the situation where we might expect the Poisson and binomial distributions to be indistinguishable, but otherwise the two models could produce substantially different results.

To make a comparison between the two kinds of model, each of the models in Tables 3.1-3.7 was refitted, both with and without the YEAR variable, under this alternative Poisson assumption. The results are summarized in Table 3.9. Note that the model selection process was not repeated, though it is possible that the optimum choice of variables under the Poisson model will be different from that under the binomial model. The results do not change any of our broad conclusions about the significance of YEAR, but quite a few of the individual parameter estimates and their levels of significance were changed under the new model. Table 3.10 gives one instance of this, in which the complete results for the model of Table 3.2 are repeated under the Poisson assumption.

Fig. 5.5: Station P exceedances over several levels



Level 99.9

Level 119.9

Level 139.9

Level 159.9

Table 3.9: Neg. log. likelihoods under Poisson model

Model of Table	NLLH with YEAR	NLLH without YEAR
3.1	122.979	125.788
3.2	122.487	125.181
3.3	105.784	106.446
3.4	71.571	74.118
3.5	248.234	249.952
3.6	226.114	226.276
3.7	133.627	134.319

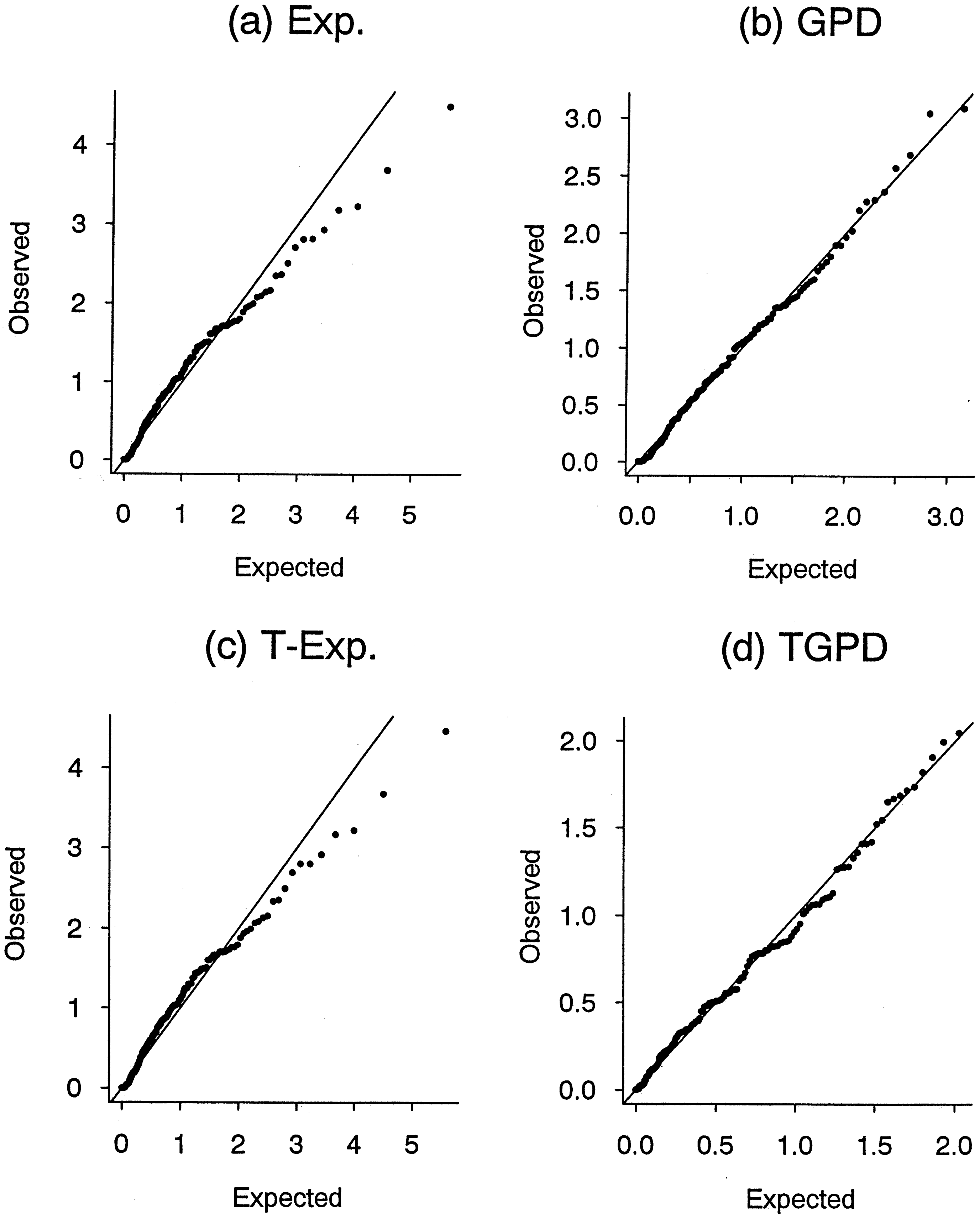
Table 3.10: Model of Table 3.2, under Poisson assumption

Variable	Estimate	Stand. error	t Ratio
CONST	−28.11	5.443	−5.17
YEAR	−0.1405	0.06103	−2.30
CDAY	−2.241	1.311	−1.71
SDAY	−0.4288	0.5317	−0.81
PDAY	0.4277	0.4212	1.01
T	0.3157	0.06302	5.01
Q	−0.1535	0.05165	−2.97
WSPD	0.2571	0.3325	0.77
T.WSPD	−0.02072	0.0118	−1.76

These results suggest that the comparability between the binomial and Poisson models is not as good as had been previously supposed. The binomial model is preferred, both on the reported NLLH values and because it more directly reflects the nature of the data, being based on daily maxima so that by definition there is no more than one exceedance per day. The reason for the difference between the models would seem to be that the p_i or λ_i are not *uniformly* small. There are some days when the weather will be conducive to a high ozone level and for these days the difference between the Poisson and binomial distributions is important.

For this reason, we do not pursue the Poisson model any further in the present paper, though in order to establish a firm framework for future studies of the same nature, it would seem to be important to continue to consider both approaches.

Fig. 5.4: Network maxima excesses (u=119.9)


(a) Exp.

(b) GPD

(c) T-Exp.

(d) TGPD

Analysis of network maxima

As explained in Section 2, we have also considered model fits to a data file of network maxima constructed from a subset of 16 ozone stations (April-October only), using a median polish technique to interpolate missing values (Bloomfield *et al.* 1993, Section 6.6). The same analysis was applied to these data, using a threshold 119.9. The best model found is given in Table 3.11. Note that YEAR, YEAR2 and PDAY are all included in this analysis and all appear significant as judged by the standard errors.

Table 3.11: Network maxima, threshold 119.9

Variable	Estimate	Stand. error	*t* Ratio
CONST	−22.9	2.357	−9.72
YEAR	.3416	.1632	2.09
YEAR2	−.04268	.01395	−3.06
CDAY	−.8855	.7036	−1.26
SDAY	.5628	.2911	1.93
PDAY	.6829	.2736	2.50
T	.2928	.02747	10.66
RH	−.01799	.01058	−1.70
WIND.U	−.03688	.04011	−.92
WIND.V	−.1887	.0449	−4.20
VIS	−.06959	.02065	−3.37
T.WSPD	−.01945	.002992	−6.50

As a comparison of different models, Table 3.12 gives the NLLH values for six models in which PDAY is either present or absent, and there is no trend, a linear trend (YEAR) or a quadratic trend (YEAR and YEAR2).

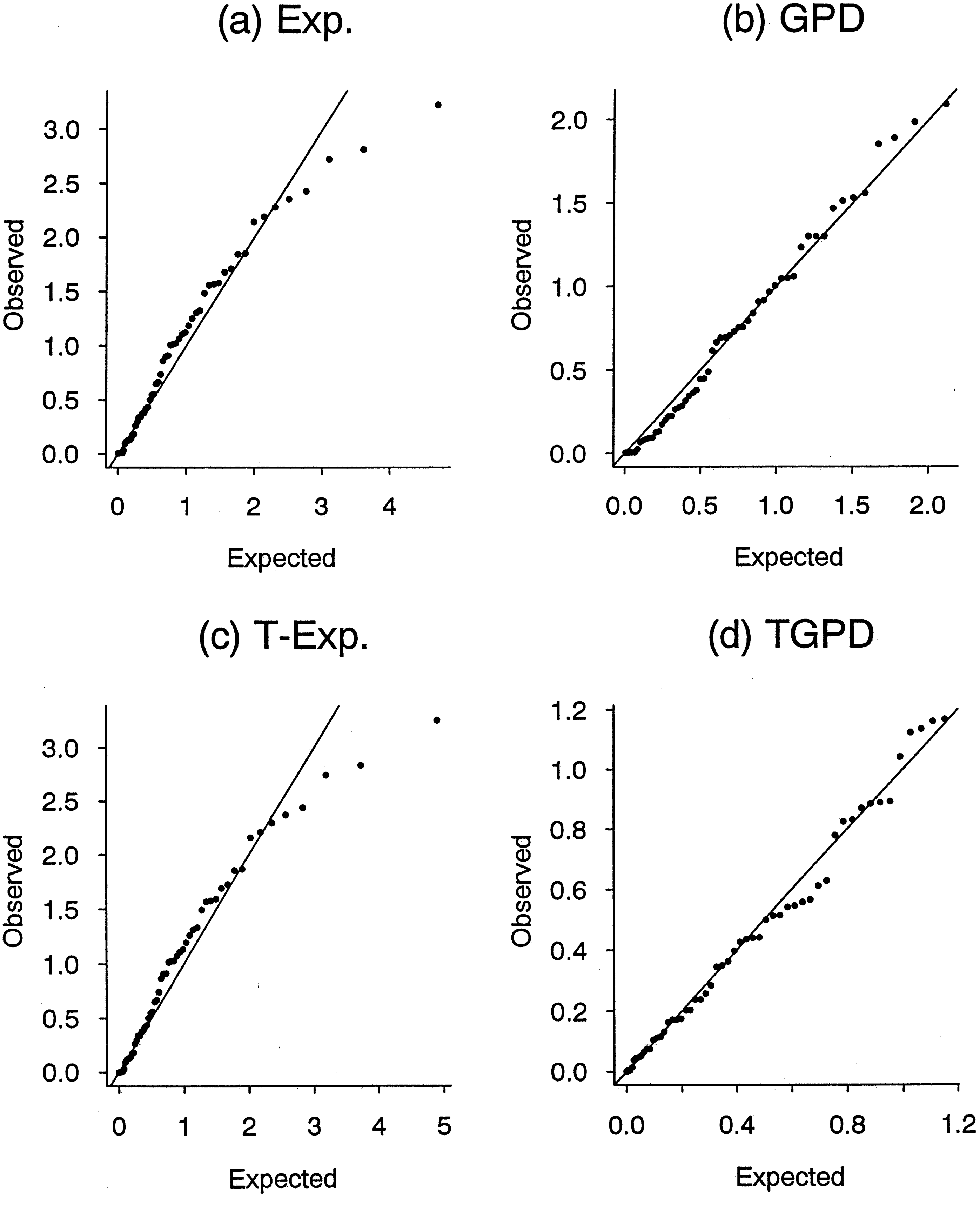
Table 3.12: Network maxima, threshold 119.9
Comparison of models using NLLH

Trend/PDAY	No	Yes
None	319.299	313.211
Linear	309.023	305.025
Quadratic	302.984	299.934

Table 3.12 reinforces the conclusion that YEAR, YEAR2 and PDAY are all significant. However, adding PDAY2 did not improve the fit.

Our results up to this point reinforce the overall conclusion that there has been a downward trend in the rate of ozone exceedances, after accounting for meteorological

Fig. 5.3: Station R excesses (u=99.9)



effects, during the period of study. In the two cases (station P and the network maxima) where we found evidence of a quadratic trend, the fitted model supports a slight increase in ozone levels up to 1984 or 1985, followed by a decrease in the subsequent years.

Predictive assessment of model fit

An important part of any statistical modeling is to be able to assess the fit of the model adopted. In the case of continuous data, there are a number of approaches, both informal ones such as graphical plots of residuals, and more formal procedures such as the Kolmogorov-Smirnov and Cramer-von Mises tests, which are widely used. For inhomogeneous binary data, such as we have here, these approaches are not so easily applied.

Nevertheless, there is an extensive literature of the subject of probability forecasting, well exemplified by the references Dawid (1982, 1984, 1986), DeGroot & Fienberg (1983) and Seillier-Moiseiwitsch & Dawid (1993). The usual paradigm is weather forecasting. Each day, a forecaster quotes the percentage probability of rainfall, in practice invariably expressed to the nearest 10%. After many days, we are able to compare the sequence of forecasts with the sequence of binary variables representing the observation of whether or not it rained on each day. How should we assess how well the forecaster is performing?

One criterion is that the forecaster should be *well-calibrated*. This means that observed frequencies should be close to forecast probabilities. For example, if we were to look at all the days on which the forecaster had quoted a 30% chance of rain, then the actual frequency of rain on those days should be close to 30%.

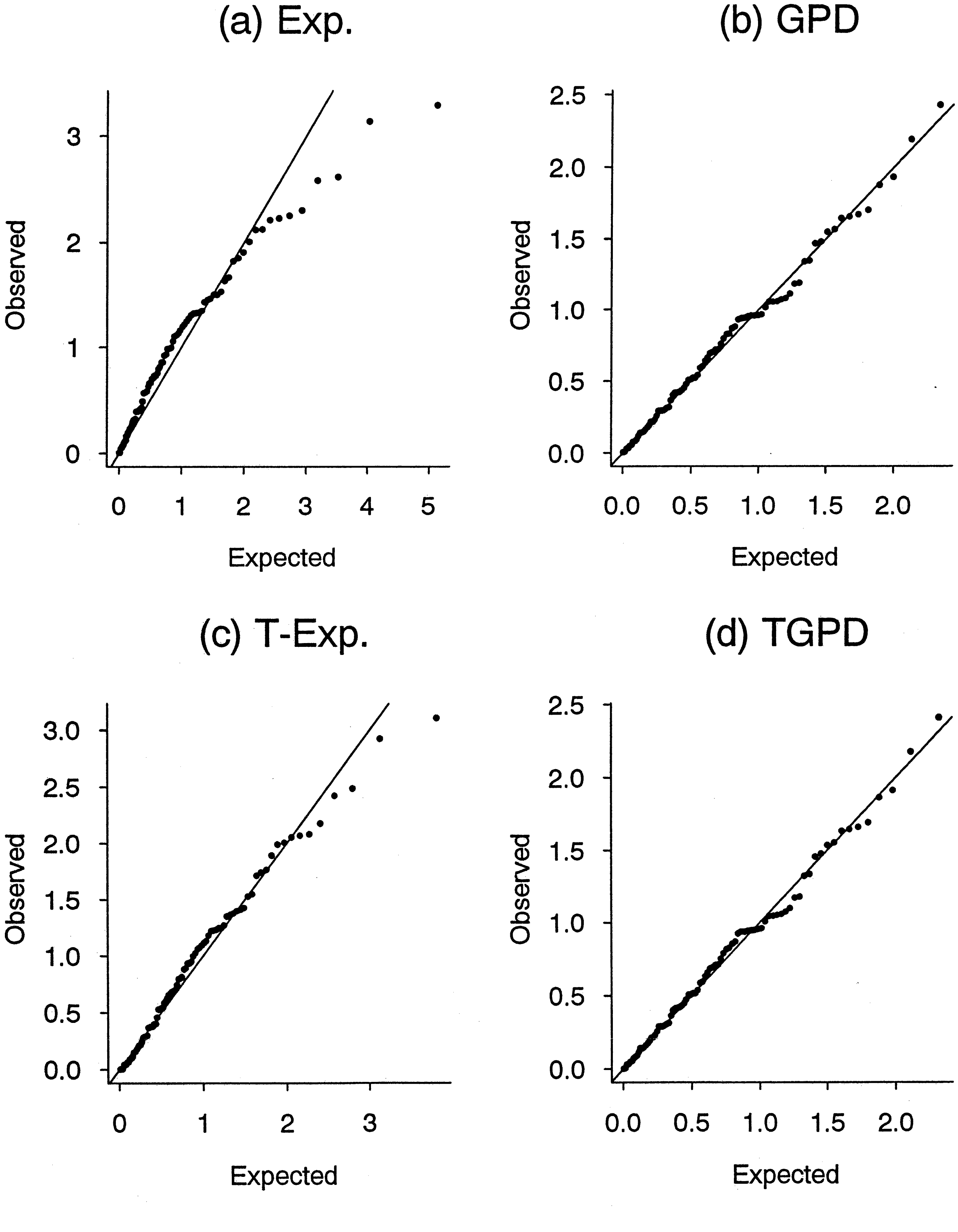
However, being well-calibrated is not enough to make a good forecaster. For example, a forecaster could be perfectly calibrated by making exactly the same forecast every day. If this forecast probability matched the true long-run probability of rainfall, then the forecaster would indeed be well calibrated, but of course such a forecast would be entirely useless in practice.

A second criterion is *resolution* (also called *refinement*). This is a measure of how well the forecaster discriminates. A forecaster who always quoted the probability of rain as either 0 or 1 would have maximum resolution, while one who always quoted the same probability would have the minimum.

It is not too difficult to measure how well a forecaster is calibrated. Chi-squared goodness of fit statistics can be constructed based on groupings of forecast values, and we shall give examples of this below. The main technical difficulty associated with this method is the justification of an asymptotic chi-squared distribution when the forecasts are sequential and possibly dependent on past data. However, much is now known about this problem, cf. Seillier-Moiseiwitsch & Dawid (1993).

Resolution is much harder to measure. Dawid (1982, 1986) argued that a criterion for a forecaster to be both well-calibrated and to have good resolution should be that her

Fig. 5.2: Station Q excesses (u=99.9)



(a) Exp.

(b) GPD

(c) T-Exp.

(d) TGPD

forecasts remain well-calibrated when restricted to any subsequence of the data, subject to a selection rule that requires the decision whether or not to include a particular day in the subsequence should depend only on the information available to the forecaster herself, which in most cases means the data available from previous days. DeGroot & Fienberg defined rules for comparing two forecasters in terms of whether one was more refined than the other. It is not so easy, however, to construct specific tests based on these concepts.

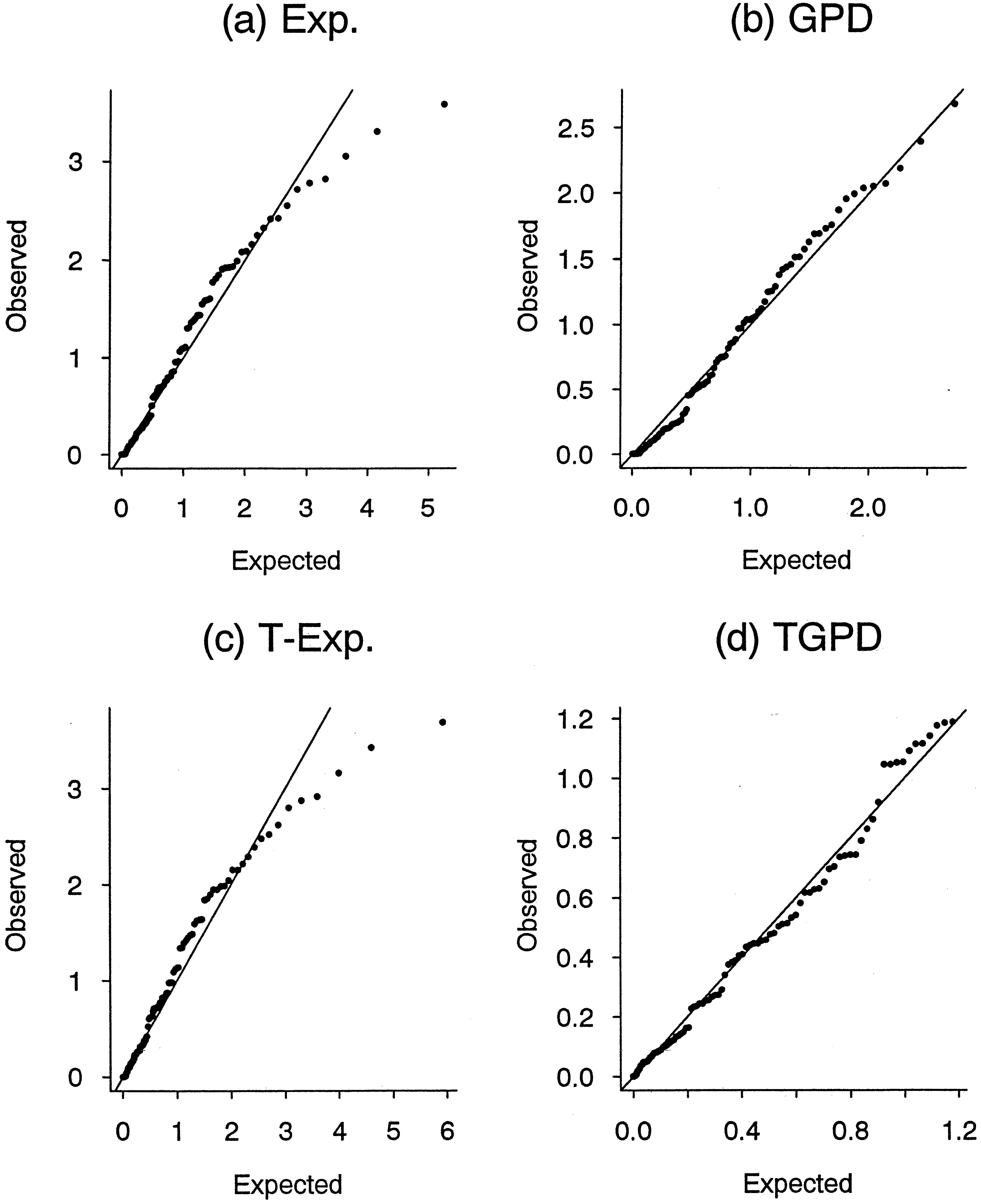
An older idea is that of *scoring rules.* A scoring rule is simply a measure or score of how well a forecaster is performing. There are a number of such rules known, but the best known are the *Brier scoring rule* $\sum(a_i - p_i)^2$ and the *logarithmic scoring rule* $\sum\{-a_i \log p_i - (1-a_i)\log(1-p_i)\}$. Here a_i is the observed value (0 or 1) on day i. These are both *proper* scoring rules, which have the property that if there is such a thing as a "true model" generating the data, then the minimum score will be achieved by quoting the probabilities given by that model.

In the context of the present study, prediction of ozone exceedances is not one of our primary aims. We are much more concerned with long-term issues such as whether there is a trend in the data or whether a particular year such as 1988 represents unusual conditions when judged against long-term climatology. Nevertheless, the models that have been used in this section are very much of a predictive nature, since they allow us to quote a "probability of exceedance" based on current weather conditions and (if PDAY is included) past ozone exceedances. Therefore, we can use ideas from the probability forecasting literature to assess their goodness of fit. In the models with PDAY, this is a genuinely sequential problem, so we do need the recent theory of sequential tests developed by Dawid (1984) and Seillier-Moiseiwitsch & Dawid (1993).

There is, however, one further complication in applying these ideas here. The literature we have described is concerned with an "honest" forecasting procedure in which the forecast for day i depends only on data observed preceding day i. This is not the case if we are considering a parametric model whose parameters have been estimated from the whole data set. The issue is similar to the familiar one in goodness of fit testing, that tests constructed on the assumption that the model is known are not valid without some adjustment if the model depends on parameters which are estimated from the data.

One version of this problem has been considered by Seillier-Moiseiwitsch (1993) for the case of a linear logistic model. She considers a procedure in which, before making a forecast for day i, the parameters of the model are re-estimated based on all the data up to time $i-1$. Under this set up, the asymptotic properties of the procedure are identical to those of a sequential forecasting scheme as in Seillier-Moiseiwitsch & Dawid (1993). The difference between the two papers is that the analysis of Seillier-Moiseiwitsch & Dawid (1993) is based on the null hypothesis that the sequential forecasting scheme is indeed the exact model that generated the data, whereas Seillier-Moiseiwitsch (1993) assumes that the parametric model is correct and makes explicit allowance for the fact that the parameters at each stage are estimated.

Fig. 5.1: Station P excesses (u=99.9)



(a) Exp.

(b) GPD

(c) T-Exp.

(d) TGPD

In the present context, it is not practicable to update the model after every observations but we have employed a variant which seems to achieve the same effect: for each year, the model was refitted to the data omitting that year and then used to produce probability forecasts for the year. All the resulting forecasts were then combined to assess how well they performed on the observed data on threshold exceedances. Thus the analysis is honest in the sense that no forecast probability of exceedance depends on the exceedance itself, or on nearby correlated values.

As an example of these ideas, Table 3.13 is concerned with the calibration of probability forecasts produced by the model of Table 3.1. This table is very similar to Table 1 of Seillier-Moiseiwitsch & Dawid (1993). Each row of the data represents a specific interval of forecast probabilities, denoted $(p_{min}, p_{max}]$. The frequency n denotes the number of days for which the forecast lay in that interval, and r is the observed number of exceedances based on forecasts within the interval. The next value e is the expected number of exceedances $\sum p_i \; I(p_{min} < p_i \leq p_{max})$ and w is its variance $\sum p_i(1-p_i) \; I(p_{min} < p_i \leq p_{max})$, under the assumption that the p_i do indeed represent true forecast probabilities. The final column gives the test statistic $z = (r-e)/\sqrt{w}$. In large samples, these will have approximately standard normal distributions, independent for non-overlapping probability intervals. The main difference from Seillier-Moiseiwitsch & Dawid (1993) is that we have used unequal probability intervals to reflect the fact that the forecast probabilities are heavily weighted towards 0. Finally, the last row of Table 3.13 gives an overall assessment of calibration by combining the intervals together.

Table 3.13: Calibration of probability forecasts
Station P, threshold 119.9, linear trend without PDAY

p_{min}	p_{max}	n	r	e	w	z
0.000	0.050	3210	9	8.602	8.406	0.137
0.050	0.100	76	2	5.508	5.095	−1.554
0.100	0.300	72	17	11.182	9.302	1.908
0.300	0.500	20	9	8.042	4.723	0.441
0.500	1.000	14	4	9.084	3.091	−2.892
0.000	1.000	3392	41	42.418	30.617	−0.256

Considering the z values in the last column, it can be seen that there is some indication of significant values in the third and fifth rows. Computing $\sum z^2$ based on the first 5 rows, we obtain 14.63, which is significant against the χ^2_5 null distribution (5% point 11.07, 1% point 15.09). Thus there is some indication that the model is not well calibrated.

The same calculation has been repeated for the models with no trend, a quadratic trend, and a linear trend including PDAY. The overall pattern of results was similar the Table 3.13 in each case, but the values of $\sum z^2$ were respectively 3.35, 14.27 and 7.62.

Fig. 4.3: Ratio of parameters under two fits
Network maxima: Logit-quadratic model

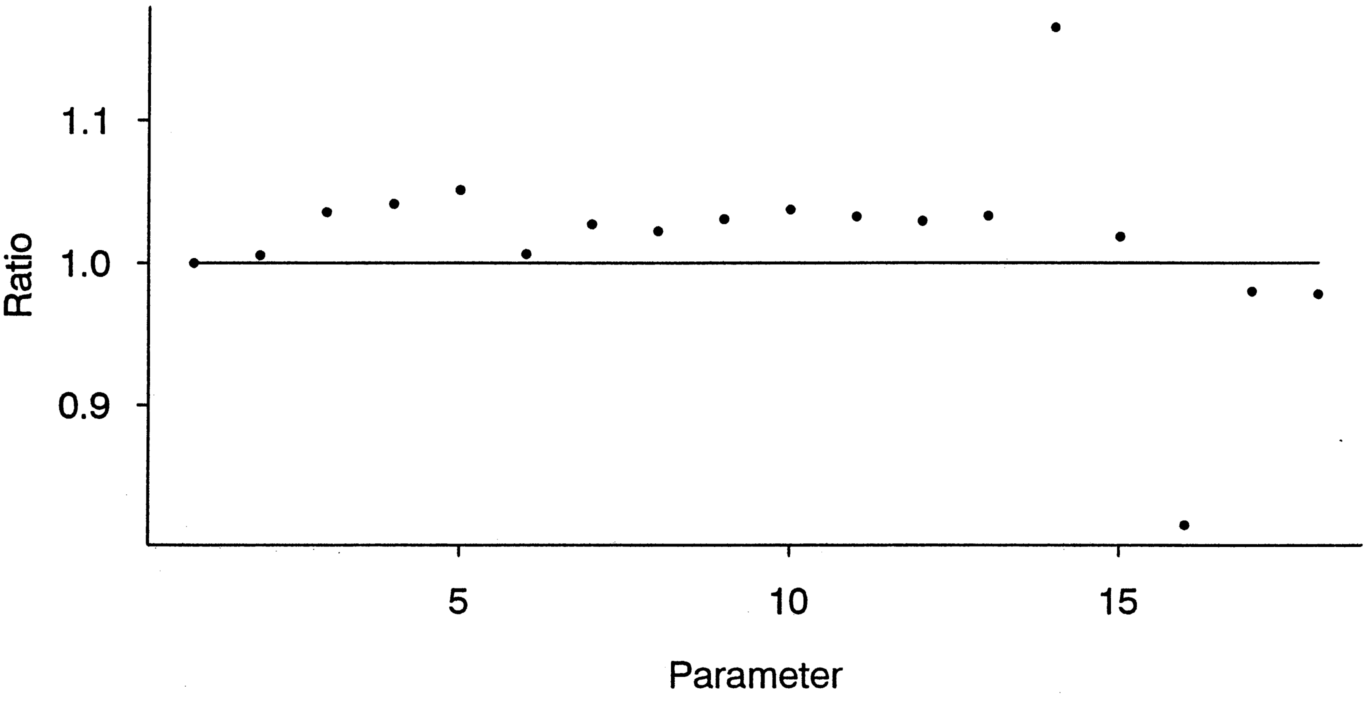
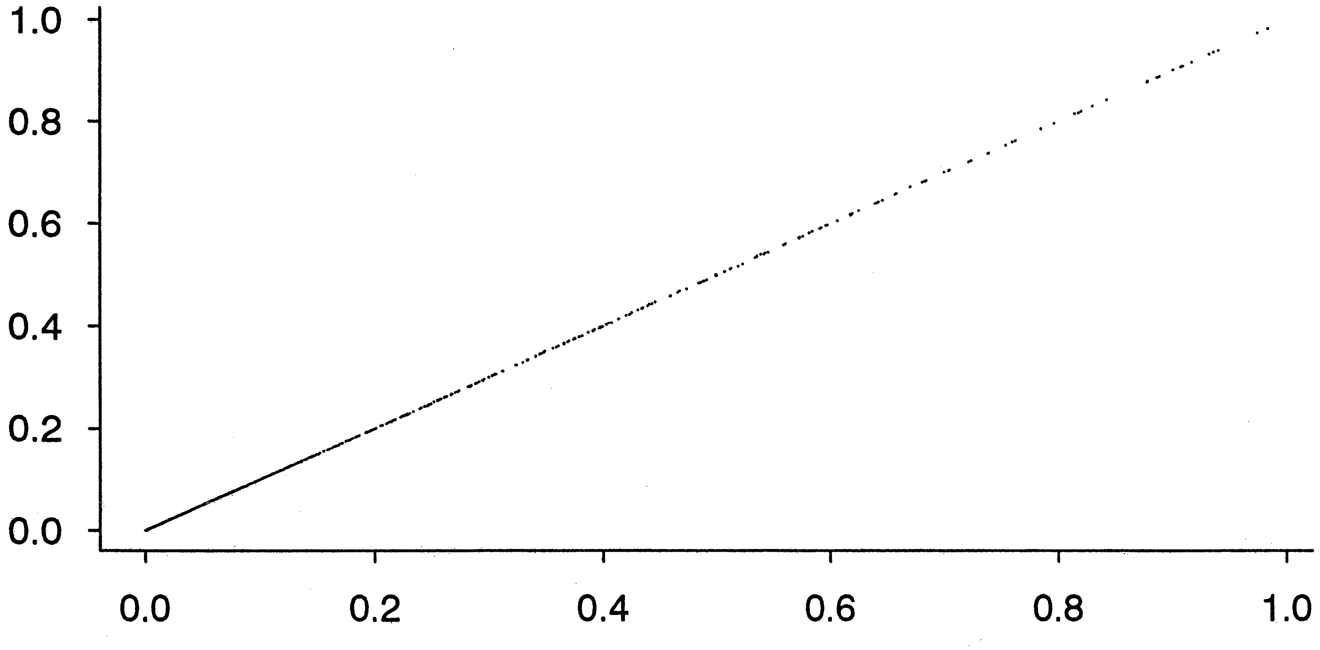


Fig. 4.4: Scatterplot of forecast probabilities
from two models in Fig. 4.3



Thus it seems, contrary to our earlier evidence, that the model with no trend is the best calibrated amongst these four, but also that the model with linear trend and PDAY passes the test. For the moment we treat these result cautiously since it is not at all clear how sensitive a test this is (for example, we do not know how sensitive it is to the grouping intervals). The other values for the overall z value (last row) are 0.137, −0.090 and −0.245, none of them remotely significant.

A second way of classifying the data is by year. Table 3.14 shows a table constructed similarly to Table 3.13 but where the rows represent individual years of data. Recall that each year's entry is based on a model fit excluding that year's data. The z values suggest that the model significantly underpredicted in 1988, and (surprisingly) overpredicted in 1991. The other clear feature of the z values is that they go from negative to positive and back to negative – it was in fact this feature which induced us to consider the quadratic trend.

Table 3.14: Calibration of probability forecasts by year
Station P, threshold 119.9, linear trend without PDAY

Year	n	r	e	w	z
1981	268	2	2.775	2.022	−0.545
1982	280	1	3.274	2.754	−1.370
1983	307	7	10.209	6.164	−1.293
1984	329	4	2.507	2.320	0.980
1985	324	3	1.648	1.553	1.085
1986	324	2	1.690	1.588	0.246
1987	323	6	3.130	2.791	1.718
1988	328	14	8.466	5.293	2.406
1989	331	1	2.726	2.095	−1.193
1990	317	0	0.677	0.653	−0.838
1991	261	1	5.316	3.384	−2.346

Table 3.15–3.17 show the corresponding results for no trend and for the quadratic trend without PDAY, and for the linear trend with PDAY. For no trend (Table 3.15), the individual z values again do not look too bad, except for 1991, but there is a clear trend. Except for 1982, the z values are all positive at the beginning and negative at the end, and this reinforces the fact that we really do need a model with trend to fit these data adequately. For the quadratic trend (Table 3.16), there is no evident trend and the only apparently significant z value is the first (1981), though since this is based on a very small expected number of exceedances it should probably not be taken too seriously. Finally, the linear trend with PDAY (Table 3.17) shows similar results to Table 3.14 but generally a slightly better fit – in particular, the z values for 1988 and 1991, though still the largest two in absolute value, are reduced compared with Table 3.14.

Fig. 4.1: Exceedances and projections, station P
Nonlinear models

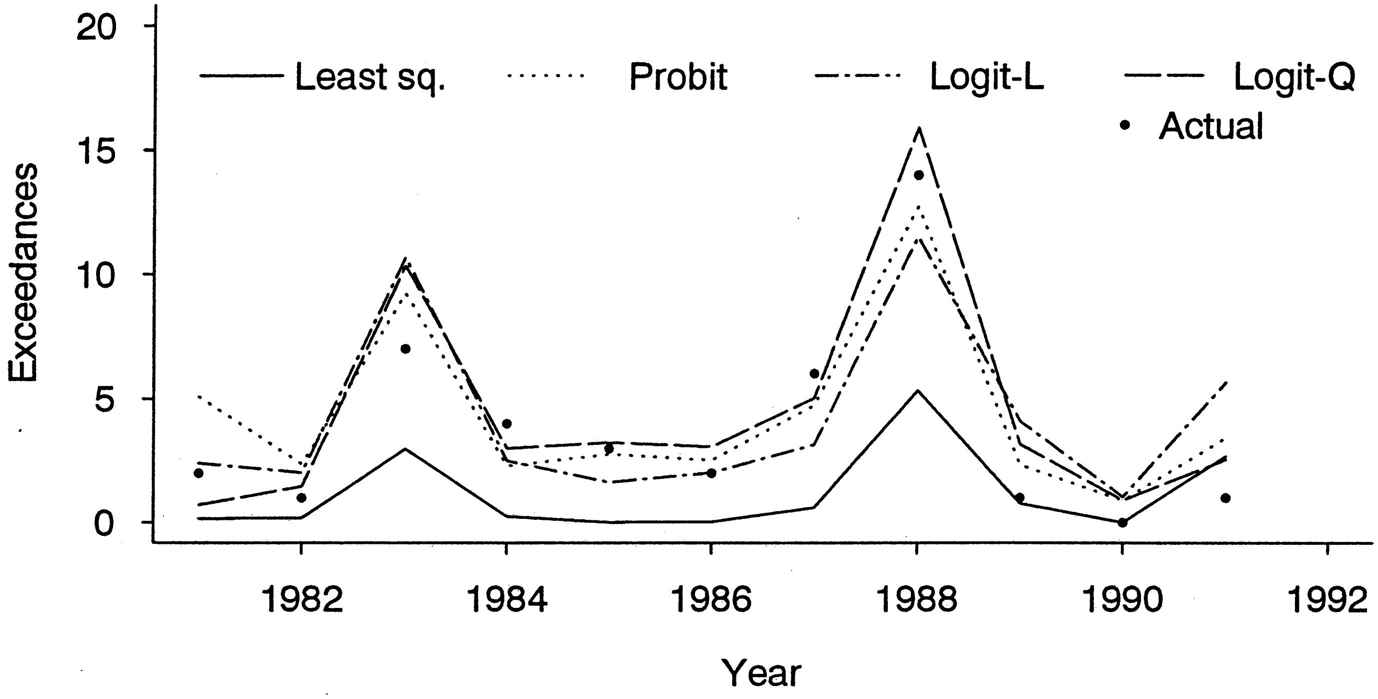


Fig. 4.2: Exceedances and projections, network maxima
Nonlinear models

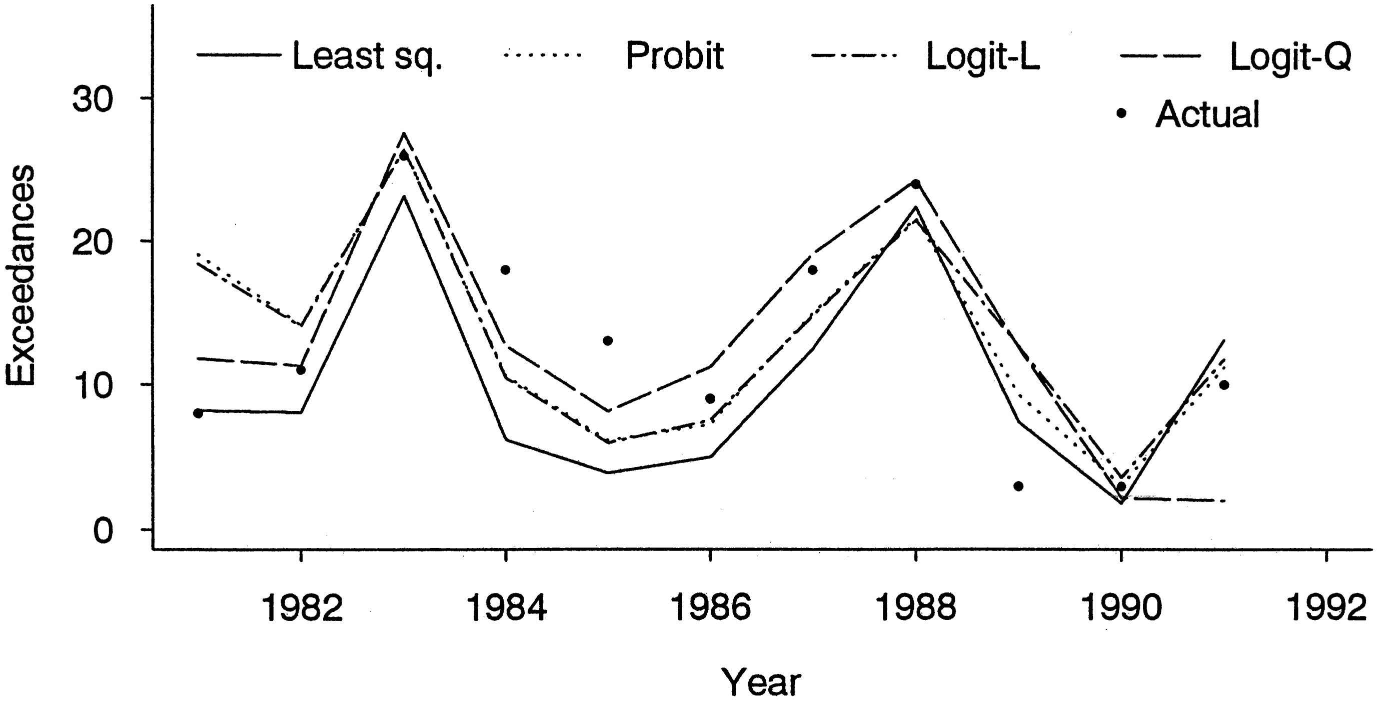


Table 3.15: Calibration of probability forecasts by year Station P, threshold 119.9, no trend, no PDAY

Year	n	r	e	w	z
1981	268	2	1.400	1.199	0.548
1982	280	1	1.664	1.531	−0.537
1983	307	7	6.168	4.447	0.395
1984	329	4	1.941	1.832	1.521
1985	324	3	1.483	1.410	1.278
1986	324	2	1.766	1.664	0.181
1987	323	6	3.851	3.375	1.170
1988	328	14	9.974	6.258	1.609
1989	331	1	3.821	2.812	−1.682
1990	317	0	1.292	1.219	−1.170
1991	261	1	6.891	4.226	−2.866

Table 3.16: Calibration of probability forecasts by year Station P, threshold 119.9, quadratic trend without PDAY

Year	n	r	e	w	z
1981	268	2	0.509	0.473	2.168
1982	280	1	1.842	1.635	−0.659
1983	307	7	10.833	6.054	−1.558
1984	329	4	3.390	3.050	0.349
1985	324	3	2.727	2.469	0.174
1986	324	2	3.020	2.697	−0.621
1987	323	6	4.862	4.047	0.566
1988	328	14	10.168	5.760	1.597
1989	331	1	2.573	1.833	−1.162
1990	317	0	0.287	0.282	−0.541
1991	261	1	1.276	1.062	−0.268

Fig. 3.1: Exceedances and projections, station P

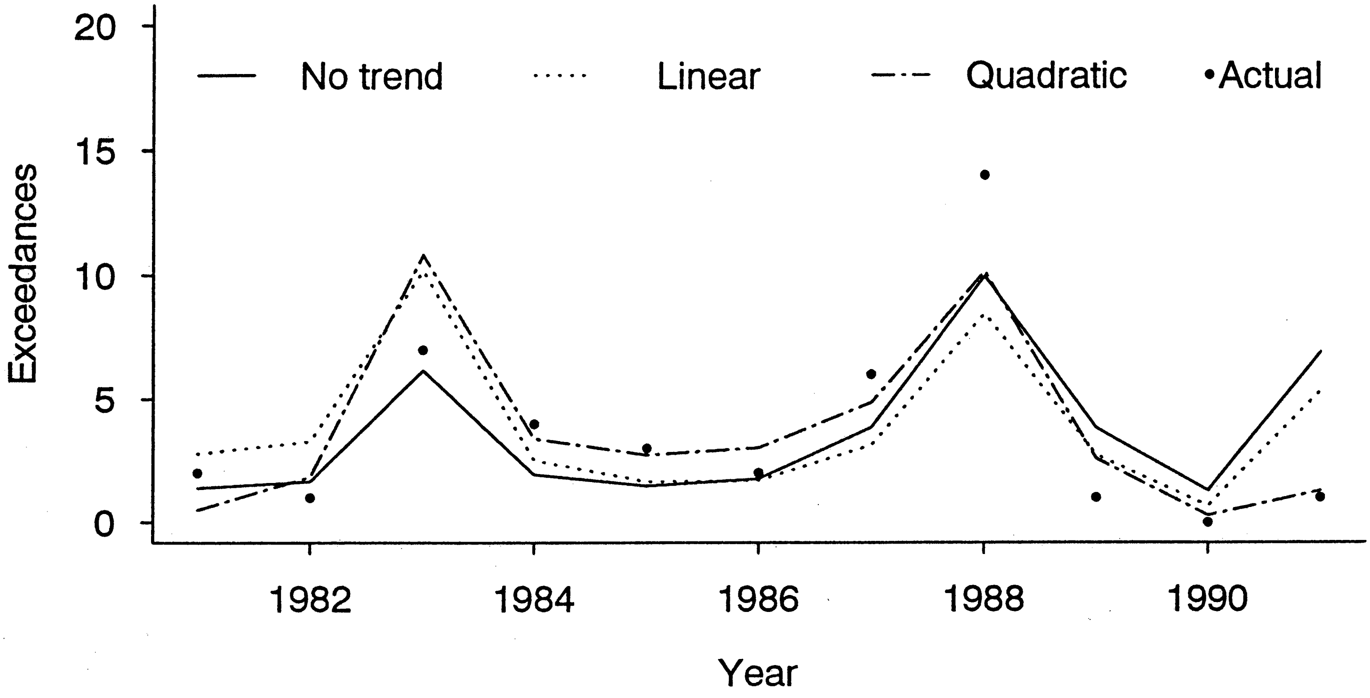


Fig. 3.2: Exceedances and projections, network maxima

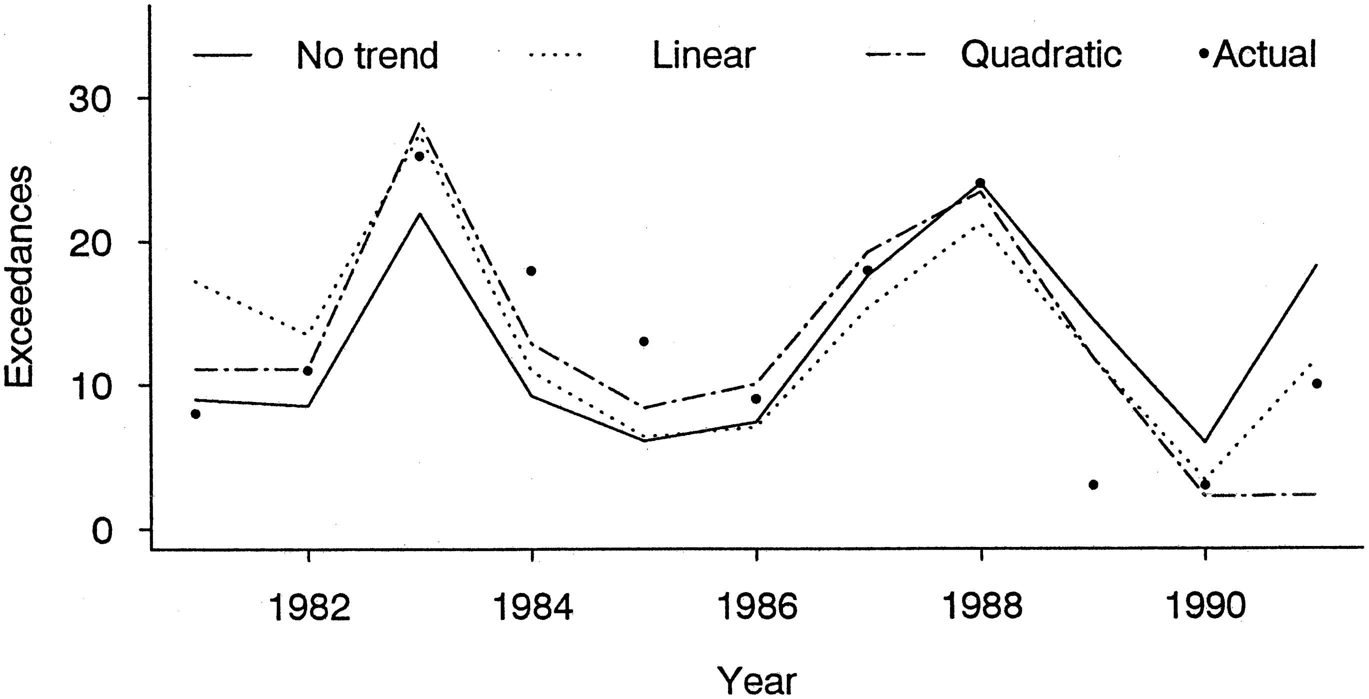


Table 3.17: Calibration of probability forecasts by year
Station P, threshold 119.9, linear trend with PDAY

Year	n	r	e	w	z
1981	268	2	2.514	1.932	−0.370
1982	280	1	3.088	2.664	−1.279
1983	307	7	10.079	5.494	−1.314
1984	329	4	2.653	2.390	0.871
1985	324	3	1.903	1.734	0.833
1986	324	2	1.604	1.519	0.321
1987	323	6	3.085	2.755	1.756
1988	328	14	9.589	5.481	1.884
1989	331	1	2.398	1.971	−0.996
1990	317	0	0.632	0.613	−0.807
1991	261	1	4.792	3.180	−2.127

Finally, we consider the Brier scores. For the four models of Tables 3.14-3.17, the Brier scores are 33.95 (0.97), 34.63 (1.00), 31.50 (0.55) and 34.36 (1.15). Here the figures in parentheses are z values for the Brier scores, corresponding to the quantity called Y_n^B in Seillier-Moiseiwitsch & Dawid (1993), page 359. None of these values are significant, but on the other hand this is just another measure of calibration. Of more interest than the z value is the Brier score itself, and we note again that the model with no trend appears to do better than either of the models with a linear trend, but from this point of view the quadratic trend model really does perform the best of the four.

Now let us consider the same analysis based on network maxima. For the best model we found (quadratic trend including PDAY), the calibration results are summarized in Tables 3.18 and 3.19. In Table 3.18, the forecast probabilities have been classified into a greater number of intervals to reflect the greater number of exceedances. The only z value to look remotely significant is that for the first row, but the overall value of $\sum z^2$ (9.51) is clearly not significant against its nominal χ^2_{10} distribution. Table 3.19 is a little more disturbing since the model has clearly overpredicted the exceedances for 1989 and greatly underpredicted for 1991, though for the most important years, 1983 and 1988, the agreement of observed and predicted numbers of exceedances is remarkably good. The underprediction for 1991 should make us cautious about extrapolations based on the quadratic trend, especially when contrasted with the overprediction for 1991 that we observed in Table 3.14 when using a linear trend for station P.

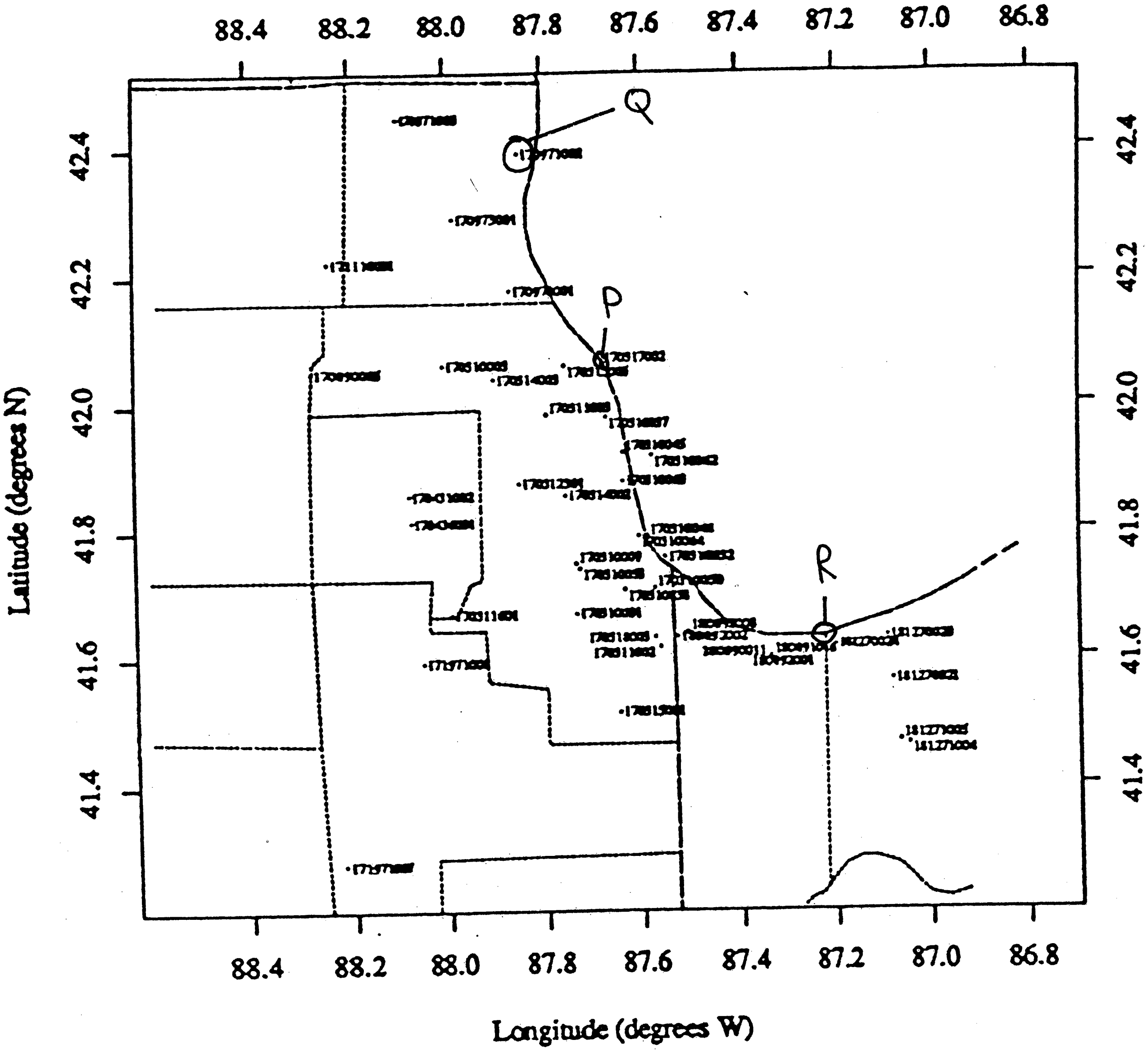


Figure 2.1: Location of measuring stations with stations P, Q and R marked.

Table 3.18: Calibration of probability forecasts
Network maxima, threshold 119.9, quadratic trend with PDAY

p_{min}	p_{max}	n	r	e	w	z
0.000	0.025	1660	13	7.724	7.621	1.911
0.025	0.050	182	10	6.574	6.327	1.362
0.050	0.075	102	3	6.282	5.890	−1.352
0.075	0.100	65	5	5.640	5.147	−0.282
0.100	0.200	132	18	18.949	16.131	−0.236
0.200	0.300	70	21	17.692	13.161	0.912
0.300	0.400	45	15	15.832	10.227	−0.260
0.400	0.500	28	11	12.629	6.913	−0.620
0.500	0.750	46	26	28.813	10.551	−0.866
0.750	1.000	24	21	20.889	2.614	0.069
0.000	1.000	2354	143	141.024	84.581	0.215

Table 3.19: Calibration of probability forecasts by year
Network maxima, threshold 119.9, quadratic trend with PDAY

Year	n	r	e	w	z
1981	214	8	11.119	7.642	−1.128
1982	214	11	11.129	7.847	−0.046
1983	214	26	28.326	12.358	−0.662
1984	214	18	12.870	8.695	1.740
1985	214	13	8.398	6.597	1.792
1986	214	9	10.058	7.907	−0.376
1987	214	18	19.274	10.312	−0.397
1988	214	24	23.433	11.509	0.167
1989	214	3	11.909	7.750	−3.200
1990	214	3	2.207	2.026	0.557
1991	214	10	2.301	1.938	5.530

This analysis was repeated for all six models in Table 3.12, and in all six cases the basic calibration table was similar to Table 3.18, with no significant discrepancies. For the analysis by year, the model with no trend showed a distinct pattern of z values similar to that for station P, with a significant underprediction of exceedances in 1984 and 1985 and a significant overprediction in 1989 and 1991. The model with linear trend again showed a pattern of negative z values at the beginning, then positive, and then negative again, with the most significant z values being −2.823 (1981), 2.546 (1984), 2.886 (1985) year are based on refitting the model to the rest of the data with that year omitted, and using the refitted model to determine the expected number of exceedances in that year.

Figure 5.6 Same as Figure 5.5, for station Q with no-trend and linear-trend model fits.

Figure 5.7 Same as Figure 5.5, for station R with no-trend and linear-trend model fits.

Figure 5.8 Same as Figure 5.5, for network maxima fitted to threshold 119.9, with no-trend, linear-trend and quadratic-trend model fits.

Figure 5.9 Same as Figure 5.5, for station P with model fits based on the dependent model with linear trends in both the probability of exceedance and excess value components.

Figure 5.10 Same as Figure 5.8, for network maxima with model fits based on the dependent model with linear trends in both the probability of exceedance and excess value components.

Figure 5.11 Simulations of the final (dependent) model for station P. Shown is a plot of all values over 120, plotted against day from 1/1/81, for the real data (top left plot) and five independent simulations.

Figure 5.12 Same as Figure 5.11, but for network maxima and the corresponding fitted model.

Figure 7.1 Expected exceedances for four thresholds in each year from 1959 (year 1) to 1991 (year 33). The calculations use actual meteorological data and predicted exceedances according to the model. The model in this case is the one fitted to network maxima assuming independent days, but refitted without any trend component.

Figure 7.2 Same as Figure 7.1, but using the model fitted to Station P.

Figure 8.1 Probability plots for normalized residuals for sums of excesses over the threshold u_0 =60, station P, based on stepwise selection of covariates. Four models: (a) exponential distribution, (b) GPD, (c) exponential distribution with power transformation, (d) TGPD.

Figure 8.2 Same as Figure 8.1, but relative to $u_0 = 80$.

Figure 8.3 Same as Figure 8.1 for station R.

Figure 8.4 Same as Figure 8.2 for station R.

and −3.192 (1989). These results are all for models including PDAY; the results without PDAY showed identical patterns in each of the three cases, but generally more extreme z values. Finally, the Brier scores and associated z values were:

No trend, no PDAY	94.82	(0.52)
No trend, include PDAY	92.54	(0.52)
Linear trend, no Pday	93.28	(0.40)
Linear trend, include PDAY	91.63	(0.40)
Quadratic trend, no PDAY	90.81	(0.75)
Quadratic trend, include PDAY	89.67	(0.86)

These results confirm the pattern that the model improves with increasing order of trend, and is better including PDAY than omitting it.

The logarithmic scoring rule has also been considered, but does not seem to produce satisfactory results. Seillier-Moiseiwitsch & Dawid (1993) comment on the difficulty of applying this rule when many of the forecasts are near 0 or 1. Our situation is better than theirs, since we are not obliged to deal with probabilities rounded to the nearest 0.1, but nevertheless we suspect that the results are highly sensitive to a few forecast probabilities very close to 0 or 1.

In Figure 3.1, we have shown the observed numbers of exceedances for station P together with the expected exceedances under each of the models with no trend, linear trend and quadratic trend. Figure 3.2 shows the same thing for the network maxima. To avoid the picture becoming too cluttered, we have shown only the models without PDAY in the case of station P, and only the models with PDAY in the case of annual maxima.

4. Comparison with the nonlinear regression approach

A natural question is how the preceding analysis ties in with the nonlinear regression approach of Bloomfield *et al.* (1993). The purpose of this section is to investigate that.

Bloomfield *et al.* (1993) constructed weighted averages of ozone over a set of the monitoring stations, and then fitted a nonlinear regression model. The model used in this section does not correspond exactly to theirs: we have used a slightly different parametrization to improve numerical stability, and have not included all the variables they did. For example, we have not included any lagged meteorological variables in our approach, and with some of the ones that are common to the two models, our definition differs slightly from theirs, for example, by being based on noon values rather than daily averages or maxima. An additional feature is that, in view of earlier results, we have included the possibility of a quadratic trend (not just linear). The model as we consider it is as follows:

Figure Captions

Figure 2.1 Locations of 43 monitoring stations around Chicago. The three specific stations identified for this study have been marked with the letters P, Q and R.

Figure 3.1 Number of exceedances per year over threshold 119.9 for Station P. Actual (observed) counts denoted by solid points; model fits by solid and broken lines.

Figure 3.2 Same as Figure 3.1, but for network maxima.

Figure 4.1 Same as Figure 3.1, but using four examples of nonlinear fits. The solid line represents a least squares fit to the whole data, analogous to the method of Bloomfield *et al.* (1993). The three broken lines represent models refitted to the binary data consisting of excesses over the threshold, using the probit model with linear trend and the logit models with both linear and quadratic trend.

Figure 4.2 Same as Figure 4.1, but for network maxima.

Figure 4.3 Illustration of instability in the model of Table 4.4. The same model was refitted to the same data starting from the previous fit. The result was a reduction in negative log likelihood of just 0.002, but with substantial changes in the individual parameter estimates. In this plot, the ratios of all 18 parameter estimates for the two fits are plotted.

Figure 4.4 Same pair of models as Figure 4.3, but the forecast probabilities of exceeding the threshold, under the two models, for each of the 2354 days in the data set, are plotted against each other. The plotted points lie almost exactly on the straight line of unit slope through the origin; maximum discrepancy 0.002. We conclude that, although the two model fits differ substantially in their parameter values, they are indistinguishable in their predictive performance.

Figure 5.1 Probability plot of normalized residuals for excesses over threshold 99.9 for station P. Four models: (a) exponential distribution, (b) GPD, (c) exponential distribution with power transformation, (d) TGPD.

Figure 5.2 Same as Figure 5.1, but for station Q.

Figure 5.3 Same as Figure 5.1, but for station R.

Figure 5.4 Same as Figure 5.1, but for network maxima and threshold 119.9.

Figure 5.5 Observed (solid points) and predicted (solid and dotted lines) numbers of exceedances of four levels, based on the logit models with no trend and with quadratic trend, for the threshold crossing probabilities and the GPD for excesses, fitted with respect to threshold 99.9 assuming independence of daily values. The predicted values for each

$$
\begin{aligned}
\mathrm{O}_3 = \Bigg[& \left\{ \frac{m_0 + m_1 \times \mathrm{WSPD} + t_1(\mathrm{T}-60) + t_2(\mathrm{T}-60)^2/1000 + t_3(\mathrm{T}-60)^3/1000}{1 + v_h \times \mathrm{WSPD}/100} \right\} \\
& \times \left\{ 1 + \frac{r(\mathrm{RH}-50)}{1000} \right\} \\
& \times \left\{ 1 + \frac{o_p(\mathrm{OPCOV}-50)}{1000} \right\} \\
& \times \left\{ 1 + \frac{v(\mathrm{VIS}-12)}{1000} \right\} \\
& \times \left\{ 1 + \frac{m_u \times \mathrm{WIND.U} + m_v \times \mathrm{WIND.V}}{1000} \right\} \\
& \times \left\{ 1 + \frac{y_1(\mathrm{YEAR}-1985)}{10} + \frac{y_2(\mathrm{YEAR}-1985)^2}{1000} \right\} \Bigg] \\
& + a_1 \cos\left(\frac{2\pi \times \mathrm{DAY}}{365.25} \right) + b_1 \sin\left(\frac{2\pi \times \mathrm{DAY}}{365.25} \right) \\
& + a_2 \cos\left(\frac{4\pi \times \mathrm{DAY}}{365.25} \right) + b_2 \sin\left(\frac{4\pi \times \mathrm{DAY}}{365.25} \right) \\
& + \text{ random error.} \qquad (4.1)
\end{aligned}
$$

Here we are following the convention of letting lower-case italic letters denote parameters, while upper case Roman letters are the covariates, defined as in Section 2. Although Bloomfield *et al.* proposed this model for weighted averages of ozone, it seems reasonable that it would apply to ozone from a single station as well. The main feature of the model is the complicated interaction term between temperature and windspeed, intended to reflect the facts that (a) ozone is a nonlinear function of temperature, the slope increasing as temperature rises, (b) high winds tend to nullify the effect of temperature. In addition, there are multiplicative terms reflecting the influence of the other meteorological variables, and additive terms reflecting seasonal effects. This form of model was determined by Bloomfield *et al.* after extensive initial examination of the data, and subsequently confirmed to be a suitable fit.

For this section we consider the network maxima first, and fit the model (4.1) to the data set, by nonlinear least squares, assuming the random errors to be independent normal with mean 0 and variance σ^2. Also, initially we assume just a linear trend ($y_2 = 0$). This produces the parameter estimates shown in Table 4.1. In this and subsequent models fits, we quote the logarithm of the scale parameter, $\log \sigma$ in place of σ, as this has again been found to improve numerical stability. The results are comparable to Table 9 of Bloomfield *et al.* (1993) — although using slightly different covariates from theirs, there is not much difference in the overall R^2 (.64 as against their .67).

Singpurwalla, N.D. (1972), Extreme values from a lognormal law with applications to air pollution problems. *Technometrics* **14**, 703-711.

Smith, R.L. (1985), Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72**, 67-90.

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

Smith, R.L. (1993), Multivariate threshold methods. Paper presented at the NIST-Temple University Conference on Extreme Value Theory and its Applications, Gaithersburg, May 1993.

Tawn, J.A. (1988), Bivariate extreme value theory - models and estimation. *Biometrika* **75**, 397-415.

Table 4.1: Nonlinear regression model of (4.1), Normal errors, fitted to all network maxima

Parameter	Estimate	Stand. error	t Ratio
m_0	56.38	2.888	19.52
m_1	13.28	1.6	8.30
t_1	2.003	0.154	13.00
t_2	63.31	6.141	10.31
t_3	-0.1989	0.221	-0.90
v_h	21.87	2.239	9.76
r	-2.933	0.4282	-6.85
o_p	-0.8051	0.171	-4.71
v	-9.19	0.8704	-10.56
m_u	-3.268	1.45	-2.25
m_v	-3.013	1.519	-1.98
y_1	-0.1317	0.01566	-8.41
a_1	-5.53	2.289	-2.42
b_1	6.715	0.9121	7.36
a_2	-1.443	1.115	-1.29
b_2	0.3204	0.9041	0.35
$\log\sigma$	2.788	0.01471	189.50

A natural question to ask is to what extent this model adequately represents the probability of crossing a high threshold. To study this, first let us rewrite equation (4.1) in the form

$$\mathrm{O}_3(i) = f(x_i;\beta) + Z_i, \tag{4.2}$$

where $\mathrm{O}_3(i)$ represents observed daily maximum ozone on day i, x_i is the vector of covariates on day i, β is the vector of unknown parameters in (4.1), and the Z_i are assumed independent N(0,σ^2). Then the probability p_i of exceeding a high threshold u, on day i, is given by

$$p_i = 1 - \Phi\left\{\frac{u - f(x_i;\beta)}{\sigma}\right\}, \tag{4.3}$$

where Φ is the standard normal distribution function. The likelihood function is defined by (3.1), with p_i defined by (4.1)-(4.3). With $u = 119.9$ as before, on the basis of the parameter values defined in Table 4.1, NLLH turns out to have the value 357.787, which is completely uncompetitive with any of the values quoted for this data set in Table 3.12.

However, this should not be surprising. We have chosen a high threshold corresponding to less than 5% of the data set; it is only natural that the model should have to be refitted to obtain good results at this threshold level. What we might expect is that the *form* of

Dawid, A.P. (1984), Statistical theory: the prequential approach (with discussion). *J.R. Statist. Soc. B* **147**, 278-292.

Dawid, A.P. (1986), Probability forecasting. In *Encyclopedia of Statistical Sciences*, Vol. 7, eds. S. Kotz, N.L. Johnson and C.B. Read. New York: Wiley-Interscience, 210-218.

DeGroot, M.H. & Fienberg, S.E. (1983), The comparison and evaluation of forecasters. *The Statistician* **32**, 12-22.

Gumbel, E.J. (1958), *Statistics of Extremes.* Columbia University Press, New York.

Karr, A.F. (1986), *Point Processes and their Statistical Inference.* Marcel Dekker, New York.

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

National Research Council (1991). *Rethinking the Ozone Problem in Urban and Regional Air Pollution.* National Academy Press, Washington, D.C.

Pickands, J. (1975), Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1986), *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press.

Resnick, S. (1987), *Extreme Values, Point Processes and Regular Variation.* Springer Verlag, New York.

Roberts, E.M. (1979), Review of statistics of extreme values with applications to air quality data, I: review, II: applications. *J. Air Pollut. Control Assoc.* **29**, 632-637 and 733-740.

Seillier-Moiseiwitsch, F. & Dawid, A.P. (1993), On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc* **88**, 355-359.

Seillier-Moiseiwitsch, F. (1993), Predictive assessment of logistic models. Unpublished technical report.

Shively, T.S. (1991), An analysis of the trend in ground-level ozone using nonhomogeneous Poisson processes. *Atmospheric Environment* **25B**, 387-396.

Shreffler, J.H. (1993), Trends in high ozone concentrations in St. Louis, 1983-1991. Department of Statistics, University of North Carolina.

the model will be an improvement on the linear dependence on covariates of the models in Section 3.

With this in mind, we have re-estimated the parameters by maximizing the likelihood defined by (3.1) and (4.1)-(4.3), using just the binary information on exceedances over u. This leads to NLLH=308.152, which is competitive with, but still not as good as, the best of the fits given in Section 3. The parameter estimates under this refitted model are given in Table 4.2, and it can be seen that they are completely different from those in Table 4.1. In effect, what we have here is a nonlinear probit model.

Table 4.2: Nonlinear regression model of (4.1), Normal errors, exceedances over threshold 119.9 NLLH=308.152

Parameter	Estimate	Stand. error	t Ratio
m_0	103.6	42.69	2.43
m_1	6.163	7.5	0.82
t_1	0.5609	1.417	0.40
t_2	16.7	45.93	0.36
t_3	0.005362	0.298	0.02
v_h	7.575	4.018	1.89
r	-0.5859	1.735	-0.34
o_p	-0.05895	0.2791	-0.21
v	-2.738	7.706	-0.36
m_u	-1.497	5.116	-0.29
m_v	-8.071	22.63	-0.36
y_1	-0.0654	0.1829	-0.36
a_1	2.197	11.2	0.20
b_1	3.186	8.811	0.36
a_2	2.33	7.283	0.32
b_2	0.7938	3.187	0.25
$\log \sigma$	2.12	2.61	0.81

An alternative model widely used in binary data analysis is based on the logistic distribution, i.e. replace $\Phi(\cdot)$ in (4.2) by $\Psi(\cdot)$, where $\Psi(x) = e^x/(1+e^x)$. In fact this is precisely the logit model of Section 3, but with a nonlinear instead of linear regression function. Optimizing this form of the likelihood produces a NLLH of 307.763, with parameter estimates shown in Table 4.3.

In Section 6, some preliminary and tentative suggestions were made as to how the results might be interpreted from the point of view of identifying extreme ozone days adjusted for meteorology.

In Section 7, we addressed the issue of "recurrence of 1988" by using the models (refitted without trend) to project ozone exceedances across the whole period for which meteorological data are available (1959-1991). The result confirm that 1988 and 1983 were the two most extreme years in the whole period, from the point of view of meteorological conditions conducive to high ozone. On the other hand, we have also seen that the models did not fully explain the very high ozone levels that actually occurred in 1988. Thus we still leave open the possibility that additional factors, over and above anything that could be explained by meteorology, were responsible for the ozone levels of that year.

Finally, in section 8 we have made some preliminary proposals for the distribution of sums over a high threshold, though we concentrated on the definition (8.1), for which some probabilistic theory is available, rather than (8.2) which corresponds to SUM06. In view of continuing work on this topic, we have not attempted a fully comprehensive analysis here.

References

Anderson, C.W. & Dancy, G. (1992), The severity of extreme events. Preprint, Sheffield University.

Bloomfield, P., Royle, A. & Yang, Q. (1993), Accounting for meteorological effects in measuring urban ozone levels and trends. National Institute of Statistical Sciences Techical Report #1.

Cox, D.R. & Lewis, P.A.W. (1966), *The Statistical Analysis of Series of Events.* Methuen, London.

Cox, W.M. & Chu, S.-H. (1992), Meteorologically adjusted ozone trends in urban areas: A probability approach. U.S. Environmental Protection Agency Technical Support Division MD-14, Research Triangle Park, NC 27711.

Davison, A.C. & Hemphill, M.W. (1987), On the statistical analysis of ambient ozone data when measurements are missing. *Atmospheric Environment* **21**, 629-639.

Davison, A.C. & Smith, R.L. (1990), Models for exceedances over high thresholds (with discussion). *J.R. Statist. Soc. B* **52**, 393-442.

Dawid, A.P. (1982), The well-calibrated Bayesian (with discussion). *J. Amer. Statist. Assoc* **77**, 605-613.

Table 4.3: Nonlinear regression model of (4.1), Logistic errors, exceedances over threshold 119.9 NLLH=307.763

Parameter	Estimate	Stand. error	t Ratio
m_0	118.4	4.196	28.23
m_1	3.939	4.213	0.93
t_1	0.04696	0.1388	0.34
t_2	1.345	3.802	0.35
t_3	-0.01847	0.0523	-0.35
v_h	3.427	3.774	0.91
r	-0.04245	0.1257	-0.34
o_p	-0.005602	0.02154	-0.26
v	-0.1829	0.5268	-0.35
m_u	-0.09017	0.3164	-0.28
m_v	-0.5198	1.46	-0.36
y_1	-0.004074	0.01139	-0.36
a_1	-0.2072	1.387	-0.15
b_1	0.2691	0.8155	0.33
a_2	0.02903	0.4356	0.07
b_2	0.08388	0.333	0.25
$\log \sigma$	-1.19	2.77	-0.43

So far, we have assumed $y_2 = 0$ in (4.1). If we add y_2 to the model, the improvement of the fit is negligible when fitted by least squares to the whole data set, but it does have a significant effect in the threshold analysis. As an example, Table 4.4 shows the same model as Table 4.3 but including y_2.

We also calculated the predictive diagnostics discussed at the end of Section 3. For the least squares model without re-estimation of parameters, the overall calibration was very poor. The calibration table corresponding to Table 3.18 produced several significant z values and an overall $\sum z^2$ of 80.95, clearly significant against χ^2_{10}. The overall e value (last row of table) was 111.8, as against $r = 143$, resulting in a highly significant $z = 4.02$. Looking at individual year's results, the greatest underprediction of exceedances was for 1984 and 1985. For the models refitted in Tables 4.2 and 4.3, the results are very similar to the results discussed in Section 3 for a linear trend: the overall calibration table is good, but there is significant overprediction of exceedances for 1981 and 1989, and significant underprediction in 1984 and 1985. The quadratic fit of Table 4.4 results in good predictions for every year except 1989, when it overpredicts, and 1991, when it underpredicts. The Brier scores resulted in 93.84 (model from Table 4.1), 91.49 (Table 4.2), 93.97 (Table 4.3) and 91.64 (Table 4.4). Thus it is seen that the original fit from Table 4.1, although by maximum likelihood, assuming the logit probability is linear in the covariates and with stepwise selection of covariates. However, we also considered a model with PDAY as a covariate, this being an indicator of whether the previous day was an ozone exceedance, so introducing an element of serial dependence into the model. For stations P and R, the linear trend represented by the covariate YEAR was significant, but not (in this analysis) for station Q. Also, in station P the variable PDAY was found to be significant, but not for the other two. An alternative analysis based on threshold 99.9 in fact reduces the significance of the YEAR variable, suggesting it is no longer significant for R and only just for P. A quadratic trend was also fitted and found significant for station P.

Comparison with an alternative Poisson process formulation showed less satisfactory fits for the Poisson model, so this was not pursued further in the paper. A corresponding analysis for network maxima showed that linear and quadratic trends were both significant, as was the PDAY variable indicating dependence.

Some ideas from the literature on probability forecasting were adapted to construct predictive measures of goodness of fit for these models. The results generally supported the conclusions that had been drawn by a likelihood analysis.

In Section 4, a direct attempt was made to compare the logit models with linear covariate structure with the nonlinear models proposed by Bloomfield *et al.* Fitting a nonlinear model by ordinary least squares, and using that to predict high exceedances, did not give good results. On the other hand, rewriting the model as a nonlinear probit or logit model, and fitting that to the exceedances over a high threshold, did produce results that were at least as good as the ones from the approach of Section 3. In some cases, the nonlinear models did better as judged by the predictive diagnostics. On the other hand, fitting the nonlinear models requires much more computing time and there are problems with numerical stability of the parameter estimates. Our conclusion is that there may be advantages in using a nonlinear model as measured by the overall fit, but these do not seem to compensate the difficulties of fitting such a model.

In Section 5, the analysis was extended to excesses over a threshold. The Generalized Pareto distribution (GPD) with covariates does uniformly well as a model for the distribution. In the case of station Q, where we earlier found no significant evidence of a trend, we do find that the trend in the excess values is significant. The predictive assessments of model fit show good results at lower levels, but suggest that the models significantly fail to predict what is happening at much higher levels. However, we also introduce a model for dependence, in which daily values are assumed to follow a Markov chain in which the joint distribution of successive pairs lies in the domain of attraction of a bivariate extreme value distribution. This model (the dependent model) does indeed seem to produce a satisfactory fit at higher levels. Although there are still observed discrepancies, especially in 1988, the overall deviation from the model is not significant as measured by the predictive diagnostics.

seemingly performing badly from every other criterion, is competitive from the point of view of Brier score, a result for which we have no ready explanation.

Table 4.4: Nonlinear regression model of (4.1), Logistic errors, exceedances over threshold 119.9 Include quadratic term: NLLH=301.613

Parameter	Estimate	Stand. error	t Ratio
m_0	118.8	2.86	41.52
m_1	5.21	4.267	1.22
t_1	0.04127	0.1042	0.40
t_2	1.175	2.946	0.40
t_3	-0.01565	0.04044	-0.39
v_h	4.473	3.756	1.19
r	-0.03539	0.09017	-0.39
o_p	-0.002989	0.01427	-0.21
v	-0.1584	0.3854	-0.41
m_u	-0.1022	0.2649	-0.39
m_v	-0.44	1.068	-0.41
y_1	-0.001986	0.005009	-0.40
y_2	-0.1007	0.2473	-0.41
a_1	-0.1613	1.014	-0.16
b_1	0.2046	0.5507	0.37
a_2	0.03245	0.3326	0.10
b_2	0.04814	0.2094	0.23
$\log \sigma$	-1.361	2.414	-0.56

For station P, the NLLH values for the nonlinear probit and logit models, corresponding to Tables 4.2–4.4, were 110.407, 109.014 and 103.292. Since the model based on a linear trend does not depend explicitly on past values (such as PDAY), the appropriate comparison is with Table 3.1, where we had NLLH=114.553. Thus, the fit does appear to be better under the nonlinear models. Under the quadratic fit, the model improves even more. Note, however, that this is at the cost of introducing more parameters into the model. In the present case it appears possible to remove t_2 and t_3 without significantly worsening the fit, but no other parameters.

The predictive diagnostics again showed that the nonlinear model in which the parameters are not re-estimated significantly underpredicts the total number of exceedances: $e = 13.050$ against $r = 41$, resulting in a highly significant $z = 10.4$, whereas the nonlinear probit model had $e = 48.375$ which is not a significant discrepancy ($z = -1.35$). The calibration by year for the probit model resulted in Table 4.4, which is to be compared with Table 3.14 for the corresponding linear-logistic model:

Table 8.8: Exceedance probability parameters
Station R, Threshold 79.9

Variable	Estimate	Stand. error	t ratio
CONST	−25.73	4.964	−5.18
YEAR	−.07884	.06826	−1.16
CDAY	−.177	.5151	−.34
SDAY	.8467	.2346	3.61
T	.3789	.06297	6.02
RH	.07593	.03169	2.40
Q	−.5981	.1476	−4.05
WIND.U	.04774	.03499	1.36
WIND.V	−.156	.03975	−3.93
VIS	−.1117	.02276	−4.91

9. Conclusions and summary

In this section we give an overall summary of what the objectives of the study were, what methods were used, and what we think has been achieved as a result. The section can be read independently of the rest of the paper.

The study was conceived as a follow-on to the study by Bloomfield, Royle & Yang (1993), in which nonlinear regression techniques were used to predict ozone levels as a function of a suite of meteorological variables. Our selection of meteorological variables does not exactly correspond to theirs – we do not include lagged meteorological variables or upper-air data, and we base all results on 12:00 noon measurements rather than any kind of daily average – but the main variables are the same and achieve similar R^2 in the only direct comparison made between the two papers, when the nonlinear model of equation (4.1) is fitted to the network maxima by least squares.

For our study we have concentrated on ozone daily maxima from three stations labelled P, Q and R, chosen because they had high ozone levels, and also the network maxima as defined by Bloomfield *et al.* The overall objective of the study is to fit models that are based solely on exceedances of a high threshold, so as to answer questions relating to trends in the high levels of the ozone series. For much of the study, the threshold used was 119.9 (i.e. just below the standard level 120), though for parts of the study which used excess values as well (Section 5 onwards), it was often found convenient to use a threshold 99.9. At the lower threshold, there are more exceedances and this helps us fit a good model.

The simplest model of this kind is to fit a logit model to the binary data consisting of 1 if the threshold is exceeded on a given day, 0 otherwise. In Section 3, we fitted this model

Table 4.5: Calibration of probability forecasts by year
Station P, threshold 119.9, model of Table 4.2

Year	n	r	e	w	z
1981	268	2	5.071	2.681	−1.876
1982	280	1	2.366	2.002	−0.965
1983	307	7	9.245	4.868	−1.018
1984	329	4	2.293	1.997	1.208
1985	324	3	2.771	1.517	0.186
1986	324	2	2.523	1.959	−0.374
1987	323	6	4.732	3.358	0.692
1988	328	14	12.729	5.994	0.519
1989	331	1	2.332	1.977	−0.947
1990	317	0	0.894	0.873	−0.957
1991	261	1	3.420	2.763	−1.456

Looking at the z values in the last column, the overall pattern is similar to that of Table 3.14, but there are fewer significant values and $\sum z^2$ is smaller (11.8 as against 22.4 for Table 3.14). The Brier scores were 37.08 (nonlinear model without refitting), 31.53 (nonlinear probit model).

Again, a logit model was fitted with both linear and quadratic trends. For the logit-linear case, the overall fit has $e = 46.686$ ($z = -1.014$), and a calibration by year as follows:

Table 4.6: Calibration of probability forecasts by year
Station P, threshold 119.9, model of Table 4.3

Year	n	r	e	w	z
1981	268	2	2.400	1.718	-0.305
1982	280	1	2.021	1.854	-0.750
1983	307	7	10.690	4.963	-1.656
1984	329	4	2.496	2.115	1.034
1985	324	3	1.641	1.525	1.101
1986	324	2	2.022	1.856	-0.016
1987	323	6	3.144	2.741	1.725
1988	328	14	11.499	7.507	0.913
1989	331	1	4.089	2.828	-1.837
1990	317	0	1.044	1.019	-1.034
1991	261	1	5.642	3.317	-2.549

Table 8.6: Exceedance probability parameters
Station P, Threshold 79.9

Variable	Estimate	Stand. error	*t* ratio
CONST	−25.89	2.372	−10.92
YEAR	−.06021	.02917	−2.06
CDAY	−1.644	.4253	−3.87
SDAY	.01026	.2043	.05
OPCOV	−.008993	.003259	−2.76
PR	.03602	.02308	1.56
T	.3	.02319	12.94
TD	.1617	.05357	3.02
Q	−.5966	.1331	−4.48
WIND.U	.04669	.03084	1.51
WIND.V	.1414	.03447	4.10
VIS	−.09568	.01674	−5.72
T.WSPD	−.02015	.002457	−8.20

Table 8.7: Exceedance probability parameters
Station R, Threshold 59.9

Variable	Estimate	Stand. error	*t* ratio
CONST	−16.21	2.218	−7.31
YEAR	−.07686	.05045	−1.52
CDAY	−.3045	.2978	−1.02
SDAY	.6717	.1599	4.20
T	.2864	.03318	8.63
Q	−.2503	.04173	−6.00
VIS	−.1051	.01786	−5.88
WSPD	.1216	.07249	1.68
T.WSPD	−.01032	.004714	−2.19

In the case of the logit-quadratic model, we had $e = 49.361$ ($z = -1.507$), and the following table for calibration by year:

Table 4.7: Calibration of probability forecasts by year
Station P, threshold 119.9, model of Table 4.4

Year	n	r	e	w	z
1981	268	2	0.728	0.722	1.498
1982	280	1	1.466	1.401	-0.394
1983	307	7	10.345	5.099	-1.481
1984	329	4	3.002	2.444	0.639
1985	324	3	3.235	2.478	-0.150
1986	324	2	3.064	2.516	-0.671
1987	323	6	5.025	3.733	0.505
1988	328	14	15.899	7.076	-0.714
1989	331	1	3.158	1.978	-1.534
1990	317	0	0.901	0.888	-0.956
1991	261	1	2.539	2.456	-0.982

The Brier scores for the two logit models were 36.49 (linear trend model) and 32.54 (quadratic trend model). In this case, the linear trend model seems to pass all the tests except for the clear pattern of z values in Table 4.6, while the quadratic trend model seems fine from every point of view.

Figure 4.1 illustrates these results for the four main models we have applied to station P. Thus, we show the actual number of exceedances in each year (solid dot), and the predicted numbers under each of the models, in each case dropping the year being predicted from the data so that the prediction is based on the model refitted to the remaining years. The two logit models being fitted here omitted the terms t_2 and t_3. It can be seen that the original least squares fit significantly underpredicts most of the exceedances, but the other three fits all follow the actual counts reasonably well, with the one based on the quadratic model perhaps identifiable as the best. Figure 4.2 shows the same thing for the network maxima, with a less clear-cut distinction between the different fits but the quadratic model still doing best if we exclude the one year (1991) for which it performs badly. However, since 1991 is the most recent year in the data, we should be concerned about its failure to predict correctly in that year.

For station Q, we obtained NLLH=93.402 for the nonlinear probit model, which is to be compared with 99.317 in Table 3.3, but were unable to obtain a fit under the logit model. For station R, we have NLLH=66.184 (probit), 66.788 (logit) against 67.524 (Table 3.4). Thus the nonlinear model does appear to improve the fit for station Q, but not for

Table 8.4: u_0=80, Station R

Variable	Estimate	Stand. error	*t* ratio
CONST	−1.84	1.223	−1.50
YEAR	−.1212	.06906	−1.76
CDAY	−1.249	.5977	−2.09
SDAY	−.5489	.2709	−2.03
T	.09555	.01805	5.29
Q	−.1306	.04333	−3.02
WIND.U	.06009	.03706	1.62
WIND.V	−.07609	.03556	−2.14
VIS	−.05623	.01981	−2.84
ξ	.1531	.1921	.80
θ	1.018	.1232	8.27

Table 8.5: Exceedance probability parameters Station P, Threshold 59.9

Variable	Estimate	Stand. error	*t* ratio
CONST	−18.34	1.417	−12.95
YEAR	−.04824	.0223	−2.16
CDAY	−.9727	.2512	−3.87
SDAY	.7299	.1332	5.48
OPCOV	−.009412	.002484	−3.79
PR	.04982	.0162	3.07
T	.2499	.01799	13.89
TD	.07584	.03041	2.49
Q	−.3279	.08207	−4.00
WIND.U	−.01097	.02112	−.52
WIND.V	.1102	.02204	5.00
VIS	−.0896	.01331	−6.73
T.WSPD	−.01122	.002127	−5.27

station R. Because of the computational time required, we did not compute the predictive diagnostics for stations Q and R.

We can now attempt some interpretation of these results. The failure of the direct attempt to apply (4.3), with the parameter estimates from the nonlinear least squares fit, is disappointing but was anticipated in view of two facts made perfectly clear by Bloomfield *et al.* (1993): σ is not constant, and the normal distribution does not fit the extreme residuals. One could, in theory, extend the Bloomfield *et al.* analysis to model both of these features, but then the model runs the danger of becoming seriously overparametrized. At any rate, Bloomfield *et al.* themselves made no attempt to do this.

The nonlinear probit and logit models might be considered a more reasonable attempt to profit from the work of Bloomfield *et al.*, using it to establish the appropriate form of a binary data regression model. The results are mixed: for stations P and Q the evidence is that the nonlinear model really does improve things, whereas for station R and the network maxima it does not. The apparent improvement in the fit when we use a quadratic trend model, in the case of station P and the network maxima, should also be noted.

One point we have not emphasized, however, is that there are computational stability problems with these models that do not appear to be present with the models of section 3. We pointed out already that we were unable to obtain a fit for the nonlinear logistic model with station Q, and the same difficulty arose in one of the instances of station P, when we attempted to refit the nonlinear logit model (with either linear or quadratic trend) with the year 1982 omitted. For this reason, the 1982 projections for these two models in Tables 4.6, 4.7 and Figure 4.1 are based on the full data set.

Moreover, even in the cases when we were able to fit a model, it appears that small perturbations of the model fit resulted in substantial changes in the parameters. As an example of this, after obtaining Table 4.4, we tried rerunning the nonlinear optimization program with the previous parameter estimates as starting values. The reduction in NLLH when we did this was minimal (from 301.613 to 301.611) but the changes in individual parameters were non-trivial. Figure 4.3 shows the ratios of all 18 parameters under the two fits, and it can be seen that several differ substantially from 1. However, the forecast probabilities for the two fits are very similar: Figure 4.4 plots forecast probabilities of exceedance from one model against those from the other. The two sets are in very good agreement and in fact the maximum discrepancy is less than 0.002. Thus there appear to be instabilities in the model which do not affect the final conclusions. One consequence of this should be noted, however: although standard errors have been quoted in this section just as they have in other sections of the paper, they should not be trusted! We might also add that the computational time required for these models was substantially larger than for those of Section 3.

It is difficult to reach any definitive conclusions from these results, but in comparing the models of this section with those of section 3, we can note that the models of this section do indeed appear to be slightly better when exploited to their fullest extent, but there is

Table 8.2: u_0=80, Station P

Variable	Estimate	Stand. error	t ratio
CONST	−14.77	2.957	−4.99
YEAR	−.09933	.02954	−3.36
CDAY	−.749	.4882	−1.53
SDAY	.1772	.2132	.83
PR	.03685	.02101	1.75
T	.2131	.03049	6.99
TD	.05107	.05694	.90
Q	−.1762	.133	−1.33
WSPD	.5068	.1681	3.02
T.WSPD	−.02725	.006264	−4.35
ξ	−.04198	.1634	−.26
θ	.8169	.08419	9.70

Table 8.3: u_0=60, Station R

Variable	Estimate	Stand. error	t ratio
CONST	−2.648	1.009	−2.62
YEAR	−.05715	.03077	−1.86
CDAY	−.4452	.2597	−1.72
SDAY	.1314	.1178	1.12
T	.09399	.009624	9.77
TD	.05699	.0258	2.21
Q	−.2594	.06871	−3.78
WIND.U	.03126	.01559	2.01
WIND.V	−.05123	.01589	−3.22
VIS	−.03928	.01057	−3.72
ξ	−.04785	.08586	−.56
θ	1.039	.07002	14.84

not a substantial difference, and the associated computational and stability difficulties might well lead one in practice to prefer the approach of Section 3.

Of course, there remain many points of contact between the two approaches, especially concerning the identification of suitable covariates — as will already be apparent from the discussion in Sections 2 and 3, we have made extensive use of the results of Bloomfield *et al.* in this respect. Where the comparison is less certain is in trying to use their functional form directly to model threshold crossings.

5. Excesses over a threshold

So far, the analysis has been concerned solely with the probability of crossing a high threshold. We now extend the analysis so that the *excess*, or the amount by which the ozone level exceeds the threshold when it crossed, is also modeled. This is important for a number of reasons, all connected with the desirability of calculating probabilities of crossing levels higher than the ones to which the methods of Sections 3 and 4 would be applicable. For example, even if there are still too many exceedances of the official ozone standard of 120 ppb, it is possible that the excesses are getting smaller so that we can say that, from the point of view of its effect on human health, the ozone situation is improving. A further reason for considering excesses is to allow us to perform comparisons based on the maximum levels achieved in different years (Sections 6 and 7). The method developed here provides a possible means of extrapolating to higher thresholds than could be justified in a direct analysis. At the same time, it is important to be aware that an extrapolation is involved in this, and it would be unwise to attempt to extrapolate too far. The discussion that follows will, it is hoped, provide some insight into how far it is reasonable to extrapolate.

To motivate the development of models for excesses over a threshold, consider first the simple case in which the underlying observations, $X_1, ..., X_n$ say, are independent with common distribution function F. The conditional probability that $Y_i = X_i - u < y$, given $Y_i > 0$, is represented by

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)}. \tag{5.1}$$

The question is what parametric form would be appropriate for F_u. In this connection, Pickands (1975) introduced the *generalized Pareto distribution* (henceforth GPD) whose distribution function is given by

$$G(y; \sigma, \xi) = 1 - \left(1 + \xi\frac{y}{\sigma}\right)_+^{-1/\xi}, \quad y > 0, \tag{5.2}$$

where $\sigma > 0$, ξ is any real number, and $x_+ = \max(x, 0)$. Thus the range of y is $0 < y < \infty$ for $\xi \geq 0$ and $0 < y < -\sigma/\xi$ if $\xi < 0$. The exponential distribution, $1 - e^{-y/\sigma}$, appears as a limiting case when $\xi \to 0$. This is of historical importance, because most early attempts

Probability plots are shown for all four data sets and all four models in Figures 8.1–8.4. In Figure 8.1, there is no discernable difference among the models but there are three apparent outliers, corresponding to (from the top) July 12 1982, July 22 1983 and June 19 1983. From Table 6.6 we see that the first and third of these were also events corresponding to a very low tail probability with respect to their daily maxima, while the July 22 1983 event also appears in both Tables 6.4 and 6.6. Thus it seems, as we might expect, that there is a high association between extreme events from the daily maximum analysis and ones from the present results. On the other hand there are no obvious outliers in the other plots, except in Figure 8.4 where it appears that there is one outlier for the date of June 19 1986. No attempt has been made to remove outliers in this analysis.

To complete the description of sums over a high threshold, we also need a model for the probability of crossing the threshold. Tables 5.5–5.8 present models that have been identified for that, using the same methods as in section 3 but repeating the variable identification with the lower thresholds.

Table 8.1: u_0=60, Station P

Variable	Estimate	Stand. error	*t* ratio
CONST	−4.66	1.537	−3.03
YEAR	−.05252	.01584	−3.32
CDAY	−.8205	.1987	−4.13
SDAY	.06538	.09482	.69
PR	.02687	.01149	2.34
T	.1094	.01648	6.64
TD	.04833	.02229	2.17
Q	−.2004	.05481	−3.66
VIS	−.01439	.008711	−1.65
WSPD	−.05544	.06709	−.83
T.WSPD	−.004823	.002938	−1.64
ξ	−.0318	.07157	−.44
θ	.9293	.05059	18.37

at this kind of analysis (e.g. in connection with Dutch sea level modeling after 1953) were based on the exponential distribution.

The theoretical justification for the GPD is based on a theorem proved by Pickands (1975), which may be interpreted as saying that the GPD occurs as a limiting distribution of $F_u(\cdot)$, for high u, if and only if the distribution F is in the domain of attraction of one of the classical extreme value distributions described for example in the books of Gumbel (1958) or Leadbetter *et al.* (1983). Since the "domain of attraction" condition is ubiquitous in extreme value analysis, this implies that the GPD is a natural choice for exceesses over high thresholds in any context where one might try to apply extreme value theory. Estimation of the parameters σ and ξ can be carried out by the method of maximum likelihood, which has regular properties whenever $\xi > -\frac{1}{2}$, a condition which is almost always met in environmental applications.

In the present context, of course, we want to extend this to include covariates in the analysis. A general framework for this was laid out by Davison & Smith (1990). They considered regression models in which the excess (if there is one) on day i, say, is representated by the $G(\cdot;\sigma_i,\xi_i)$ with σ_i and ξ_i depending on covariates. In practice it is usually adequate (and a lot simpler) to assume ξ constant, while the interpretation of σ_i as a scale parameter suggests naturally that a logarithmic link function would be appropriate. Thus we are lead to consider models of the form

$$\log\sigma_i = \sum_j x_{ij}\gamma_j, \quad \xi_i = \xi \tag{5.3}$$

in terms of new coefficients $\{\gamma_j,\ j = 1,2,...\}$. There is no reason why the significant covariates should be the same as in the binary analysis of Section 3, so in general we would expect to repeat the variable-selection procedure based on the excesses over the threshold. For the time being, we assume the daily values are independent.

A more general model, whose significance will become clear later on, is to extend the GPD to include a power-law transformation of Y_i. That is, we assume Y_i^θ, for some $\theta > 0$, follows the GPD, so the distribution function of the excess Y_i is given by

$$H(y_i;\sigma_i,\xi,\theta) = 1 - \left\{1 + \xi\left(\frac{y_i}{\sigma_i}\right)^\theta\right\}_+^{-1/\xi}, \quad y > 0, \tag{5.4}$$

To give this a specific name we call it the TGPD (transformed GPD).

We now consider the application of these models to the three stations we have been considering, as well as the network maxima. In the case of the three stations P, Q and R, it was found more satisfactory to use a lower threshold, so we present most of our results with respect to the threshold 99.9 rather than 119.9. Again, we did not use 100 or 120 as a threshold so as to avoid the difficulties caused by having observations exactly

8. Sums of excesses over a threshold

A different characterization of the "extremeness" of an ozone day, other than the daily maximum, is some form of criterion based on the persistence of hourly ozone readings over a lower threshold than the ozone standard 120. For example, the sum of exceedances over the threshold 60 is used in calculating the SUMO6 criterion. In general, we may define a threshold u_0 (e.g. $u_0 = 60$) and consider either of the criteria

$$S_i = \sum_k (Y_{ik} - u_0)_+ \tag{8.1}$$

or

$$S_i = \sum_k Y_{ik} I(Y_{ik} \geq u_0), \tag{8.2}$$

where Y_{ik} is the k'th hourly value on day i. In (8.1) we use the notation, as previously, $x_+ = \max(x, 0)$, and in (8.2), $I(Y_{ik} \geq u_0)$ denotes the indicator function (1 if $Y_{ik} \geq u_0$, 0 otherwise).

Note: This is the first time in the whole report that we consider hourly ozone values. Up to now, all the analysis has been in terms of daily maxima.

In this analysis we confine ourselves to (8.1), whose distribution is easier to characterize than (8.2). However, current interest in SUMO6 is focussed on (8.2), so it would seem to be important in future work to try to characterize that distribution as well.

The current status of work in extreme value theory is that it has a good deal less to say about quantities such as (8.1) than it does about overall maxima. Nevertheless, a recent paper by Anderson & Dancy (1992) has provided a theoretical characterization and some practical results. In particular, they proposed the TGPD (equation 5.4) as a distributional form that was consistent with the theoretical characterization and performed well in a practical (hydrological) data study. Although their result should not be considered a complete treatment of the problem, it does suggest that an analysis based on the TGPD would have both theoretical and practical justification, and we therefore pursue that here.

The possibility has also been suggested of trying other values of u_0 besides 60. In particular, one could define a similar criterion based on $u_0 = 80$, so that is tried as well.

In Tables 8.1–8.4, the results of the TGPD model are given for $u_0 = 60$ and 80, for stations P and R. In both cases the full model (including both ξ and θ) is quoted, though in no case are both of these parameters actually required. In all four cases, if θ is included in the model, then a test of $H_0 : \xi = 0$ produces a result which is not significant. In fact, only in the case of station P and $u_0 = 80$ is the null hypothesis of an exponential distribution ($\xi = 0,\ \theta = 1$) rejected by a likelihood ratio test. In all cases the other parameters were chosen by a variable selection procedure similar to that used in section 3, though assuming a linear trend in the YEAR variable.

on the threshold. One reason for using the lower threshold was that there are many more exceedances in the data set – 94, 84 and 55 respectively for stations P, Q and R, as compared with 41, 38 and 23 for threshold 119.9 – and this improves our ability to fit and compare different models. For the network maxima, there are plenty of exceedances of the higher threshold – 143 in all – so we continued to use threshold 119.9 for that data set.

We now present specific results, starting with station P. All the analyses have used YEAR as a covariate, and a process of variable selection, similar to that carried out in Section 3, showed that T, WSPD and T.WSPD were also significant variables. Here are the results for the model defined by equations (5.2) and (5.3):

Table 5.1: Station P: Excesses over threshold 99.9

Variable	Estimate	Stand. error	*t* ratio
CONST	−19.03	5.293	−3.60
YEAR	−.09418	.0331	−2.85
T	.2631	.06151	4.28
WSPD	1.06	.3693	2.87
T.WSPD	−.03967	.01468	−2.70
ξ	−.2846	.138	−2.06

Here NLLH=365.376. If YEAR is omitted, this rises to 368.904. Thus, the evidence in a downward trend in the excesses, and not just in the probability of crossing the threshold, seems clear from this.

As an indication of the effect of ξ and θ, with the other covariates remaining the same, the following results were obtained:

Table 5.2: Station P, threshold 119.9
Comparison of different models for ξ and θ

$\hat{\xi}$	Stand. err.	$\hat{\theta}$	Stand. err.	NLLH
—	—	—	—	367.016
−.285	.137	—	—	365.376
—	—	.930	.080	366.650

No fit was obtained for the TGPD model including both parameters ξ and θ. With this model, the iterative numerical routine moved into the region $\xi < -1$ for which no defined as the probability of crossing level t on day n, given the meteorology vector x_n for that day. We can therefore define an expected number of crossings for year N by the formula

$$S_N(t) = \sum_{\text{day } n \in \text{ year } N} p_n(t|x_n).$$

This quantity can be calculated with respect to each year N for which the meteorological data are available, and can be thought of as the "ozone potential" for that year.

For an analysis of this nature, it would not make sense to include trends in the model, so we use only models for which no trend is included. Also, since the model is not being applied directly to the ozone record, we cannot include PDAY in the model. However, we can calculate the function $S_N(t)$ with respect to different ozone models that have been fitted to different stations.

For our main evaluation, we used the models refitted to the network maxima data as in Tables 3.11 and 5.5, omitting YEAR, YEAR2 and PDAY from the models. The function $S_N(t)$ was then plotted against N, for each of t =120, 140, 160, 180. The results are shown in Figure 7.1. The figure shows substantial fluctuations, with high projected ozone levels for 1966, 1975, 1983 and 1988. Overall 1988 has the highest projected exceedances, though at the higher levels (160, 180), 1983 is higher. This is consistent with what we have seen in earlier comparisons. A similar figure computed using model fits for station P (Figure 7.2) gives a very similar message, though in this case with 1988 giving the highest number of projected exceedances at all levels.

These models were based on independent days. However, the calculations were repeated using the Markov models of Section 5, with results indistinguishable from those in Figures 7.1 and 7.2. There is a distinction between using the Markov model in this context and for the predictive diagnostics of Section 5, because there, the diagnostics were computed sequentially (one day's ozone reading affects the diagnostic for the next). Our analysis in this section is based purely on expected numbers of exceedances, so it is only to be anticipated that the dependence in the model has minimal influence on the conclusions.

There are a number of possible interpretations of the results. There appears to be a general increase of ozone potential over the whole period of the study, which implies climatological variability. This could therefore be tied in with much broader issues of climate change, including the greenhouse effect and its possible effect on global warming. The results do lend some support to the contention that 1988 was the most extreme year in terms of meteorology in the record. On the other hand, we have seen in Section 5 that the models still underpredict the actual ozone levels that occurred in 1988, so there is still some support for the notion that additional factors were responsible for the exceptionally high levels of that year.

local maximum of the GPD likelihood exists (Smith 1985). Although there are theoretical procedures for obtaining model estimates in the situation (Smith 1993), past experience with these kinds of models has suggested that the phenomenon is more likely to be due to having too small a sample to fit the models adequately. Indeed, with the original threshold of 119.9, this happened even with the GPD model (without the transformation parameter θ), and with several of the data sets. Since a probability plot (Figure 5.1) suggests that the GPD model fits perfectly well, for the rest of the analysis we adopt that model and do not attempt to resolve the difficulty concerning the TGPD model.

To construct Figure 5.1, we defined a "residual" $Y_i/\widehat{\sigma}_i$ with respect to the i'th excess Y_i and the associated estimated scale parameter $\widehat{\sigma}_i$. The residuals were then arranged in order and plotted against expected order statistics under each of the four models. In the case of the TGPD model, although not finding a maximum likelihood fit, we did plot the residuals corresponding to the best model found (with $\xi \approx -1$). If the model fits the data, then the residuals should be tightly clustered around a straight line of unit slope through the origin (also shown on the plots). The results show clear deviation from this in the upper tail of the exponential and transformed exponential data sets, but a good fit for the GPD. In the light of this, we believe that the GPD forms a good fit overall.

For the other stations, in all cases the same model seems to produce a good fit. Tables 5.3–5.5 show GPD model fits for stations Q, R and the newtork maxima, and Figures 5.2–5.4 are the corresponding probability plots for residuals under each of the four models for the marginal distribution.

Table 5.3: Station Q: Excesses over threshold 99.9

Variable	Estimate	Stand. error	t ratio
CONST	−8.511	5.414	−1.57
YEAR	−.09546	.02657	−3.59
T	.1502	.06346	2.37
WSPD	.1658	.3652	.45
T.WSPD	−.01252	.01424	−.88
ξ	−.3603	.09193	−3.92

Table 6.6: Highest ozone days by joint criterion
All days with Ozone $\geq$ 120 and $\eta_i < 0.1$ are tabulated
Station P

Date	Ozone	Probability η_i
June 20 1983	122	0.0012
June 6 1988	164	0.0022
July 12 1982	144	0.0026
August 8 1985	133	0.0048
September 2 1983	141	0.0049
July 27 1985	128	0.0068
May 15 1991	123	0.0101
August 1 1981	151	0.0102
June 20 1987	124	0.0123
August 11 1988	149	0.0143
June 23 1983	161	0.0169
July 22 1983	158	0.0171
August 4 1984	172	0.0195
July 28 1986	123	0.0210
August 3 1984	145	0.0258
July 7 1988	215	0.0292
August 7 1988	144	0.0312
August 9 1986	117	0.0327
July 24 1988	144	0.0368
June 8 1985	143	0.0413
August 15 1984	126	0.0433
August 3 1988	160	0.0473
July 13 1984	136	0.0576
July 8 1988	215	0.0607
June 19 1987	133	0.0736

7. Recurrence of 1988

One question that has repeatedly occurred in discussion of the ozone problem is "how extreme a year was 1988?". Of course we could ask the same question with respect to any other year in the study, but 1988 is the focus because it is clearly the worst ozone year in the current record. In this section we attempt to answer this question with respect to historical climatological data going back to 1959.

Each of the models we have considered can be regarded as providing an algebraic expression for

$$p_n(t|x_n)$$

Table 5.4: Station R: Excesses over threshold 99.9

Variable	Estimate	Stand. error	*t* ratio
CONST	−6.911	3.654	−1.89
YEAR	−.2613	.1245	−2.10
T	.1424	.04434	3.21
WSPD	.4443	.3097	1.44
T.WSPD	−.01896	.008808	−2.15
ξ	−.4019	.1718	−2.34

Table 5.5: Network maxima: Excesses over threshold 119.9

Variable	Estimate	Stand. error	*t* ratio
CONST	−.8527	1.685	−.51
YEAR	−.04	.03098	−1.29
T	.05795	.02026	2.86
WSPD	−.1005	.1307	−.77
T.WSPD	−.004424	.005415	−.82
ξ	−.2333	.08884	−2.63

In all four cases, the GPD model was a significant improvement on the exponential model, and better than the transformed exponential model. Results for the TGPD were somewhat mixed: in the case of station Q, the TGPD model was fitted successfully, but with a negligible improvement in log likelihood (0.003) over the GPD model. For station R, we encountered the same difficulty as with station P, no fit being obtained. For the network maxima, the TGPD was successfully fitted and did provide a significant improvement in log likelihood over the GPD (3.363). On the other hand, the probability plot in Figure 5.4 still shows an adequate fit for the GPD, and in view of the difficulties in fitting the TGPD in the other cases, it was decided to adopt the GPD as the principal model for this case also. As a comparison, here is the detailed table of results for the TGPD model:

Table 6.4: 11 highest ozone days and associated probabilities
Station P

Date	Ozone	Prob. 1	Prob. 2
July 7 1988	215	0.0292	0.0784
July 8 1988	215	0.0607	0.1192
July 6 1988	186	0.3230	0.4052
August 4 1984	172	0.0195	0.0129
July 5 1988	170	0.1568	0.1789
June 6 1988	164	0.0022	0.0104
June 23 1983	161	0.0169	0.0078
August 3 1988	160	0.0473	0.0676
July 22 1983	158	0.0171	0.0060
July 29 1983	157	0.1712	0.0567
August 1 1981	151	0.0102	0.0016

Table 6.5: Annual highest ozone days and associated probabilities
Station P

Date	Ozone	Prob. 1	Prob. 2
August 1 1981	151	0.0102	0.0016
July 12 1982	144	0.0026	0.0020
June 23 1983	161	0.0169	0.0078
August 4 1984	172	0.0195	0.0129
June 8 1985	143	0.0413	0.0319
July 28 1986	123	0.0210	0.0198
June 18 1987	149	0.1150	0.1176
June 6 1988	164	0.0022	0.0104
July 6 1989	121	0.3644	0.4480
September 10 1990	107	0.0311	0.0573
May 15 1991	123	0.0101	0.0273

Table 5.6: Network maxima: Excesses over threshold 119.9
TGPD model

Variable	Estimate	Stand. error	*t* ratio
CONST	.2615	1.226	.21
YEAR	-.002885	.05484	-.05
T	.04926	.01385	3.56
WSPD	-.1785	.09756	-1.83
T.WSPD	-.004063	.004196	-.97
ξ	-.5786	.2435	-2.38
θ	.7222	.1252	5.77

In three of the four cases, there appears to be a significant improvement as a result of including YEAR in the model. Using the GPD model as the standard, the difference in log likelihood when YEAR is omitted are 3.528 (station P), 5.107 (Q), 2.115 (R) and 0.860 (network maxima). Only the last of these is not clearly significant. The result is especially interesting for station Q, since in that case our earlier analyses based on exceedances of a single threshold have cast doubt on whether the trend is significant. The present analysis shows that, for high enough ozone levels, there is a significant downward trend in that case also.

One additional comment should be made concerning the GPD. Earlier analyses of ozone data, including those of Smith (1989), have suggested that the high-level exceedances of ozone have an approximately exponential tail. The present analysis contradicts that, since in all four cases the GPD is a significant improvement on the exponential tail. This is clear from the plots in Figures 5.1–5.4, and from the differences in log likelihoods (1.640, 4.420, 2.157 and 2.829 respectively for stations P, Q, R and the network maxima). As a comparison, we reran all the models without any covariates, focussing just on the comparison of the exponential and GPD models, obtaining log likelihood differences of 0.574, 1.041, 0.000 and 0.122. None of these are significant, so we would accept an exponential tail of the distribution if we ignored the covariates. Thus, our conclusion is that the form of tail function adopted is dependent on whether covariates are included in the model, but in the case of interest to us, where the covariates are included, the exponential is too long a tail and the GPD with $\xi < 0$ is a significantly better fit.

Table 6.3: Highest ozone days by joint criterion
All days with Ozone $\geq$ 130 and $\eta_i < 0.05$ are tabulated
Network maxima

Date	Ozone	Probability η_i
May 8 1982	170	0.0002
June 30 1981	136	0.0008
July 7 1988	223	0.0012
August 22 1990	130	0.0015
June 6 1988	164	0.0048
July 23 1983	162	0.0050
August 15 1982	135	0.0051
July 8 1988	215	0.0055
August 17 1986	131	0.0063
August 11 1988	173	0.0063
July 19 1991	135	0.0078
May 28 1982	132	0.0087
June 23 1984	160	0.0088
April 18 1987	139	0.0101
August 1 1981	172	0.0116
May 25 1985	146	0.0126
June 18 1987	178	0.0194
August 7 1988	144	0.0201
June 21 1991	134	0.0205
June 28 1984	130	0.0212
August 18 1984	151	0.0233
July 7 1985	130	0.0236
August 16 1982	147	0.0311
June 24 1987	177	0.0312
August 14 1991	134	0.0348
July 3 1982	134	0.0356
July 28 1986	142	0.0356
July 12 1982	144	0.0366
August 15 1984	136	0.0420
June 19 1986	131	0.0444
August 29 1983	155	0.0450

Predictive assessment of model fit

Now that we have a model for both threshold crossing probabilities and the distribution of excesses over a threshold, we can compute crossing probabilities and expected numbers of exceedances with respect to any level larger than the original threshold. This makes it possible to repeat the predictive diagnostics which were first used in Section 3, with respect to higher ozone levels.

As an example, here is what happens when the models of Tables 3.11 (quadratic trend for exceedance probabilities) and 5.5 (linear trend for the distribution of excesses), both of which were calculated with to respect to the threshold 119.9, are combined and used to model exceedances of the level 139.9. At this level, the expected numbers of exceedances are still large enough for standard χ^2 tests to be applicable.

The overall calibration table (compare Table 3.13) is as follows:

Table 5.7: Calibration of probability forecasts
Network maxima, level 139.9, models with trend

p_{min}	p_{max}	n	r	e	w	z
0.000	0.050	2123	10	8.004	7.813	0.714
0.050	0.100	84	5	6.108	5.647	−0.466
0.100	0.300	106	22	18.572	14.974	0.886
0.300	0.500	30	10	10.947	6.864	−0.362
0.500	1.000	11	8	6.663	2.537	0.839
0.000	1.000	2354	55	50.295	37.834	0.765

Here, based on the first five rows, $\sum z^2 = 2.35$, clearly not significant as a χ^2_5 variable. The final z value (0.765) represents the overall fit, while the Brier score is 40.42 with a corresponding z (Y_n^B) value of 0.61. All of these point to the model as being fully satisfactory in terms of its overall calibration.

The calibrations by year, corresponding to Table 3.14, are as follows:

to be a sensible thing to do on its own, however, for two reasons. The first is that it is quite possible for a very small η_i to arise on a day when the ozone is not at all high in absolute terms, if it is merely a lot higher than would have been expected from the background conditions. Identifying such days would appear to be of little use in dealing with the overall ozone problem. The second reason for not giving too much attention to the very small values of η_i is that they are likely to be very sensitive to arbitrary features of the model, such as which threshold was used or which covariates were included in the models for p_i and σ_i.

However, another possibility is to adopt a joint criterion: select some intermediate ozone level, say 130, and some intermediate critical value of η_i, say .05, and flag all days where both criteria are violated. The idea is that this will help identify days where the ozone was high but, in a sense, it should not have been. The results of this are shown in Table 6.3, arranged in order of increasing η_i.

Similar results for station P are presented in Tables 6.4–6.6. In this case, for the models with trend, a linear trend in both the exceedance probability and excess value components was adopted. In Table 6.4, it can be seen that four of the top five values are from July 5–8 1988, and that despite the fact that we are using the dependent model, so that the quoted probabilities take previous high values into account, the η_i values for July 7 and 8 are still very small, whereas those for July 5 and 6 are much more moderate. The conclusion from this is that the meteorology through our models will "explain" the values for July 5 and 6, which are still very high compared with the rest of the record, but that it still cannot explain the exceptionally high ozone readings of July 7 and 8. Table 6.5 again presents the annual maxima, and Table 6.6 lists all the days for which the ozone level was at least 120 and $\eta_i < 0.1$.

The results are necessarily highly tentative. As is already clear from comparing the results for station P and the network maxima, the results are highly volatile to the selection of model, and the interpretation of a small value of η_i is far from clear. Nevertheless, it seems valuable to try to identify those days on which a high ozone level occured which could not be explained by meteorology, as this gives at least indirect evidence that additional factors may have been responsible for the high ozone level on those days, which could in turn point the way towards improved control strategies for the future.

Table 5.8: Calibration of probability forecasts by year
Network maxima, level 139.9, models with trend

Year	n	r	e	w	z
1981	214	4	4.346	3.430	−0.187
1982	214	4	3.873	3.273	0.070
1983	214	13	14.527	8.803	−0.515
1984	214	8	4.151	3.489	2.061
1985	214	3	2.482	2.191	0.350
1986	214	1	2.749	2.478	−1.111
1987	214	7	7.002	5.212	−0.001
1988	214	13	6.104	4.742	3.167
1989	214	0	3.357	2.707	−2.040
1990	214	0	0.532	0.513	−0.742
1991	214	2	1.172	0.996	0.830

Here there do appear to be some discrepancies, especially in 1988 where the model underpredicts the true rate of exceedance, and to a lesser extent in 1984 and 1989. We have $\sum z^2 = 21.34$, which is significant as a χ^2_{11} variable (5% point 19.68, 2.5% point 21.92), though this is somewhat questionable as several of the expected values in the e column are less than 5. If we group the years together as $\{1981,\ 1982\}$, 1983, $\{1984, 1985, 1986\}$, 1987, 1988, $\{1989, 1990, 1991\}$ then we obtain the following table:

Table 5.9: Calibration of forecasts by year (regrouped)
Network maxima, level 139.9, models with trend

Year	n	r	e	w	z
1981–2	428	8	8.219	6.703	−0.085
1983	214	13	14.527	8.803	−0.515
1984–6	642	12	9.382	8.158	0.917
1987	214	7	7.002	5.212	−0.001
1988	214	13	6.104	4.742	3.167
1989–91	642	2	5.061	4.216	−1.491

for $\sum z^2 = 13.36$, which is significant against χ^2_6 at the 5% level but not at the 2.5% level.

On the other hand, if we fit the same models without any trend parameters (YEAR and YEAR2) and repeat the same calculations, we obtain the following results. For the overall calibration we have:

Table 6.1: 14 highest ozone days and associated probabilities
Network maxima

Date	Ozone	Prob. 1	Prob. 2
July 7 1988	223	0.0012	0.0019
July 8 1988	215	0.0055	0.0071
June 23 1983	188	0.0533	0.0239
July 6 1988	186	0.1387	0.1573
July 15 1983	180	0.3319	0.2919
June 18 1987	178	0.0194	0.0170
June 24 1987	177	0.0312	0.0345
July 7 1981	176	0.1036	0.0485
July 19 1983	175	0.3272	0.2875
August 11 1988	173	0.0063	0.0104
August 1 1981	172	0.0116	0.0062
August 4 1984	172	0.0714	0.0422
July 5 1988	170	0.0685	0.0679
May 8 1982	170	0.0002	0.0002

Table 6.2: Annual highest ozone days and associated probabilities
Network maxima

Date	Ozone	Prob. 1	Prob. 2
July 7 1981	176	0.1036	0.0485
May 8 1982	170	0.0002	0.0002
June 23 1983	188	0.0533	0.0239
August 4 1984	172	0.0714	0.0422
June 7 1985	148	0.0699	0.0485
June 8 1985	148	0.0844	0.0592
July 28 1986	142	0.0356	0.0295
June 18 1987	178	0.0194	0.0170
July 7 1988	223	0.0012	0.0019
July 6 1989	139	0.4565	0.5130
August 22 1990	130	0.0015	0.0080
June 1 1991	152	0.0701	0.2589
June 20 1991	152	0.2049	0.4038

An alternative approach to identifying high ozone days might be to identify any day with an exceptionally small value of η_i as an extreme ozone event. This does not appear

Table 5.10: Calibration of probability forecasts
Network maxima, level 139.9, models without trend

p_{min}	p_{max}	n	r	e	w	z
0.000	0.050	2098	8	8.562	8.378	−0.194
0.050	0.100	110	11	7.942	7.348	1.128
0.100	0.300	108	16	18.760	15.185	−0.708
0.300	0.500	28	15	10.739	6.545	1.666
0.500	1.000	10	5	6.137	2.311	−0.748
0.000	1.000	2354	55	52.141	39.767	0.453

Here $\sum z^2 = 5.15$, the Brier score is 41.47, and the z (Y_n^B) value corresponding to that is 0.38, none of them significant. The calibration by year is as follows:

Table 5.11: Calibration of probability forecasts by year
Network maxima, level 139.9, models without trend

Year	n	r	e	w	z
1981	214	4	3.006	2.498	0.629
1982	214	4	2.831	2.503	0.739
1983	214	13	10.134	6.618	1.114
1984	214	8	2.875	2.528	3.224
1985	214	3	1.895	1.721	0.842
1986	214	1	2.179	2.014	−0.831
1987	214	7	6.942	5.208	0.025
1988	214	13	8.522	6.368	1.775
1989	214	0	4.711	3.724	−2.441
1990	214	0	1.763	1.621	−1.385
1991	214	2	7.283	4.964	−2.371

with $\sum z^2 = 30.62$ if the data are grouped as shown, or $\sum z^2 = 22.81$ if the data are grouped in the same way as Table 5.9. Clearly, these values are much more highly significant than for the fit with the YEAR and YEAR2 values. Even though the fit for 1988 is better, the overall calibration shows a clear downward pattern in the z values.

Similar, but mixed, results were obtained for stations P, Q and R. In every case, the predictive fit of the model becomes questionable at higher levels when assessed on a "calibration by year" basis, and it is not always the case that the model including a trend dominates over the one that does not. As a graphical comparison, Figures 5.5–5.8 show

6. Identifying the extreme ozone days adjusted for meteorology

Here, some preliminary thoughts are given as to how the preceding results could be interpreted as a means of identifying extreme ozone days when allowance is made for meteorology.

Suppose we have a model which defines the probability $p_i(x)$ that the ozone on day i is over a level x. For the models in this paper, this function is defined for any x above the initial threshold, by a combination of the models defined in Sections 3 and 5, and in the case of a dependent model as at the end of Section 5, it is to be interpreted as a conditional probability given past ozone values. If there is indeed an exceedance of the threshold on that day, with daily maximum at level Y_i, then the quantity $\eta_i = p_i(Y_i)$ can be thought of as the tail probability associated with the value Y_i. A very small value of η_i thus indicates some additional factor that must have produced extreme ozone on that day.

This η_i has a rather different interpretation according to whether or not trend is included in the model. If trend is not included, then it represents purely an adjustment for meteorology, not allowing for any other factors that may have affected ozone levels. If trend is included then, at least within the confines of the current study, it represents the best available indicator of how extreme a particular day is. The qualification "within the confines of the current study" is added because, of course, this is one aspect which could be changed considerably if additional information, such as data on precursor emissions, were included.

Table 6.1 lists all ozone days over 170 for the network maxima, with associated tail probabilities calculated both with a trend in the model (probability 1) and without (probability 2). The model with trend was the same as in Section 5 for threshold 119.9: quadratic trend in the binomial exceedance probabilities, linear trend in the GPD for excesses, with the dependence parameter α also included. Thus, the probability η_i is actually a conditional probability given the previous day's ozone. The model without trend is the same model refitted without the YEAR and YEAR2 covariates.

It can be seen that there are considerable differences among the probabilities associated with the highest ozone days. The table includes several values from 1988 with very small η_i values, reinforcing the fact that this year's very high ozone readings cannot be exaplined solely in terms of meteorology. In contrast, the 1983 values are associated with much more modest values of η_i.

Table 6.2 is similar to Table 6.1 but computed for, not the overall highest ozone days, but the highest within each year (including some ties). The effect of the trend on the tail probabilities can be more clearly seen in this table.

actual versus predicted numbers of exceedances for several ozone levels. In each case the models were based on those of Section 3 and the present section, with different assumptions about trend. In all cases, the predicted numbers follow the general pattern well, but the results for stations P, R and the network maxima clearly fail to reflect the very high values for 1988, while for station Q, 1988 was not a particularly extreme year.

Some attempt has also been made to do the same thing using nonlinear models, as in Section 4, for the probability of crossing the initial threshold, but this does not appear to improve the results. The difficulty is that the nonlinear models we have tried are still based on a constant scale parameter, so they cannot reflect relative changes in crossing probabilities as the level gets higher.

Allowing for dependence

The preceding analysis has allowed for serial dependence in a crude way, in allowing the variable PDAY as a covariate in the probability of crossing the initial threshold. However, this does not adequately reflect the dependence that may occur at higher thresholds. In this section we outline an alternative approach based on a first-order Markov structure incorporated into a threshold analysis. A recent review paper by Smith (1993) contains necessary background material.

Suppose we observe data $\{X_1,...,X_n\}$ from a first-order discrete-time Markov chain, i.e. the conditional distribution of each X_i given $X_1,...,X_{i-1}$ depends just on X_{i-1} through a conditional density $f(X_i|X_{i-1})$; here the variables may be discrete or continuous, the density referring to a transition probability in the discrete case or a conditional p.d.f. in the continuous case. We also assume that the process is stationary and that its stationary distribution is represented by a density f_1 say. Then the joint distribution of $X_1,...,X_n$ is

$$f_1(X_1)\prod_{i=2}^{n} f(X_i|X_{i-1}) \tag{5.5}$$

and this may be used to define a likelihood for the data.

A more convenient way to write (5.5) is in the form

$$L = \frac{\prod_{i=2}^{n} f_2(X_{i-1},X_i)}{\prod_{i=2}^{n-1} f_1(X_i)} \tag{5.6}$$

where $f_2(\cdot,\cdot)$ denotes the joint density of two consecutive values of the series. We write out equations (5.5) and (5.6) merely to make explicit an obvious point: that for a stationary first-order Markov chain, to define the joint likelihood of the data, it is necessary to specify the first- and second-order marginal densities of the process.

The conclusion at this point is that the dependent analysis *does* appear to deal with the extrapolation to higher thresholds, in the sense that any discrepancies between predictions and data are consistent with normal statistical variability. To extend the comparison, for example, at level 159.9, the calibration by year yields $\sum z^2 = 10.97$, not significant as χ^2_{11}, or if we group the years in the same way as in Table 5.16, $\sum z^2 = 8.03$, still nowhere near significant against χ^2_6. On the other hand, the individual values for 1988 in this case are $r = 7$, $e = 3.187$, $w = 2.58$ for $z = 2.37$. So the 1988 value is significant when taken on its own, but not when viewed as just one year amongst 11 years' data.

A similar analysis has also been carried out for station P, based on threshold 99.9, with linear trends (the YEAR covariate) in both the exceedance probabilities and excess values. In this case we found $\hat{\alpha} = 0.94$, standard error 0.048, for a log likelihood difference of 0.832 compared with the independent ($\alpha = 1$) case. In this case we do not believe the difference in the model fits to be significant, but again the predictive diagnostics do point to some improvement. For example, at level 139.9, we find $\sum z^2 = 10.29$ for the calibration by year, with 1988 values $r = 10$, $e = 5.093$, $z = 2.635$, and a Brier score for this level of 17.72. The results for the corresponding model with no dependence are $\sum z^2 = 23.17$, $e = 3.421$ and $z = 4.252$ for 1988, and Brier score 18.54.

Plots of the year-by-year comparisons of observed with expected exceedances, based on the dependent models for station P and the network maxima, are shown in Figures 5.9 and 5.10. In each case the agreement is better than for the corresponding plots in Figures 5.5 and 5.8. (For station P, the value for 1988 in the dependent linear trend model of Figure 5.9 is similar to that for the independent quadratic trend model of Figure 5.5, but the overall fit is much better in the case of Figure 5.9.)

As a final evaluation of the dependent models, Figures 5.11 and 5.12 show plots of the actual data on exceedances over 120, for station P and the network maxima, together with five simulations from our final dependent model. The purpose of this is to allow us to judge by eye how well the simulations are producing data sets that look like the real thing. The plot here is of day within the period (day 1=1/1/81) on the x axis, ozone in ppb on the y axis. For station P, the simulations generally produce similar 1988 values to those observed, and higher ones for 1983. In the case of network maxima (Figure 5.12), none of the simulations produces values as high as those actually observed for 1988, though the agreement elsewhere is reasonably good.

The overall conclusion from this section is that the independent models, or those that merely incorporate dependence through the inclusion of the PDAY variable in the model of Section 3, seem to fit the data when judged by probability plots of the excesses, but do not do an entirely satisfactory job of reproducing the higher levels as judged by the predictive diagnostics. The Markovian model introduced in this last part of the section does seem to get around that difficulty as well: although the models still underpredict the 1988 values, they are consistent with the data as judged by overall goodness of fit criteria.

In the present context we are, in effect, working not with the original process $\{X_i\}$ of daily ozone maxima, but with a censored process $\{\delta_i, Y_i\}$ where

$$\delta_i = \begin{cases} 1 & \text{if } X_i > u \\ 0 & \text{if } X_i \le u \end{cases}$$

and $Y_i = X_i - u$ if $\delta_i = 1$ and is undefined otherwise. To model this within the framework of (5.6), we therefore have to think of the observed process as a mixed discrete-continuous process. Suppose, on day i, the probability of exceeding the threshold u is p_i, and the conditional distribution of the excess, given the threshold is exceeded, is represented by a GPD $G(y; \sigma_i, \xi)$ from (5.2), with density $g = \partial G/\partial y$. Then the relevant density component f_1 is

$$\begin{cases} 1 - p_i & \text{if } \delta_i = 0, \\ p_i g(Y_i; \sigma_i, \xi) & \text{if } \delta_i = 1. \end{cases} \tag{5.7}$$

To complete the specification of the problem, then, we need to define a joint density f_2. To do this, we need to find a suitable form of the joint distribution function $F_2(x_1, x_2) = \Pr\{X_1 \le x_1,\ X_2 \le x_2\}$, which will be well-defined for $x_1 \ge u, x_2 \ge u$, that ideally will have the properties (a) the marginal distributions of the excesses have the GPD, (b) the distribution will contain independence as a special case, (c) the model should also incorporate some meaningful dependence among the extreme values. One model that has all these properties, given as equation (3.5) in Smith's (1993) review paper, is defined by

$$\begin{aligned} F_2(x_1, x_2) = \exp\Bigg[-\Bigg\{ & \left(-\log\left(1 - p_1\left(1 + \xi_1 \frac{x_1 - u}{\sigma_1}\right)_+^{-1/\xi_1}\right)\right)^{1/\alpha} \\ & + \left(-\log\left(1 - p_2\left(1 + \xi_2 \frac{x_2 - u}{\sigma_2}\right)_+^{-1/\xi_2}\right)\right)^{1/\alpha} \Bigg\}^{\alpha}\Bigg], \end{aligned} \tag{5.8}$$

valid in $x_1 \ge u,\ x_2 \ge u$. Here α is a parameter representing the strength of dependence between the two components. The permissible range is $0 \le \alpha \le 1$, where $\alpha = 1$ represents independence and the limit $\alpha \downarrow 0$ the total-dependence case in which the two components are equal with probability 1. The justification of the parametric form of (5.8) is that it represents a threshold version of the "logistic" dependence structure which has proved to be a valuable contribution to the literature on bivariate extreme value theory (see e.g. Tawn, 1988).

With F_2 defined by (5.8), the appropriate form of f_2 for insertion in (5.6) is

$$\begin{cases} F_2(u, u) & \text{if } \delta_1 = 0,\ \delta_2 = 0, \\ \partial F_2(u + Y_1, u)/\partial Y_1 & \text{if } \delta_1 = 1,\ \delta_2 = 0, \\ \partial F_2(u, u + Y_2)/\partial Y_2 & \text{if } \delta_1 = 0,\ \delta_2 = 1, \\ \partial^2 F_2(u + Y_1, u + Y_2)/\partial Y_1 \partial Y_2 & \text{if } \delta_1 = 1,\ \delta_2 = 1. \end{cases} \tag{5.9}$$

We thus have a model whose components are

Table 5.15: Calibration of probability forecasts by year
Network maxima, level 139.9, dependent model

Year	n	r	e	w	z
1981	214	4	4.039	3.268	−0.021
1982	214	4	4.365	3.736	−0.189
1983	214	13	12.284	8.009	0.253
1984	214	8	4.520	3.778	1.790
1985	214	3	2.613	2.301	0.255
1986	214	1	2.580	2.349	−1.031
1987	214	7	6.548	4.982	0.203
1988	214	13	8.930	6.217	1.632
1989	214	0	2.463	2.053	−1.719
1990	214	0	0.561	0.540	−0.763
1991	214	2	2.306	1.809	−0.227

The z values do appear to be better than those in the earlier Table 5.8. In this case $\sum z^2 = 10.728$, clearly not significant against ξ^2_{11}, though the interpretation of this is still a little dubious as several years have small e values. Regrouping in the same way as in Table 5.9, we have:

Table 5.16: Calibration of forecasts by year (regrouped)
Network maxima, level 139.9, dependent model

Year	n	r	e	w	z
1981–2	428	8	8.404	7.004	−0.153
1983	214	13	12.284	8.009	0.253
1984–6	642	12	9.713	8.428	0.788
1987	214	7	6.548	4.982	0.203
1988	214	13	8.930	6.217	1.632
1989–91	642	2	5.330	4.402	−1.587

for $\sum z^2 = 5.93$, not significant against χ^2_6. Note that the e values here are computed sequentially as we go through the data, i.e. the predicted value for day n does depend on the observed value for day $n-1$. However, this does appear to be a valid method of comparison, since this kind of sequential analysis is precisely the situation that the theory of Seillier-Moiseiwitsch and Dawid (1993) is designed for. As in our earlier analyses, we have dealt with the problem of parameter estimation by dropping one year at a time, each year's predictions being calculated using the model fitted to the other ten years' data.

- the probability of exceedance p_i, of the form (3.2) (with no PDAY, PDAY2 covariates in this case),

- the GPD $G(\cdot;\sigma_i,\xi)$ for the daily excesses, defined by (5.2) and (5.3),

- a first-order dependence parameter α, defined by (5.8), that captures the strength of dependence from day to day.

This model was then fitted to the network maxima data using the models of Tables 3.11 (without PDAY) and 5.5 to define starting values for, respectively, the exceedance probability parameters and the excess parameters. The final values from the fitted models are given in Tables 5.12 and 5.13. The estimated value of α is .940 with a standard error of .032, and the value of NLLH is 859.564, compared with an initial value of 862.729 with $\alpha = 1$. For technical reasons (J. Tawn, private communication) the likelihood ratio test statistic of the null hypothesis $\alpha = 1$ against the alternative $\alpha < 1$ does not follow the usual asymptotic χ_1^2 distribution, but the drop in NLLH would seem big enough to indicate a significant result. On the other hand, there is no evidence of a very strong dependence.

Table 5.12: Exceedance probability parameters under dependence model Network maxima: threshold 119.9

Parameter	Estimate	Stand. error	t Ratio
CONST	−21.85	1.575	−13.87
YEAR	−0.1545	0.03762	−4.11
CDAY	−1.047	0.7126	−1.47
SDAY	0.472	0.2954	1.60
T	0.2894	0.02157	13.42
RH	−0.01627	0.00967	−1.68
WIND.U	−0.0297	0.03868	−0.77
WIND.V	−0.1851	0.04311	−4.29
VIS	−0.06701	0.01955	−3.43
T.WSPD	−0.01795	0.002721	−6.60

Table 5.13: Excess parameters under dependence model
Network maxima: threshold 119.9

Parameter	Estimate	Stand. error	t Ratio
CONST	0.04538	1.556	0.03
YEAR	−0.04698	0.02982	−1.57
T	0.047	0.01877	2.50
WSPD	−0.1312	0.1283	−1.02
T.WSPD	−0.002606	0.005078	−0.51
ξ	−0.2367	0.07145	−3.31

Once again, we believe that the predictive diagnostics form the best guide as to how well this model is doing. Corresponding to the overall calibration in Table 5.7, we have the following:

Table 5.14: Calibration of probability forecasts
Network maxima, level 139.9, dependent model

p_{min}	p_{max}	n	r	e	w	z
0.000	0.050	2107	7	7.902	7.722	−0.325
0.050	0.100	93	5	6.477	6.009	−0.602
0.100	0.300	116	23	20.410	16.422	0.639
0.300	0.500	30	14	11.618	7.002	0.900
0.500	1.000	8	6	4.803	1.887	0.871
0.000	1.000	2354	55	51.210	39.042	0.607

Here $\sum z^2 = 2.45$ based on the first five rows, and the Brier score is 38.70 ($Y_n^B = -0.08$) compared with the earlier value of 40.42 for the model without dependence. The calibration by year leads to: