NISS

# Predicting the Urban Ozone Levels and Trends with Semiparametric Modeling

Feng Gao, Jerome Sacks and William J. Welch

Technical Report Number 14
May, 1994

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

# Predicting the Urban Ozone Levels and Trends with Semiparametric Modeling

Feng Gao[1], Jerome Sacks[2] and William J. Welch[3]

May 17, 1994

## Abstract

Ozone in the lower layer of the earth's atmosphere, the troposphere, is considered harmful to plants and human health. The surface ozone levels are determined by the strengths of sources and precursor emissions, and by the meteorological conditions. EPA has imposed regulations to encourage practices that restrict the production of ozone and has for many years monitored ozone concentration levels across the USA. Whether these regulations have effectively reduced ozone levels is in question because assessing the ozone trends is complicated by meteorological variability. In this report, a semiparametric modeling technique is used to build models which use meteorology to predict ozone levels. This technique can be used to estimate that part of the trend in ozone levels that cannot be accounted for by meteorology. The indications are that there is a decrease in ozone levels once the meteorological effects on ozone have been adjusted for. The models based on this technique are also shown to provide predictions which closely match the actual ozone levels. The results complement those of Bloomfield, Royle and Yang (1993) which produces similar conclusions.

*Key words*: Ozone concentration, meteorological adjustment, semiparametric regression.

## 1 Introduction

High ozone concentration in the troposphere is believed to be harmful to human health and to cause damage to crops (see National Research Council (1991)). The variability in surface ozone concentration levels is affected by the strengths of sources and precursor emissions, and by meteorological conditions. In order to assess that part of the trend in ozone concentration levels that cannot be accounted for by meteorology, we need to build models which relate ozone to meteorology.

[1]National Institute of Statistical Sciences, P. O. Box 14162, Research Triangle Park, NC 27709-4162. Research supported in part by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0 and National Science Foundation Grant DMS-9208758.

[2]National Institute of Statistical Sciences, P. O. Box 14162, Research Triangle Park, NC 27709-4162. Research supported in part by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0.

[3]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1 Canada. Research supported in part by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0.

In Bloomfield et al. (1993), nonlinear least squares methods were used to model the dependence of ozone on meteorological factors, and to estimate the trends. That report focuses on the urban Chicago area. The analysis and model building were based on daily data from a network of ozone monitoring stations in the Chicago area from 1981 to 1991, and on meteorological data from the same area in that period.

In this follow-on report, a semiparametric modeling technique is used to build models that relate ozone to meteorology. The approach taken produces results similar to that obtained in Bloomfield et al. (1993), but under weaker assumptions, thereby lending additional credibility to the results in Bloomfield et al. (1993). The models constructed were also used to predict behavior of ozone in future years as a function of meteorology with generally very satisfactory prediction errors.

## 2 Semiparametric Model

The ozone concentration levels to be modeled are the daily network *typical value*, the daily network *maximum* or the daily network *secondary maximum*. To obtain the daily network typical value, the least absolute deviations decomposition (or the median polish decomposition, see Chapters 10 and 11 of Tukey (1977)) of $y_{d,s}$, the maximum concentration on day $d$ at station $s$, was performed for all the 45 ozone monitoring stations in the network:

$$y_{d,s} = \mu' + \alpha'_d + \beta'_s + \epsilon'_{d,s}$$

The daily network typical value is then defined as $\mu' + \alpha'_d$. The decomposition was also used to impute the missing data. This daily network typical value is called the network average in Bloomfield et al. (1993). The daily network maximum and the daily network secondary maximum are the largest value and the second largest value of the station daily maxima in the network. They are based on the daily maxima over stations with imputed missing values. The unit for all three responses is parts per billion (ppb).

For simplicity, the same meteorological variables adopted by Bloomfield et al. (1993) are used here. The surface weather data were taken at O'Hare Airport and the upper air weather data were taken from a station at Peoria. The variables used in the models are:

- maximum temperature from 9:00 am to 6:00 pm (maxt)
- 12 noon wind speed (wspd)
- 24 hour average wind vector (meanu and meanv)
- 12 noon relative humidity (rh)
- 12 noon visibility (vis)

- 12 noon opaque cloud cover (opcov)
- 7 am wind speed at 700 mb (wspd700)
- 24 hour average temperature lagged 1 day and 2 days (tlag1 and tlag2)
- 24 hour average wind speed lagged 1 day (wlag)
- 24 hour average relative humidity lagged 1 day (rhlag)

Also used is a variable for year, which takes the integer values 1,2, ..., 11, corresponding to years 1981 - 1991, and a variable for day taking values from 1 to 365 to reflect seasonal effects.

On day $i$, in year $j$, with meteorological condition $met$, where $met$ is a 12-dimensional vector of the above meteorological variables, let $x = (met, i, j)$. So $x$ is a 14-dimensional vector $x = (\xi_1, \ldots, \xi_{14})$. The response $y(x)$ (the network typical value or the network maximum or the network secondary maximum) is assumed to be a realization of a stochastic process, $Y(x)$:

$$Y(met, i, j) = \beta_j + Z(met, i) + \varepsilon_{ij} \tag{2.1}$$

where $\beta_j$ are constants, $j = 1, 2, \ldots, 11$, $Z(x) = Z(met, i)$ is a zero mean Gaussian process with covariance function $\mathrm{Cov}(Z(x), Z(x')) = \sigma_Z^2 R(x, x')$ to be specified later, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$. For more discussion on the use of this technique for modeling response surfaces, see Sacks, Welch, Mitchell and Wynn (1989), and the references cited there.

Assume, as in Sacks et al. (1989), that the covariance between $Z(x)$ and $Z(x')$ is

$$\sigma_Z^2 R(x, x') = \sigma_Z^2 \exp(-\sum_{k=1}^{14} \theta_k |\xi_k - \xi_k'|^{p_k}) \tag{2.2}$$

where $x = (\xi_1, \ldots, \xi_{14})$, $x' = (\xi_1', \ldots, \xi_{14}')$, $\theta_k \geq 0$, $k = 1, \ldots, 13$, $\theta_{14} = 0$ and $1 \leq p_k \leq 2$, $k = 1, \ldots, 14$. $\theta_{14}$ corresponds to the variable year. This class of stationary processes provides us with a wide range of functions.

Given the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ for $p$ consecutive years starting from year 1 (1981) with $n_j$ data points in year $j$ and $n_1 + \cdots + n_p = n$ and, provided $\sigma_Z$, $\sigma_\varepsilon$ and $R(\cdot, \cdot)$ are known, the best linear unbiased predictor (BLUP) $\hat{y}(x)$ at a new point $x$ in year $j$ can be written as (see Sacks et al. (1989))

$$\hat{y}(x) = \hat{\beta}_j + \hat{Z}(x) = \hat{\beta}_j + r'(x) C^{-1} (y - F\hat{\beta}) \tag{2.3}$$

where $y = (y_1, y_2, \ldots, y_n)$, $C = \mathrm{Corr}(y) = (\sigma_Z^2/\sigma^2) R + (\sigma_\varepsilon^2/\sigma^2) I$, where $\sigma^2 = \sigma_Z^2 + \sigma_\varepsilon^2$, and $R = \{R(x_i, x_j), 1 \leq i \leq n; 1 \leq j \leq n\}$, is the $n \times n$ matrix of correlations among $Z$'s at the data

points, $r(x) = (\sigma_Z^2/\sigma^2)[R(x_1, x), \ldots, R(x_n, x)]'$,

$$F = \begin{pmatrix} \vec{1}_{n_1 \times 1} & \vec{0} & \cdots & \vec{0} \\ \vec{0} & \vec{1}_{n_2 \times 1} & \cdots & \vec{0} \\ \vdots & \vdots & \ddots & \vdots \\ \vec{0} & \vec{0} & \cdots & \vec{1}_{n_p \times 1} \end{pmatrix}_{n \times p},$$

and $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)' = (F'C^{-1}F)^{-1}F'C^{-1}y$, which is the usual generalized least-squares estimate of $\beta = (\beta_1, \ldots, \beta_p)'$.

In the model, $\sigma_\varepsilon$ represents random variability that cannot be explained by modeling. Values of $p$ indicate smoothness of the response surface as a function of the corresponding variables. Larger values of $\theta$ usually indicate greater importance of the corresponding variables if the variables are on normalized scales.

Note that the year variable $j$ does not appear in the stochastic process $Z(\cdot)$ in the model (2.1), or, equivalently, $\theta$ for year in (2.2) is assumed to be zero. The reasons are: for most of the monthly periods tried, the estimated $\theta$ for year was zero. And for the remaining few monthly periods, the estimated $\theta$'s for year were small, indicating that year was not an important variable. Because the adjusted trend of ozone could be unambiguously interpreted through the $\beta_j$'s if $j$ does not appear in $Z(\cdot)$ (see Section 3.3 for details), we choose to set the $\theta$ for year equal to zero in all cases.

To obtain the unknown parameters $\sigma_Z$, $\sigma_\varepsilon$, $\theta$'s and $p$'s in the model, maximum likelihood estimate (MLE) method is used. These estimates are then used in (2.3) to predict the response surface.

The main concern for ozone is in the summer, the period when its concentration is high. This work will focus on the period from May 15 to September 15, the middle 4 month period of the 7 month ozone season as studied in Bloomfield et al. (1993). This period is further divided into 4 smaller periods: May 15 - June 15, June 15 - July 15, July 15 - August 15 and August 15 - September 15. A model is fitted for each of the 4 periods separately. There are basically two reasons for doing so: First, the assumption of stationarity of $Z$ within a shorter time period is more plausible than it would be over longer ones. Secondly, the computational burden is reduced because a likelihood evaluation requires $O(n^3)$ operations, so cutting the data size by 1/4 produces an overall reduction factor of 1/16. It can take up to 3 hours on a Sun Sparc Classic machine to do one maximum likelihood estimate with 11 years worth of data for a one month period.

# 3 Modeling the Network Typical Value

A model is fitted using data from 1981 to 1991 for each of the 4 periods mentioned in the previous section. The ozone data used for the model are the network typical values described in the previous section.

Table 3.1: Estimates of $\theta$'s and $p$'s for models for the network typical value with meteorological variables rescaled.

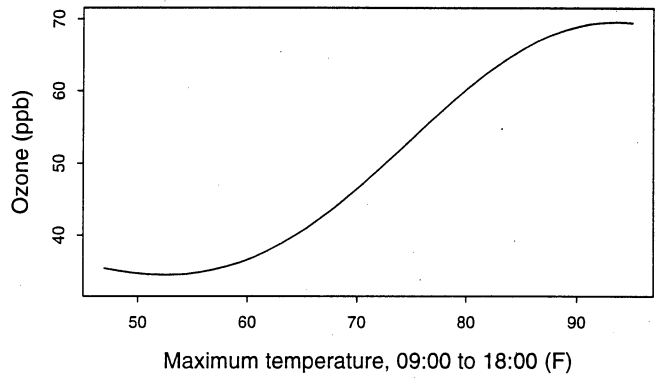
| | May 15 - June 15 | | June 15 - July 15 | | July 15 - Aug. 15 | | Aug. 15 - Sept. 15 | |
|---|---|---|---|---|---|---|---|---|
| variables | $\theta$ | $p$ | $\theta$ | $p$ | $\theta$ | $p$ | $\theta$ | $p$ |
| maxt | 3.3222 | 2 | 4.8540 | 2 | 0.8202 | 2 | 1.7082 | 2 |
| wspd | 0.2368 | 2 | 0.0000 | 2 | 0.3261 | 2 | 0.0000 | 2 |
| meanu | 0.0000 | 2 | 0.7052 | 1.185 | 0.0733 | 1.435 | 0.3849 | 1 |
| meanv | 0.0000 | 2 | 1.4913 | 2 | 0.5097 | 2 | 1.6454 | 2 |
| rh | 1.7655 | 2 | 0.7756 | 2 | 0.5506 | 2 | 1.9768 | 2 |
| vis | 0.1833 | 2 | 1.1844 | 2 | 0.0333 | 2 | 0.6480 | 2 |
| opcov | 0.0000 | 2 | 0.0731 | 2 | 0.0000 | 2 | 0.0617 | 2 |
| wspd700 | 0.5513 | 2 | 0.0000 | 2 | 0.0170 | 2 | 0.1922 | 2 |
| tlag1 | 0.0000 | 2 | 0.5948 | 2 | 0.0000 | 2 | 0.0000 | 2 |
| tlag2 | 0.0928 | 2 | 0.0000 | 2 | 0.0462 | 2 | 0.2011 | 2 |
| wlag | 0.0481 | 1 | 0.6122 | 2 | 0.0000 | 2 | 2.2078 | 2 |
| rhlag | 0.0000 | 2 | 0.3031 | 2 | 0.0000 | 2 | 0.1955 | 2 |
| day | 0.0939 | 2 | 0.1270 | 2 | 0.0000 | 2 | 0.1172 | 2 |

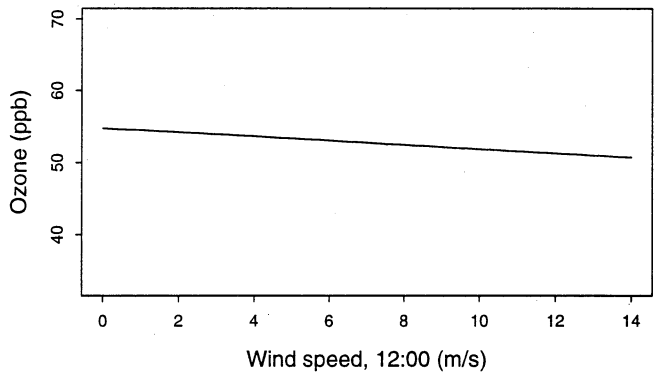
## 3.1 Important Variables

Each meteorological variable has its own scale and in order to indicate which ones have strong effects it is useful to rescale so that each meteorological variable ranges over [0,1]. The maximum likelihood estimates of the $\theta$'s and $p$'s with the rescaled meteorological variables and the rescaled variable day are given in Table 3.1. The importance of a variable is reflected by the magnitude of its corresponding $\theta$.

From the table, it can be seen that temperature, relative humidity and wind (through wspd, wlag, meanu, meanv and/or wspd700) are consistently important across the months.

It is possible and useful to get from a fitted model the "main effects" and the "joint effects" for some of the important meteorological variables. In designed experiments over a rectangle, the approach in Sacks et al. (1989) can be taken. But that approach is not transferable in the current observational setting. We take two alternative approaches. First, to obtain a main effect of one meteorological variable, fix all the rest of the meteorological variables at their median levels in the fitted model, so that ozone becomes a function of that single meteorological variable. This could be viewed as the main effect for that variable. The second way is to average the modeled response over all the $x$ variables except the one of interest. Because of correlations between the $x$ variables, we average over their empirical joint distribution in the data. Joint effects for pairs of $x$ variables are obtained similarly by the two methods. These two methods should lead to different main effects and joint effects. However, for the models fitted here for the network typical values, the two methods



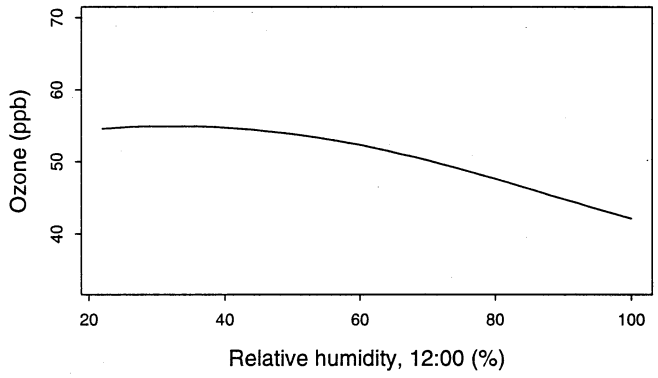




Figure 3.1: Main effect plots for the period of May 15 - June 15.

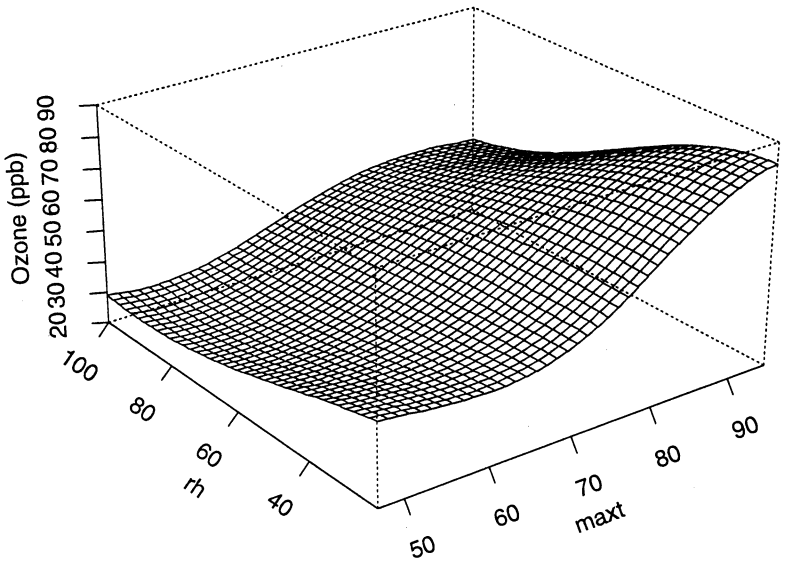



Figure 3.2: Joint effect of maximum temperature from 09:00 to 18:00 and 12 noon relative humidity for the period of May 15 - June 15.

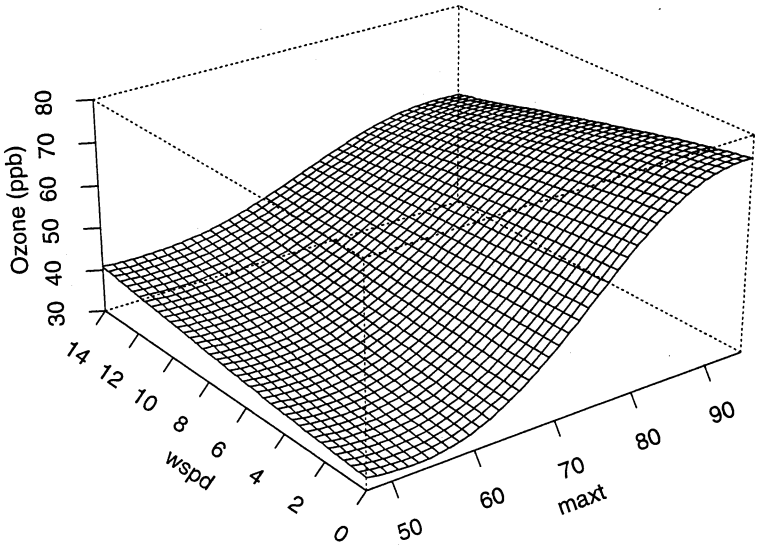



Figure 3.3: Joint effect of maximum temperature from 09:00 to 18:00 and 12 noon wind speed for the period of May 15 - June 15.

give very similar main effects and joint effects for the important variables. Figures 3.1, 3.2 and 3.3 show the main effects for maximum temperature from 9:00 am to 6:00 pm, 12 noon wind speed and 12 noon relative humidity, and the joint effects for them for the period of May 15 - June 15. Similar plots could be generated for other periods. The plots shown here were produced by the first method. These plots look similar to the plots in Figures 12, 14 and 15 in Bloomfield et al. (1993). The nonlinear model in Bloomfield et al. (1993) was developed partly based on Figures 12, 14 and 15 in that report. The main effect and the joint effect plots obtained here demonstrate that the semiparametric model provides an effective tool for discovering and interpreting important relationships between responses and variables.

## 3.2 Quality of the Fitted Models

To check the quality of the model fit, the "cross validation" (CV) prediction at the $i$th data point $x_i$ is calculated by (2.3) using all but the $i$th data point. This CV prediction is denoted by $\hat{y}_{-i}(x_i)$. The MLE estimates of the parameters are based on all the data, and are only calculated once, not re-estimated each time. For this reason, *cross validation* was put in the quotes. The CV prediction is calculated for every data point and compared with the actual network typical value at that point. Figure 3.4 shows the scatter plots of CV predictions from the fitted models vs. the actual network typical values. The points lie along the 45 degree line, indicating that the model fits are generally good, though the models underpredict when the ozone levels are very high. This is not surprising: high responses will be smoothed by models of the type used. Models are plausibly inadequate to explain solely on available meteorology what happens to cause extreme conditions. The highest CV residual, 47.8 ppb, occurs at July 30, 1983, when the actual network typical value is 105.5 ppb, and the CV predicted value is 57.7 ppb.

For a data set of 11 years, suppose the $i$th data point is from year $j$. Let $f_j$ be an 11-dimensional vector with the $j$th element 1 and the others 0. Let $F_{-i}$ be the matrix $F$ defined in Section 2 following formula (2.3) with the $i$th row deleted. Similarly, let $r_{-i}(x_i)$ be the vector $r(x_i)$ with the $i$th element removed, and let $C_{-i}$ be the matrix $C$ with the $i$th data point left out. With these definitions, the mean square error (MSE) for $\hat{y}_{-i}(x_i)$ is (see Sacks et al. (1989)):

$$\mathrm{MSE}[\hat{y}_{-i}(x_i)] = \sigma^2 \left[ 1 - (f_j', r_{-i}'(x_i)) \begin{pmatrix} 0 & F_{-i}' \\ F_{-i} & C_{-i} \end{pmatrix}^{-1} \begin{pmatrix} f_j \\ r_{-i}(x_i) \end{pmatrix} \right]. \tag{3.1}$$

Replacing $\sigma_Z$, $\sigma_\varepsilon$, $\theta$'s and $p$'s by their MLEs, an estimated value for $\mathrm{MSE}[\hat{y}_{-i}(x_i)]$ for every data point can be obtained. Then standardized CV residuals can be calculated by $[y_i - \hat{y}_{-i}(x_i)]/\{\mathrm{MSE}[\hat{y}_{-i}(x_i)]\}^{1/2}$, where $y_i$ is the $i$th data.

Figure 3.5 shows the Q-Q plots of the standardized CV residuals from the fitted models against the standard normal distribution. The plots show that these standardized CV residuals are reasonably close to standard normal, suggesting that the MSE (3.1) is a useful measure of uncertainty

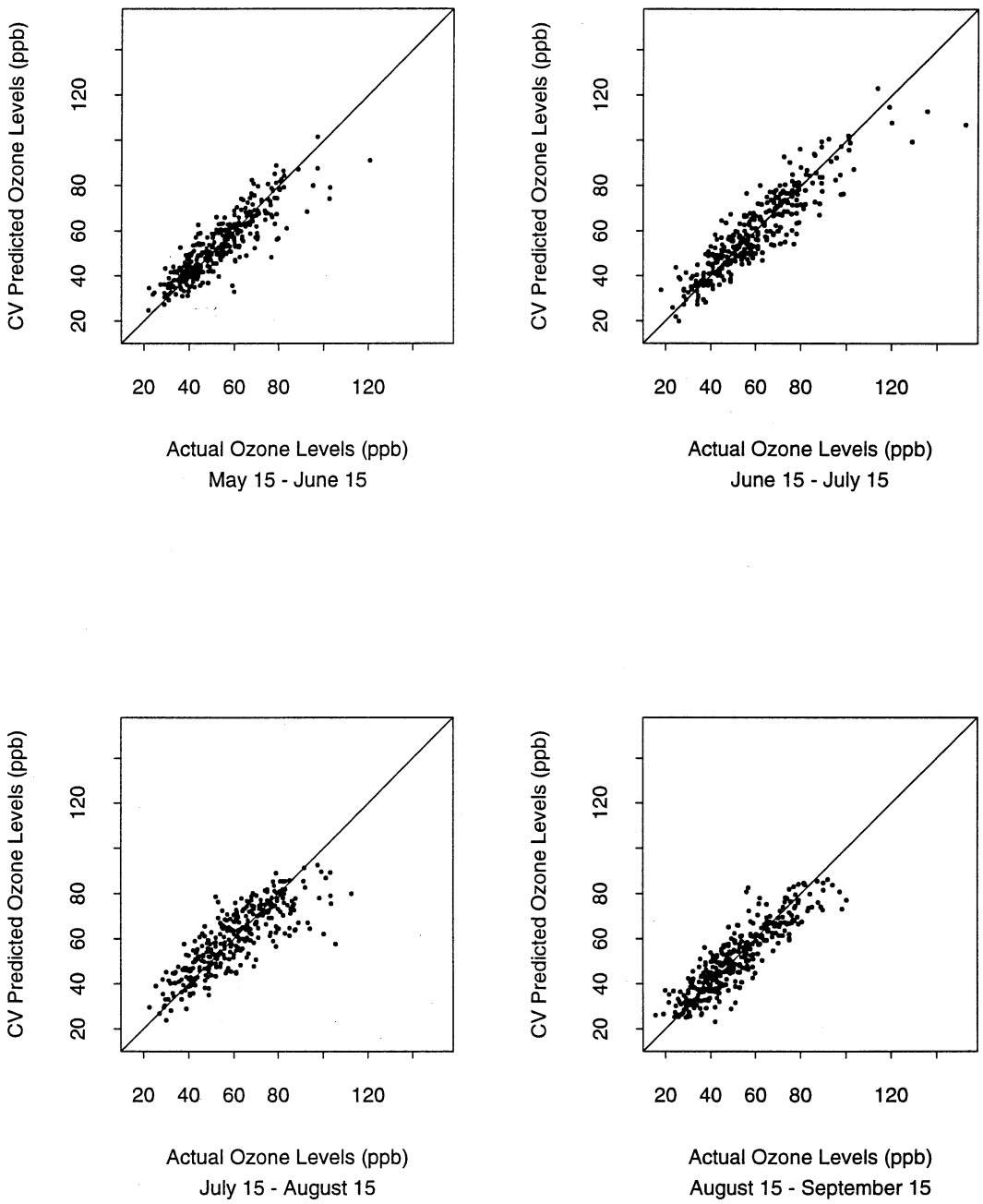



Figure 3.4: CV predicted network typical values vs. actual values.

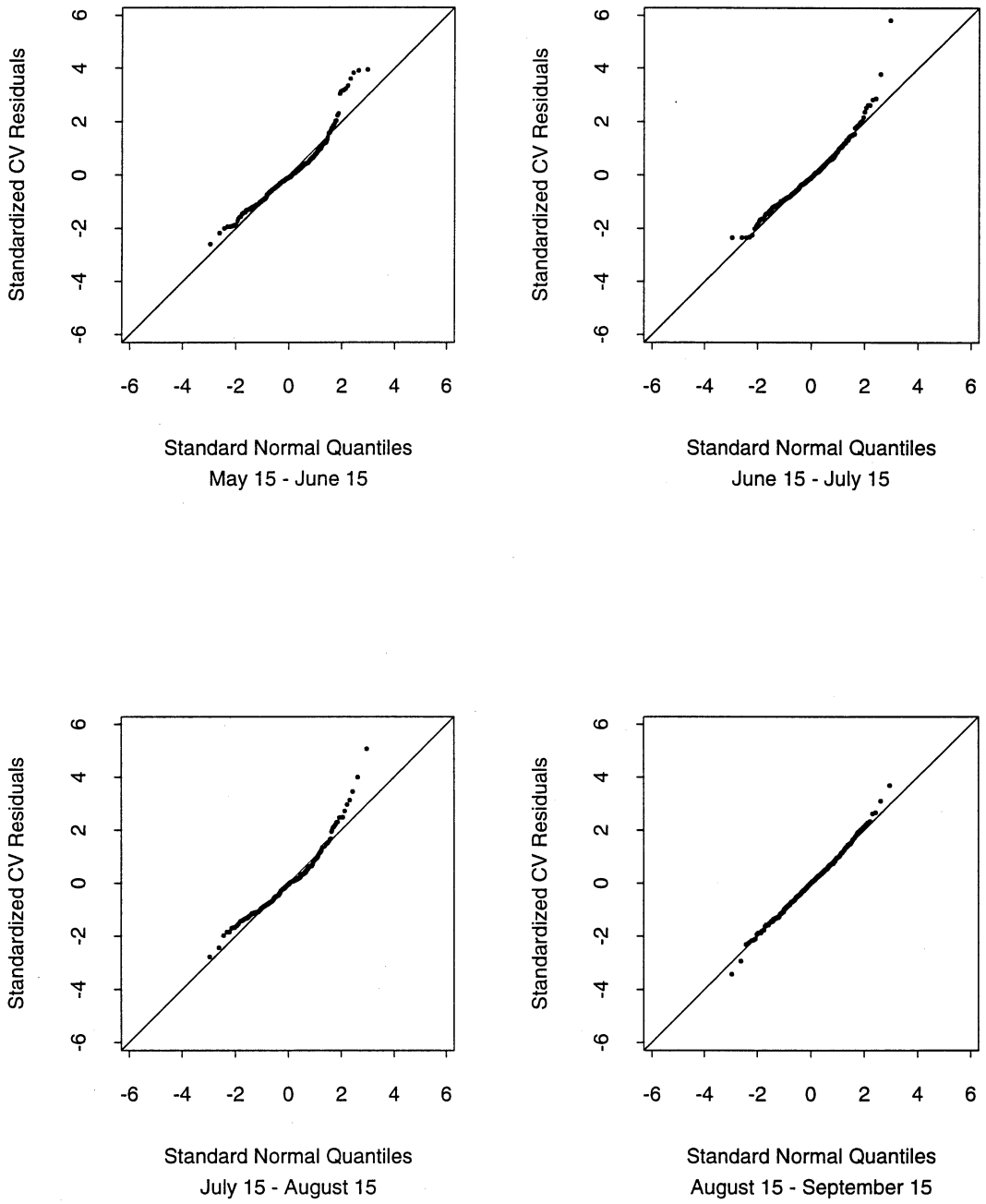



Figure 3.5: Q-Q plots of the standardized CV residuals for the network typical values vs. standard normal distribution.

Table 3.2: Estimates of $\sigma_Z$ and $\sigma_\varepsilon$, and CVRMSE for models for the network typical value.

| Models | $\hat{\sigma}_Z$ | $\hat{\sigma}_\varepsilon$ | CVRMSE |
|---|---|---|---|
| May 15 - June 15 | 17.051 | 7.028 | 7.606 |
| June 15 - July 15 | 17.584 | 7.028 | 8.433 |
| July 15 - Aug. 15 | 33.194 | 9.236 | 9.828 |
| Aug. 15 - Sept. 15 | 15.263 | 6.281 | 7.680 |

for prediction and that the model used fits well. Some transformations such as $\log(y)$ and $y^2$ of the original ozone data were also tried to fit the model. They reduced the skewness in the Q-Q plots. However they did not improve the overall quality of model fitting in terms of CV predictions. Using the transformed ozone data also makes it harder to estimate and interpret the adjusted trends and their standard errors for the ozone. Further more, the ozone data in the original scale were used in Bloomfield et al. (1993). It would be easier to compare the results in this report with those in Bloomfield et al. (1993) if data in the same scale are used. Therefore ozone concentration data in their original scale are used in this report.

The CV root mean square error (CVRMSE) can also be calculated by:

$$\text{CVRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{-i}(x_i) - y_i)^2}. \tag{3.2}$$

If the model fit is good, the CVRMSE should be close to $\sigma_\varepsilon$ or its MLE estimate.

Table 3.2 lists the MLEs of $\sigma_Z$ and $\sigma_\varepsilon$ and the CVRMSEs for the fitted models. The table shows that the CVRMSEs and the MLEs of $\sigma_\varepsilon$ for the fitted models are fairly close. The values of CVRMSE are close to the values of root mean square residual from the parametric model fitting in Table 6 of Bloomfield et al. (1993). Table 3.2 also shows that the $\hat{\sigma}_Z$ for the period of July 15 - August 15 is the largest, indicating greater variability in the regression surface for this period. The largest cross validation residual also occurs in this period.

## 3.3 Trend Estimation

The main goal in the study is to reveal trends in ozone concentration levels, if there are any, after adjusting for meteorology.

It is possible to interpret the adjusted trend through the $\beta_j$'s in the model when the variable year does not appear in the stochastic process part of the model $Z(\cdot)$. Under this circumstances, if $met$ is held fixed, the change from year to year is, except for random errors $\varepsilon$, reflected in the differences of the $\beta_j$'s. Therefore the adjusted trend is defined as the trend in the $\beta_j$'s.

Let $\beta_j^* = \hat{\beta}_j + (\bar{y} - \bar{\hat{\beta}})$, then $\bar{\beta}^* = \bar{y}$. These $\beta_j^*$'s can be interpreted as the adjusted (for meteorology) averages of ozone level across the years while the simple yearly averages $\bar{y}_j$'s are the unadjusted averages. Time series plots of the adjusted and the unadjusted averages for the network

typical values are given in Figure 3.6. The plots demonstrate that a large portion of the variability in the unadjusted averages is eliminated in the adjusted averages. This portion of the variability is caused by meteorology. The plots suggest a linear trend for the adjusted averages for all the four periods. The lines in the plots are the least square regression lines for the adjusted averages.

Use of the straight line to assess the trend is a summary estimate. Let $\hat{a}$ be the intercept at year= 81 and $\hat{b}$ be the slope of the line, then the estimate of the adjusted trend is

$$\hat{trend} = 10 \times \frac{\hat{b}}{\hat{a}} \qquad (\%/\text{decade}). \tag{3.3}$$

Let $J = (0, 1, \ldots, 10)'$, and assume the model is correct, the variances of $\hat{b}$ and $\hat{b}/\hat{a}$ can then be calculated by

$$\text{Var}(\hat{b}) = (J - \bar{J})'\text{Cov}(\hat{\beta})(J - \bar{J})/((J - \bar{J})'(J - \bar{J}))^2 \tag{3.4}$$

$$\text{Var}\left(\frac{\hat{b}}{\hat{a}}\right) \approx \left(\frac{\hat{b}}{\hat{a}}\right)^2 \left(\frac{\text{Var}(\hat{b})}{\hat{b}^2} + \frac{\text{Var}(\hat{a})}{\hat{a}^2} - \frac{2\text{Cov}(\hat{a}, \hat{b})}{\hat{a}\hat{b}}\right) \tag{3.5}$$

where

$$\text{Cov}(\hat{\beta}) = \sigma^2(F'C^{-1}F)^{-1},$$

$$\text{Var}(\hat{a}) = \text{Var}(\bar{y}) + \bar{J}^2\text{Var}(\hat{b}) - 2\bar{J}\text{Cov}(\hat{b}, \bar{y}),$$

$$\text{Cov}(\hat{a}, \hat{b}) = \text{Cov}(\hat{b}, \bar{y}) - \bar{J}\text{Var}(\hat{b}).$$

Formula (3.5) is from page 181 of Mood, Graybill and Boes (1974). Replacing $\sigma$, $\theta$'s and $p$'s by their MLEs, estimated values of $\text{Var}(\hat{b})$ and $\text{Var}(\hat{trend})$ can be obtained.

Another way to assess variances of the slopes and the trends is by jackknifing. The *leave-out-one* estimates of $\beta$ can be used to construct "pseudo-values" (see Chapter 8 of Mosteller and Tukey (1977)) of slopes and trends from:

$$\text{pseudo-value} = n(\text{estimate from all data}) - (n-1)(\text{leave out one estimate}),$$

where $n$ is the number of data in the model. The pseudo-values are then treated as a sample of values estimating the given parameters (here they are slopes and trends). Their mean is used as the "jackknifed" estimate of the parameter, and the standard error of the mean gives a standard error for either the original parameter estimate or the jackknifed estimate. Although the theoretical properties of the jackknifed estimates in semiparametric models like ours are unclear, it is interesting to calculate the jackknifed estimates and compare them with those based on the model and the MLEs.

The estimates of the slopes and trends and their standard errors are listed in Table 3.3 and Table 3.4.

The tables show that the jackknifed estimates are very close to the original estimates. From the tables, it is seen that for the periods of May 15 - June 15 and July 15 - August 15, there are no significant trends in the adjusted averages. However, for the periods of June 15 - July 15 and August

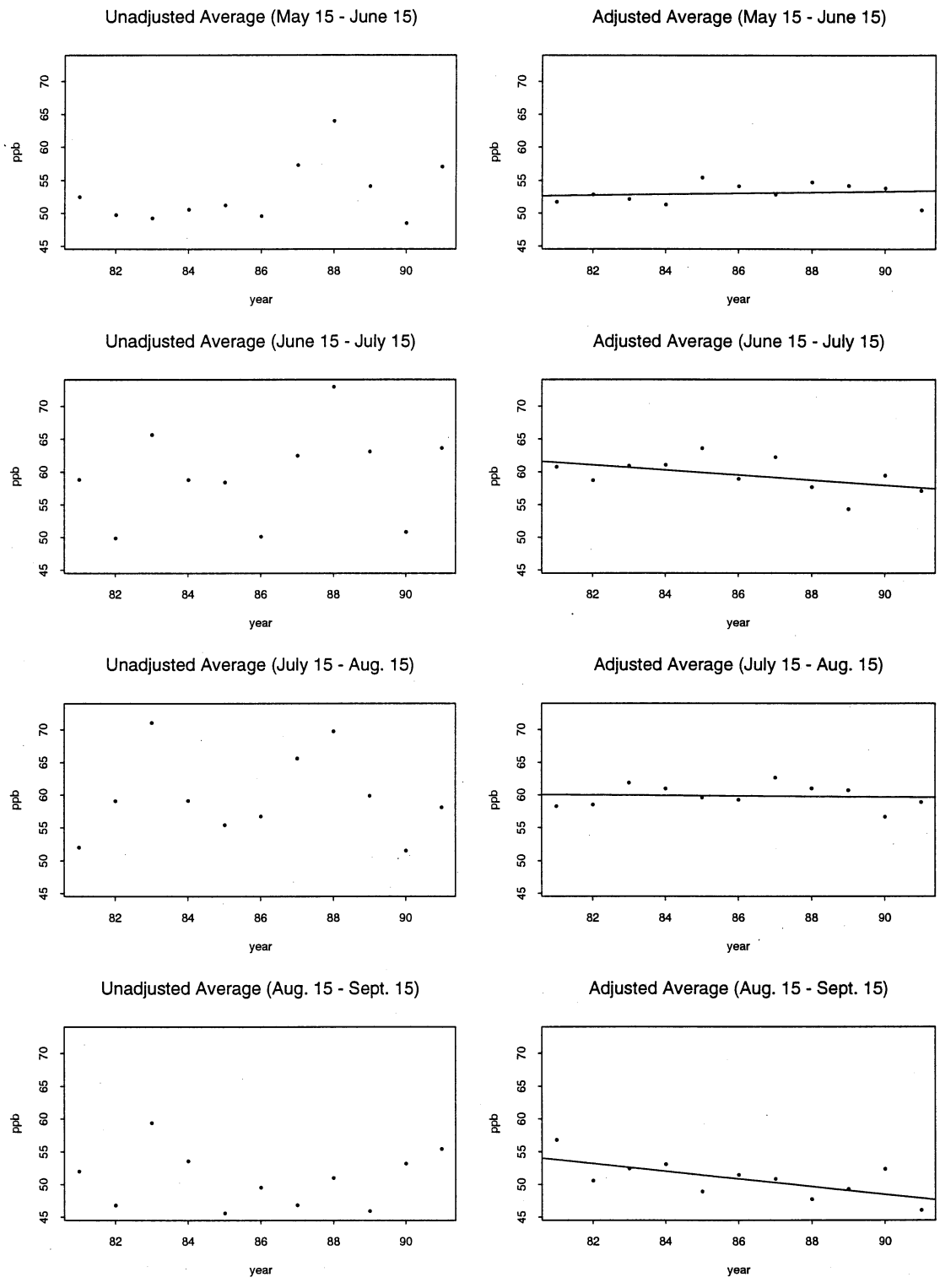



Figure 3.6: Adjusted and unadjusted averages of the network typical values.

Table 3.3: Estimates of slopes and their standard errors for the adjusted averages for the network typical value.

| | Model Estimates | | | Jackknifed Estimates | | |
|---|---|---|---|---|---|---|
| Models | Slope | Standard Error | $t$ Value | Slope | Standard Error | $t$ Value |
| May 15 - June 15 | 0.0731 | 0.1458 | 0.5014 | 0.0808 | 0.1505 | 0.5369 |
| June 15 - July 15 | −0.3909 | 0.1639 | 2.3850 | −0.3982 | 0.1829 | 2.1771 |
| July 15 - Aug. 15 | −0.0443 | 0.1882 | 0.2354 | −0.0524 | 0.1867 | 0.2807 |
| Aug. 15 - Sept. 15 | −0.5884 | 0.1410 | 4.1730 | −0.6155 | 0.1642 | 3.7485 |

Table 3.4: Estimates of trends and their standard errors for the adjusted averages for the network typical value.

| | Model Estimates | | | Jackknifed Estimates | | |
|---|---|---|---|---|---|---|
| Models | Trend | Standard Error | $t$ Value | Trend | Standard Error | $t$ Value |
| May 15 - June 15 | 0.0139 | 0.0280 | 0.4964 | 0.0149 | 0.0288 | 0.5174 |
| June 15 - July 15 | −0.0635 | 0.0285 | 2.2281 | −0.0651 | 0.0287 | 2.2683 |
| July 15 - Aug. 15 | −0.0074 | 0.0313 | 0.2364 | −0.0093 | 0.0310 | 0.3000 |
| Aug. 15 - Sept. 15 | −0.1094 | 0.0330 | 3.3152 | −0.1146 | 0.0290 | 3.9517 |

15 - September 15, there are significant decreasing trends in the adjusted averages. According to the jackknifed estimates, the adjusted network typical value decreased by about 4.0 ppb over the decade with a standard error of 1.8 ppb for the period of June 15 - July 15, and by about 6.2 ppb over the decade with a standard error of 1.6 ppb for the period of August 15 - September 15. The jackknifed estimates of trend are $-6.5\%$/decade with a standard error of 2.9%/decade for the period of June 15 - July 15, and $-11.5\%$/decade with a standard error of 2.9%/decade for the period of August 15 - September 15.

## 3.4 Predictions

Another goal in the study is to build models to predict ozone concentration levels from meteorology. To see how well the semiparametric model can predict, data from 1981 to 1987 were taken to build a model, and then 1988 meteorology was used in the model to predict 1988 ozone levels. To obtain a $\beta_8$ for 1988, linear extrapolation was used based on $\hat{\beta}_1, \ldots, \hat{\beta}_7$, the estimates from the 1981 - 1987 data set. (In the model, a different $\beta$ was given to each year to avoid assuming linear trend in year. Results from the previous section indicate that the trend is approximately linear. A linear model $a + b \times year$ could have been used in place of $\beta$'s with little loss for prediction purposes.) The predictions were then compared with the actual ozone levels for 1988. Similarly, data from 1981 to 1988 can be used as "training" data to build a model, which can then be used to

Table 4.1: Estimates of $\theta$'s and $p$'s for models for the network maximum with meteorological variables rescaled.

| | May 15 - June 15 | | June 15 - July 15 | | July 15 - Aug. 15 | | Aug. 15 - Sept. 15 | |
|---|---|---|---|---|---|---|---|---|
| variables | $\theta$ | $p$ | $\theta$ | $p$ | $\theta$ | $p$ | $\theta$ | $p$ |
| maxt | 3.6265 | 2 | 6.0549 | 2 | 1.3784 | 2 | 2.0112 | 2 |
| wspd | 0.3702 | 2 | 1.4934 | 1.582 | 0.0000 | 2 | 0.0000 | 2 |
| meanu | 0.0000 | 2 | 1.7544 | 2 | 2.4429 | 2 | 1.8713 | 2 |
| meanv | 0.7260 | 2 | 0.9142 | 1.042 | 6.4683 | 2 | 3.7599 | 2 |
| rh | 1.8617 | 2 | 0.9876 | 2 | 0.0000 | 2 | 0.8022 | 2 |
| vis | 0.0742 | 2 | 1.5417 | 2 | 0.1891 | 2 | 0.1956 | 2 |
| opcov | 0.0000 | 2 | 0.0000 | 2 | 0.0820 | 2 | 0.0673 | 2 |
| wspd700 | 0.7158 | 2 | 0.0000 | 2 | 0.0000 | 2 | 0.1638 | 2 |
| tlag1 | 0.0000 | 2 | 1.0731 | 2 | 0.0000 | 2 | 0.0000 | 2 |
| tlag2 | 0.0000 | 2 | 0.0000 | 2 | 0.0000 | 2 | 0.3565 | 2 |
| wlag | 0.3703 | 2 | 0.4233 | 1 | 0.1342 | 2 | 1.4558 | 2 |
| rhlag | 0.0000 | 2 | 0.0000 | 2 | 0.0000 | 2 | 0.0000 | 2 |
| day | 0.0643 | 2 | 0.0000 | 2 | 0.0000 | 2 | 0.1253 | 2 |

predict 1989 ozone levels, and so on. And finally, the model based on 1981 - 1991 data was used to predict 1992 ozone levels.

Figures 3.7 to 3.11 show the results. The predictions are generally a good match to the actual levels, though it is very difficult for the model to capture the peaks in the unusually severe year, 1988.

# 4 Modeling the Network Maximum

Parallel to the analysis for the network typical value, the same type of model can be fitted for the network maximum value with the same meteorology data.

## 4.1 Important Variables

The maximum likelihood estimates of the $\theta$'s and $p$'s with the rescaled meteorological variables are given in Table 4.1.

From the table, it can be seen that temperature, relative humidity and wind (through wspd, wlag, meanu, meanv and/or wspd700) are consistently important across the months (relative humidity is unimportant for the third month), as was the case for network typical values.

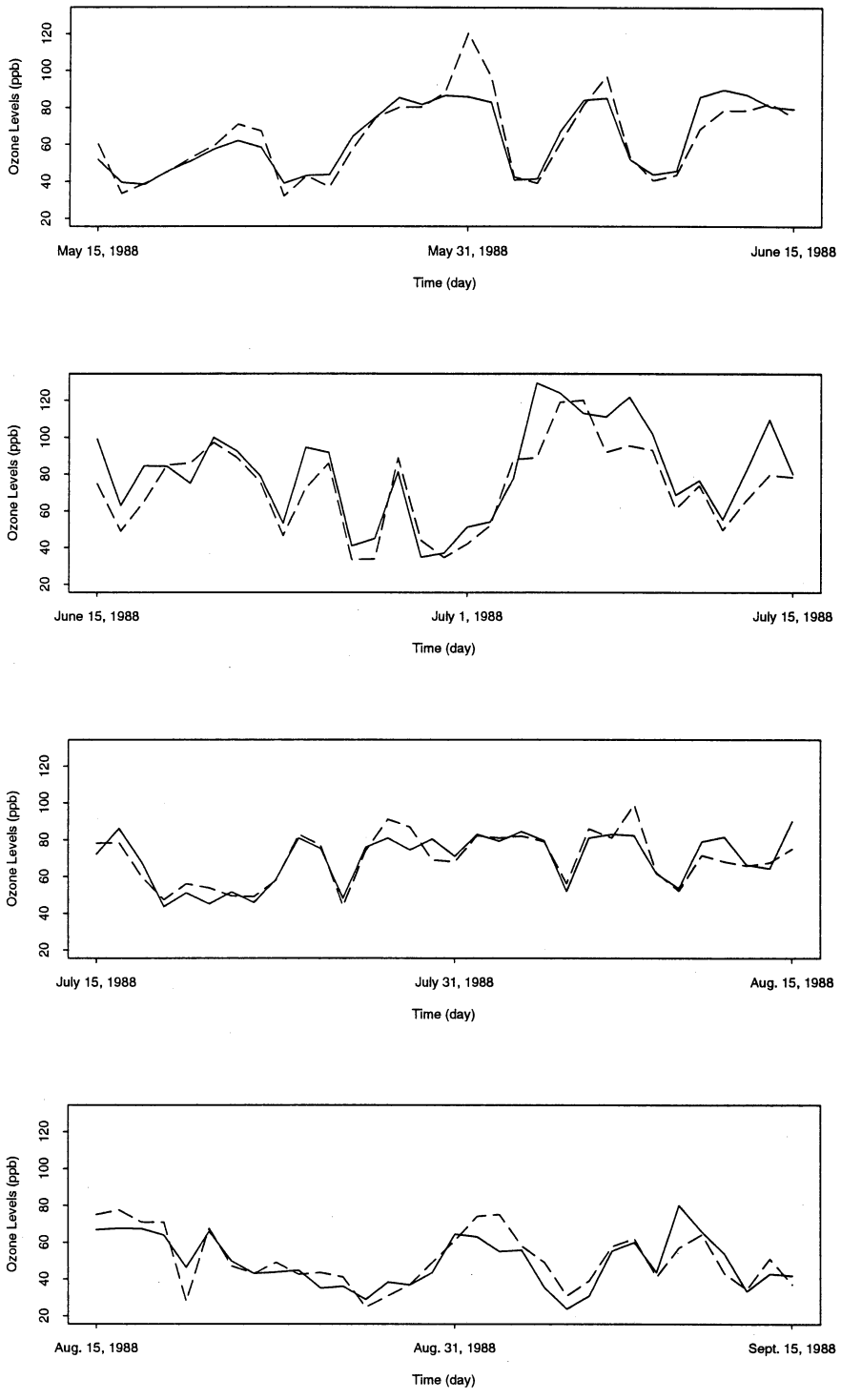



Figure 3.7: Predicted and actual network typical value for 1988. Predictions are made using models based on 1981 - 1987 data. Dashed lines are actual levels, solid lines are predicted levels.

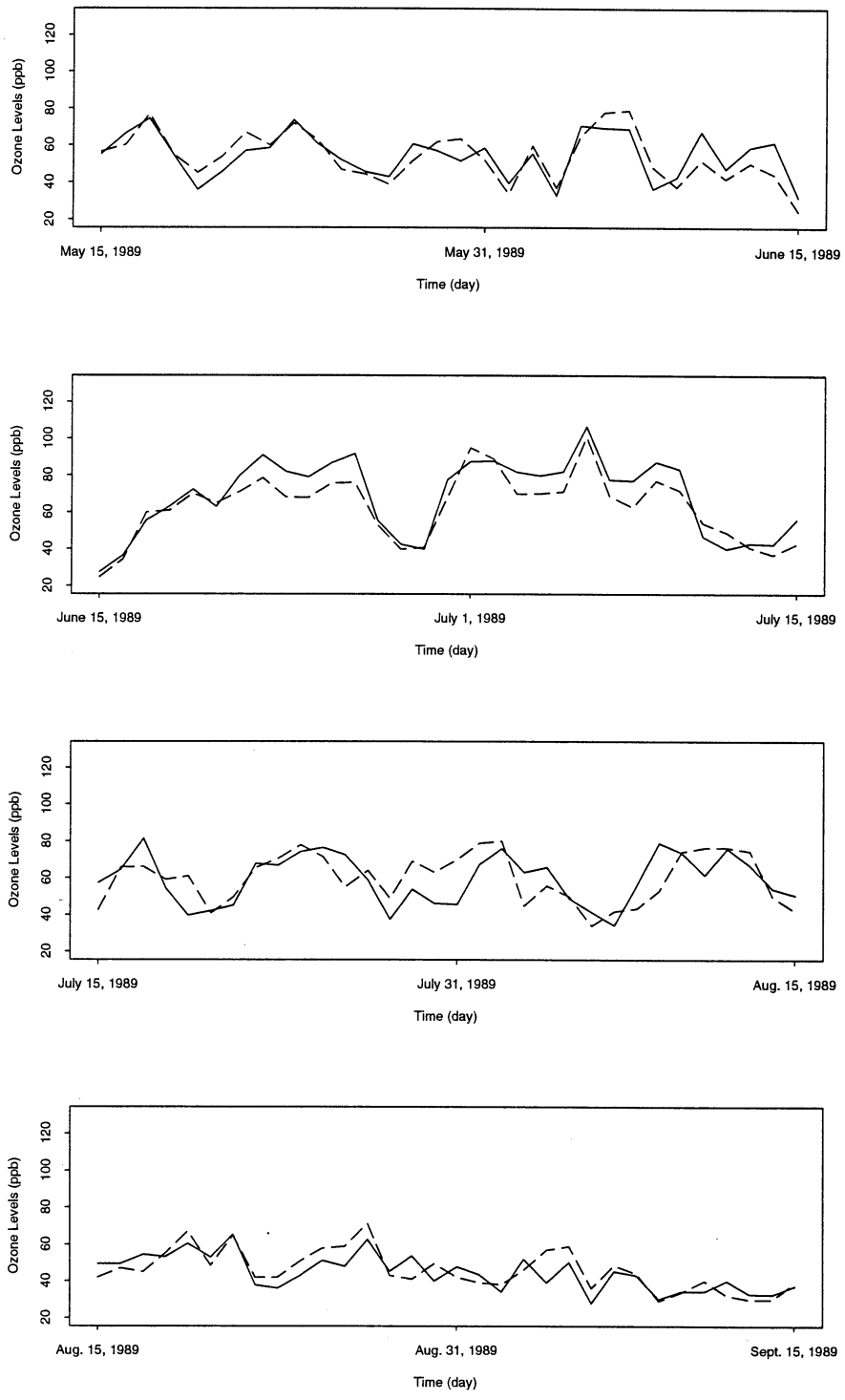



Figure 3.8: Predicted and actual network typical value for 1989. Predictions are made using models based on 1981 - 1988 data. Dashed lines are actual levels, solid lines are predicted levels.

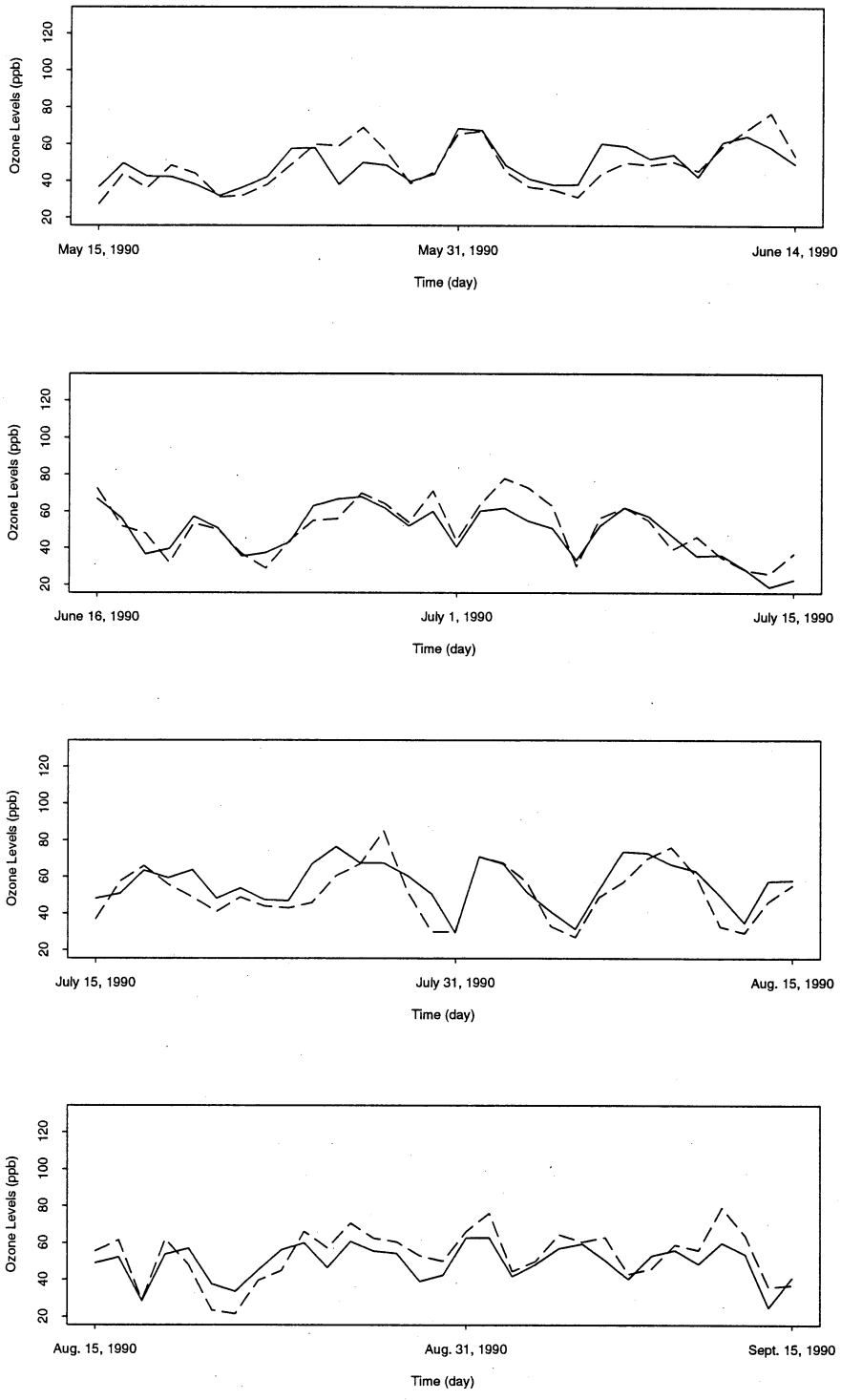



Figure 3.9: Predicted and actual network typical value for 1990. Predictions are made using models based on 1981 - 1989 data. Dashed lines are actual levels, solid lines are predicted levels.

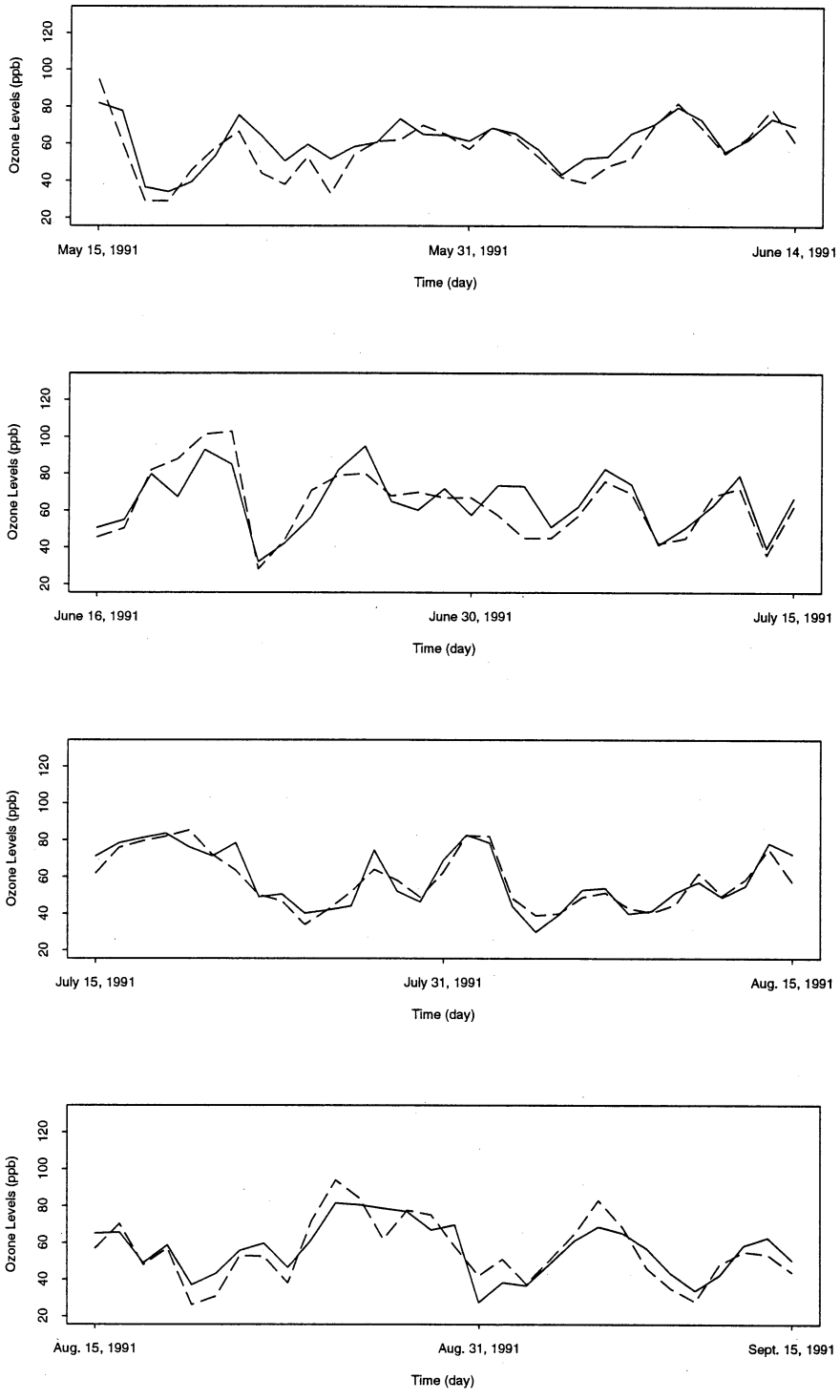



Figure 3.10: Predicted and actual network typical value for 1991. Predictions are made using models based on 1981 - 1990 data. Dashed lines are actual levels, solid lines are predicted levels.

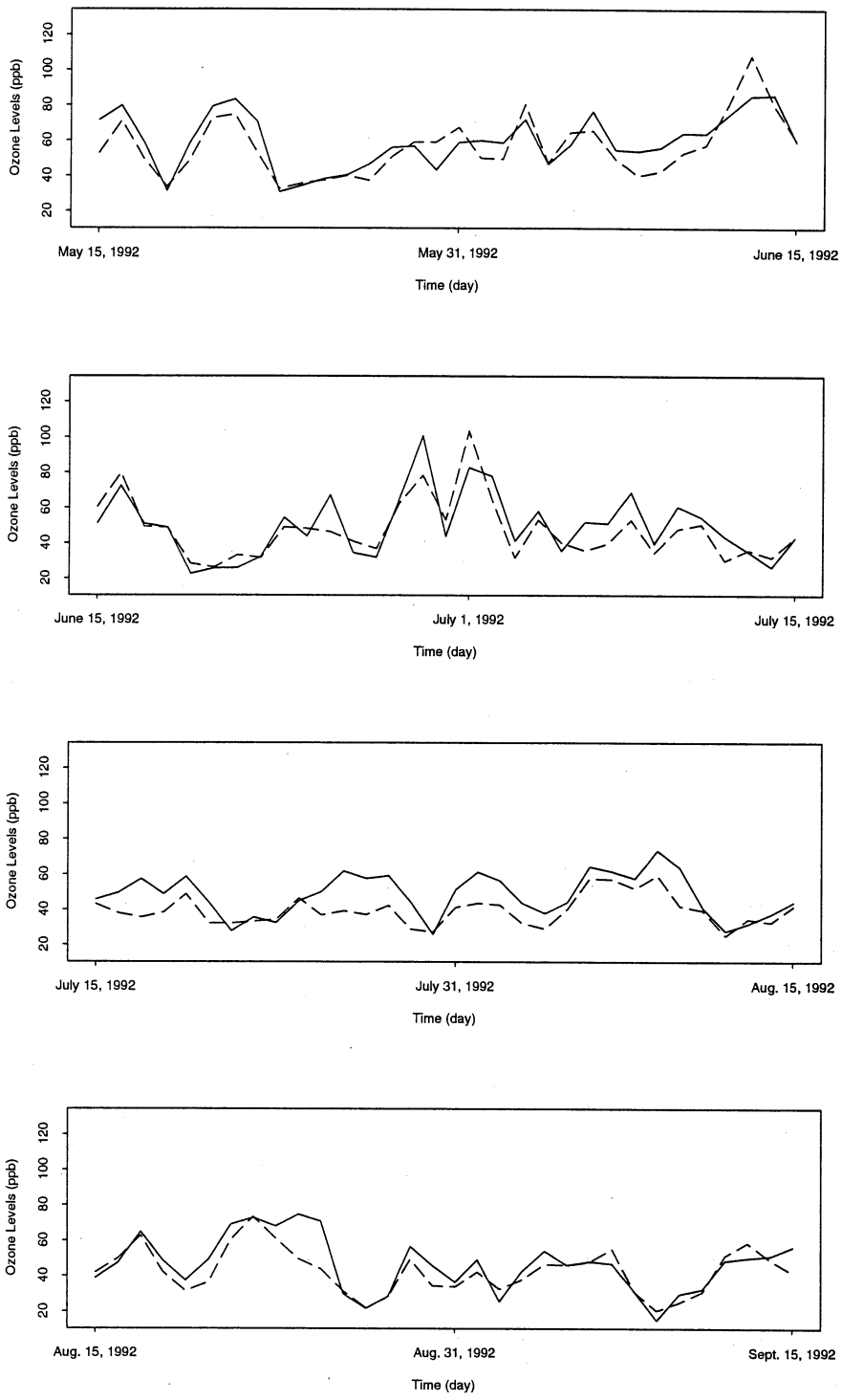



Figure 3.11: Predicted and actual network typical value for 1992. Predictions are made using models based on 1981 - 1991 data. Dashed lines are actual levels, solid lines are predicted levels.

Table 4.2: Estimates of $\sigma_Z$ and $\sigma_\varepsilon$, and CVRMSE for models for the network maximum.

| Models | $\hat{\sigma}_Z$ | $\hat{\sigma}_\varepsilon$ | CVRMSE |
|---|---|---|---|
| May 15 - June 15 | 27.364 | 13.149 | 14.404 |
| June 15 - July 15 | 25.977 | 13.388 | 16.991 |
| July 15 - Aug. 15 | 27.242 | 16.699 | 17.865 |
| Aug. 15 - Sept. 15 | 23.640 | 13.193 | 14.895 |

## 4.2 Quality of the Fitted Models

Figure 4.1 shows the scatter plots of CV predicted network maximum values vs. the actual values. The plots show that the model fits are generally good.

The Q-Q plots in Figure 4.2 show that standardized CV residuals are reasonably close to standard normal.

The MLEs of $\sigma_Z$ and $\sigma_\varepsilon$ and the CVRMSEs are listed in Table 4.2. The table shows that the CVRMSEs and the MLEs of $\sigma_\varepsilon$ for the fitted models are fairly close. These estimates of $\sigma_\varepsilon$ are close to 15.611 ppb, the root mean square residual obtained in the parametric model based on data of 11 years from April 1 to October 31 in Bloomfield et al. (1993). The variability in the network maximum values is greater than in the network typical values.

## 4.3 Trend Estimation

The same methods are used to assess the adjusted trends for the network maximum. Figure 4.3 indicates that using a straight line for the trend is appropriate.

The estimates of slopes and trends and their standard errors are listed in Table 4.3 and Table 4.4.

The tables show that for the periods of June 15 - July 15 and August 15 - September 15, there are significant decreasing trends in the adjusted averages. According to the jackknifed estimates, the adjusted network maximum decreased by about 18.3 ppb over the decade with a standard error of 3.7 ppb for the period of June 15 - July 15, and by about 7.3 ppb over the decade with a standard error of 2.9 ppb for the period of August 15 - September 15. The jackknifed estimates of trend are $-18.9\%$/decade with a standard error of 3.4%/decade for the period of June 15 - July 15, and $-9.2\%$/decade with a standard error of 3.5%/decade for the period of August 15 - September 15.

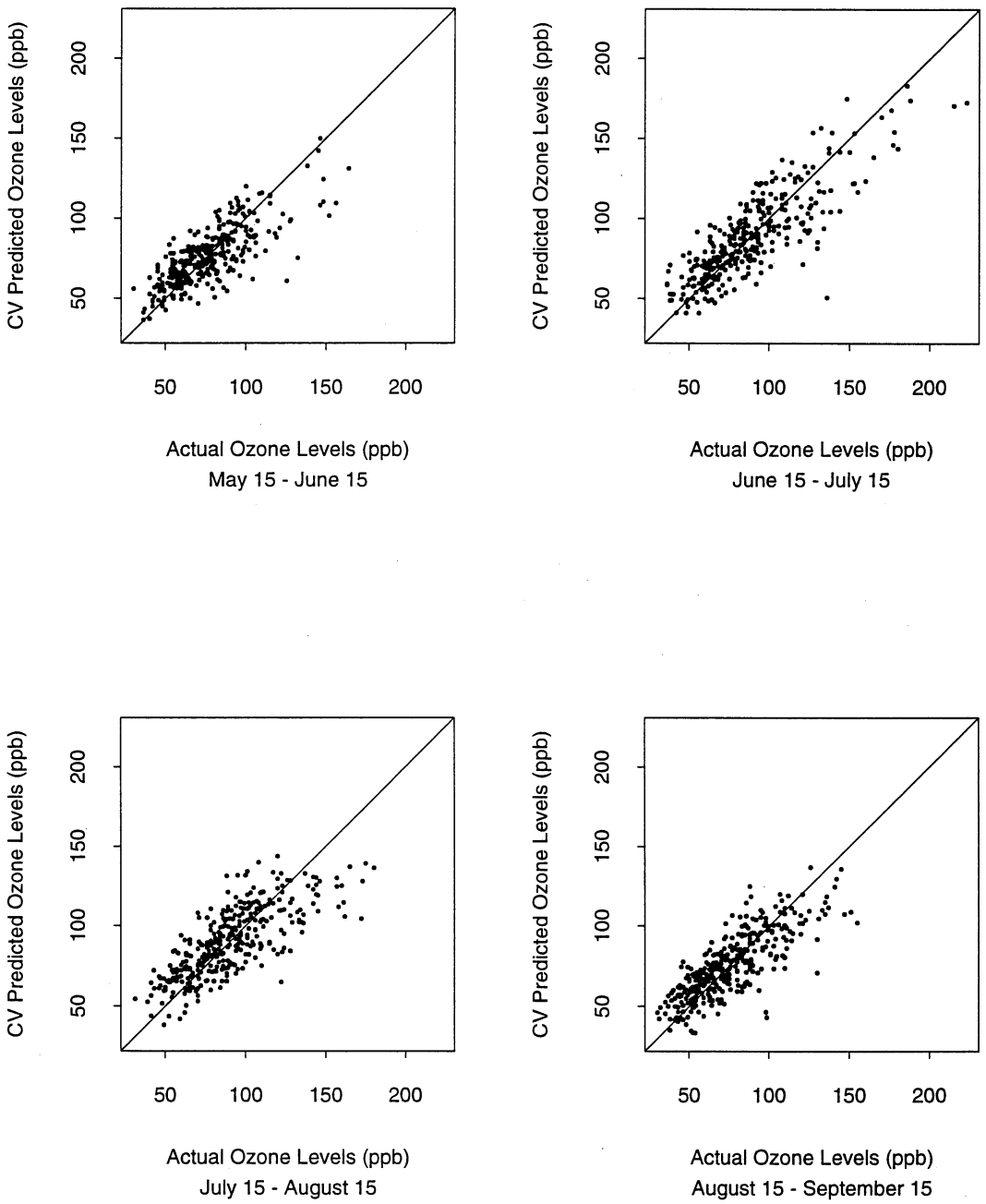



Figure 4.1: CV predicted network maximum values vs. actual values.

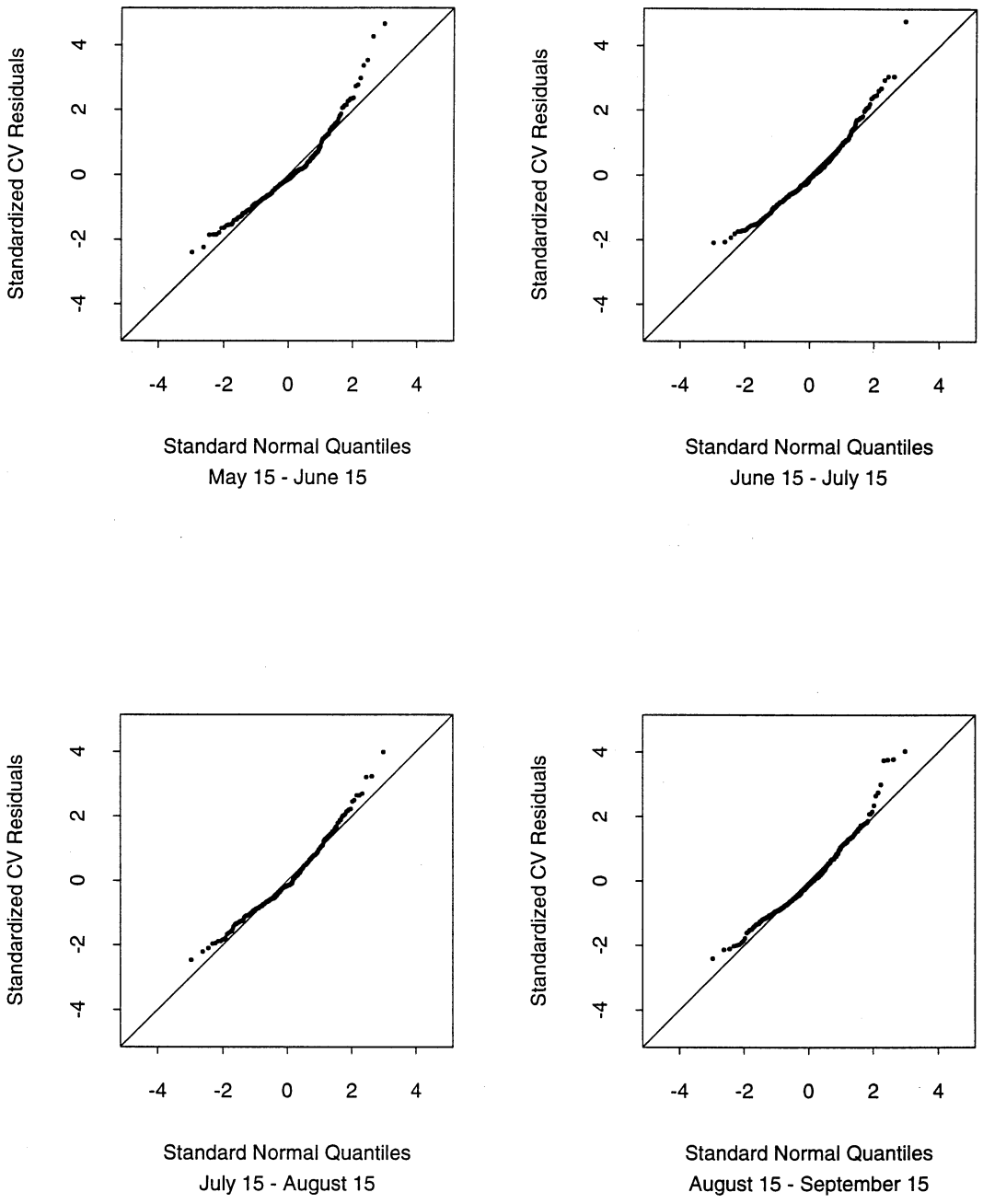



Figure 4.2: Q-Q plots of the standardized CV residuals for the network maximum values vs. standard normal distribution.

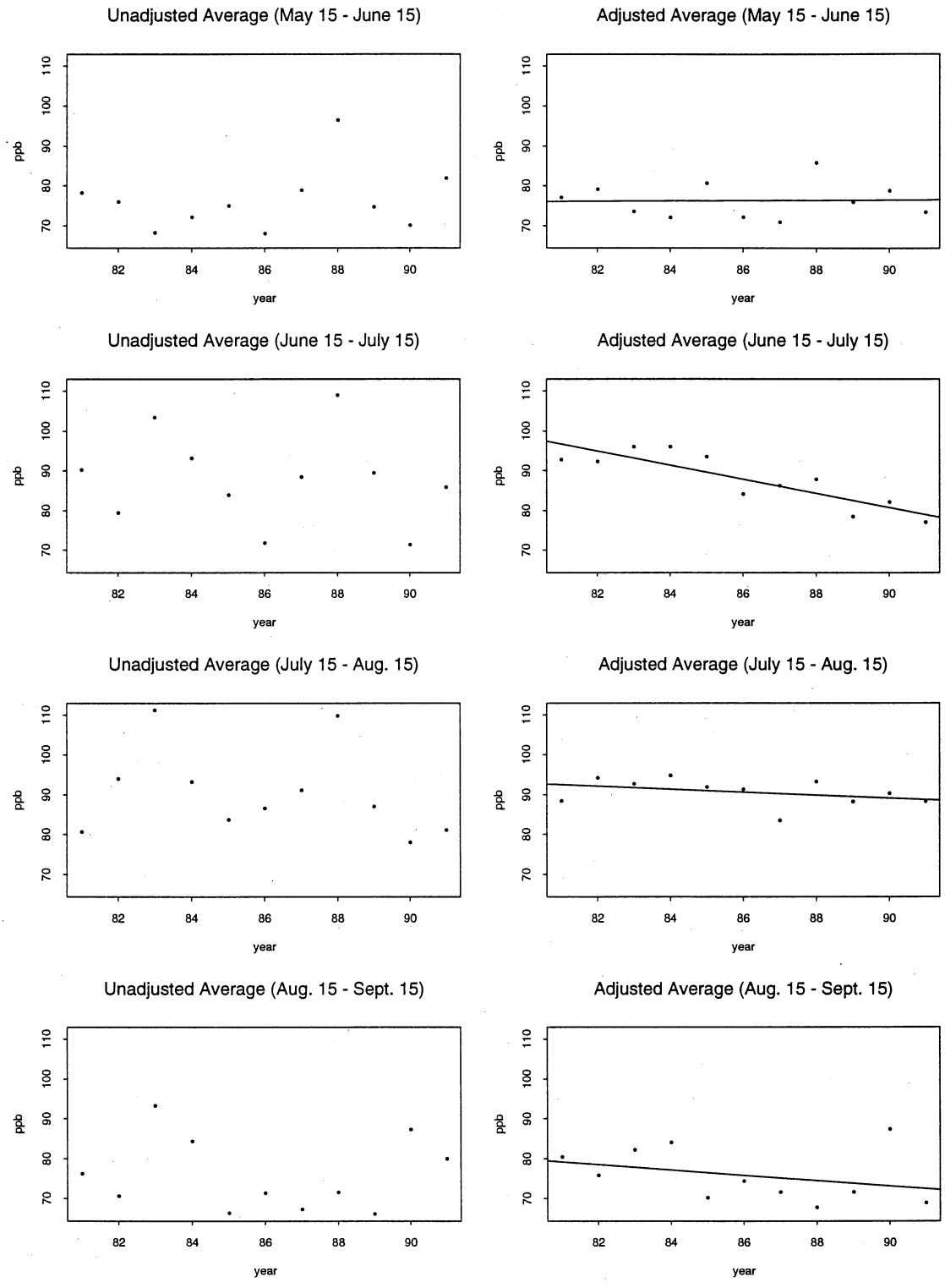



Figure 4.3: Adjusted and unadjusted averages of the network maximum.

Table 4.3: Estimates of slopes and their standard errors for the adjusted averages for the network maximum.

| | Model Estimates | | | Jackknifed Estimates | | |
|---|---|---|---|---|---|---|
| Models | Slope | Standard Error | $t$ Value | Slope | Standard Error | $t$ Value |
| May 15 - June 15 | 0.0524 | 0.2704 | 0.1938 | 0.0639 | 0.2760 | 0.2315 |
| June 15 - July 15 | −1.7798 | 0.3237 | 5.4983 | −1.8337 | 0.3690 | 4.9694 |
| July 15 - Aug. 15 | −0.3706 | 0.3093 | 1.1982 | −0.3575 | 0.3151 | 1.1346 |
| Aug. 15 - Sept. 15 | −0.6678 | 0.2663 | 2.5077 | −0.7255 | 0.2888 | 2.5121 |

Table 4.4: Estimates of trends and their standard errors for the adjusted averages for the network maximum.

| | Model Estimates | | | Jackknifed Estimates | | |
|---|---|---|---|---|---|---|
| Models | Trend | Standard Error | $t$ Value | Trend | Standard Error | $t$ Value |
| May 15 - June 15 | 0.0069 | 0.0357 | 0.1933 | 0.0078 | 0.0364 | 0.2143 |
| June 15 - July 15 | −0.1835 | 0.0411 | 4.4647 | −0.1892 | 0.0344 | 5.5000 |
| July 15 - Aug. 15 | −0.0401 | 0.0337 | 1.1899 | −0.0394 | 0.0334 | 1.1796 |
| Aug. 15 - Sept. 15 | −0.0844 | 0.0362 | 2.3315 | −0.0919 | 0.0351 | 2.6182 |

## 4.4 Predictions

The models can be used for prediction. Figures 4.4 to 4.8 show the results. The predictions are generally a good match to the actual levels.

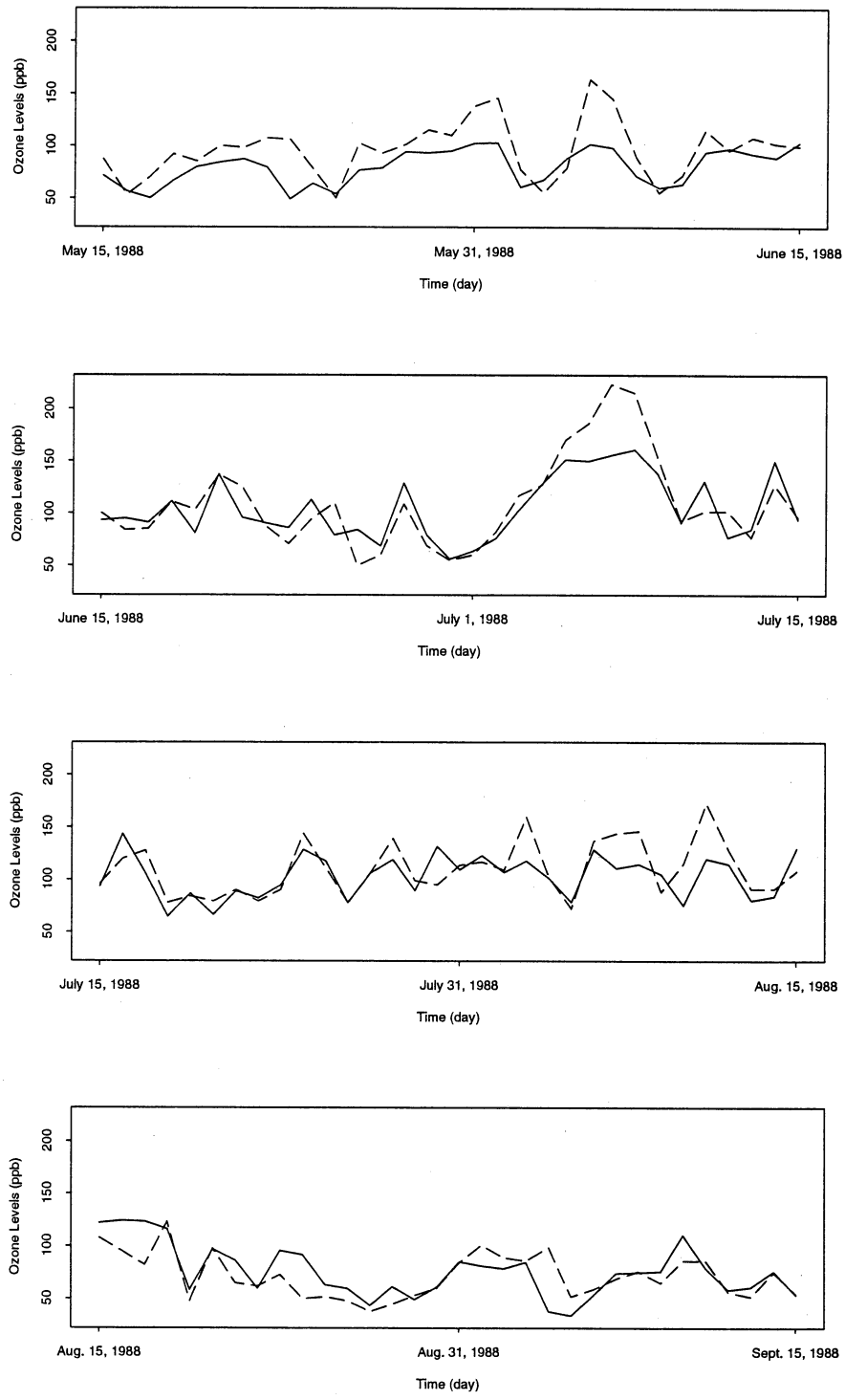



Figure 4.4: Predicted and actual network maximum for 1988. Predictions are made using models based on 1981 - 1987 data. Dashed lines are actual levels, solid lines are predicted levels.

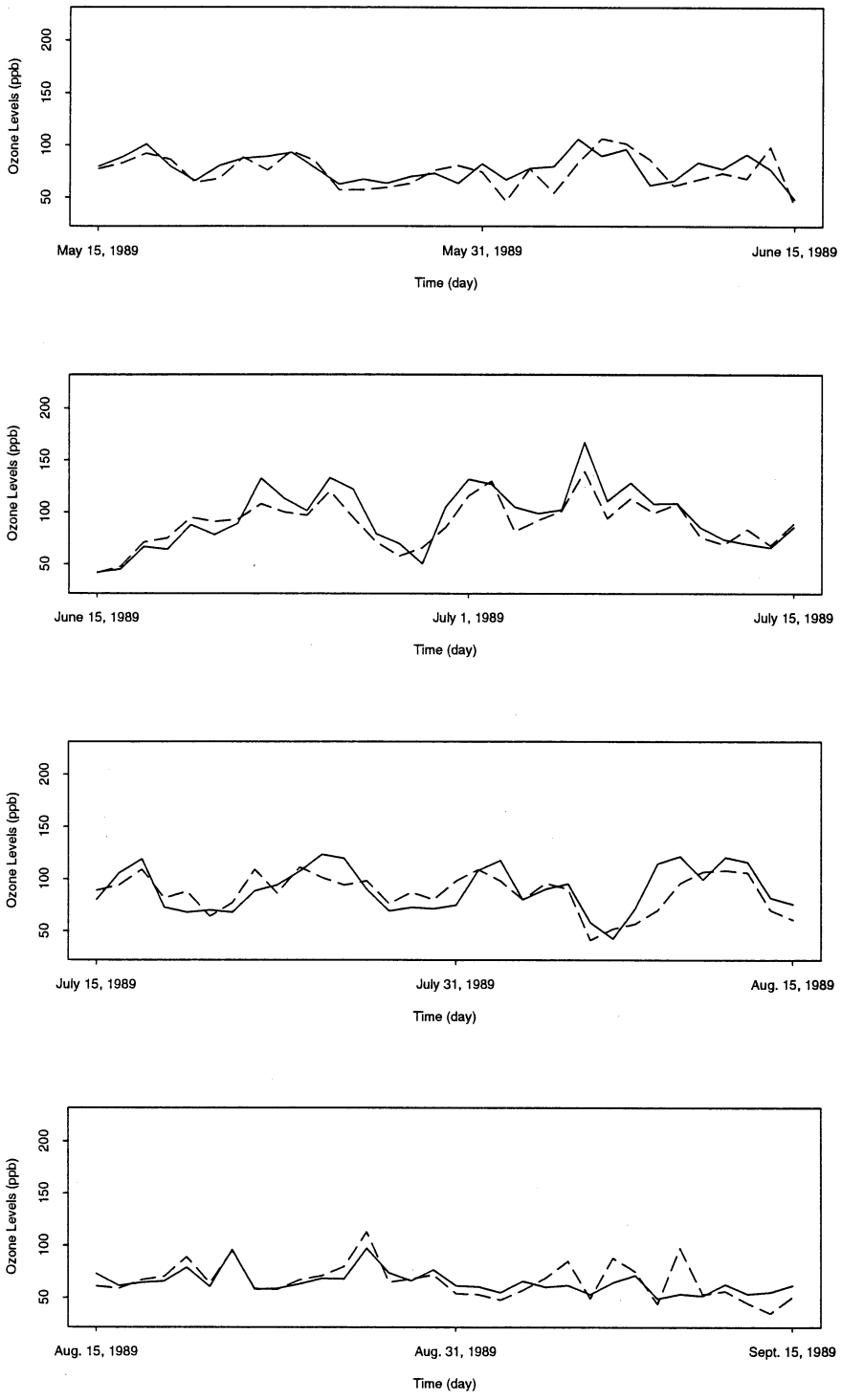



Figure 4.5: Predicted and actual network maximum for 1989. Predictions are made using models based on 1981 - 1988 data. Dashed lines are actual levels, solid lines are predicted levels.

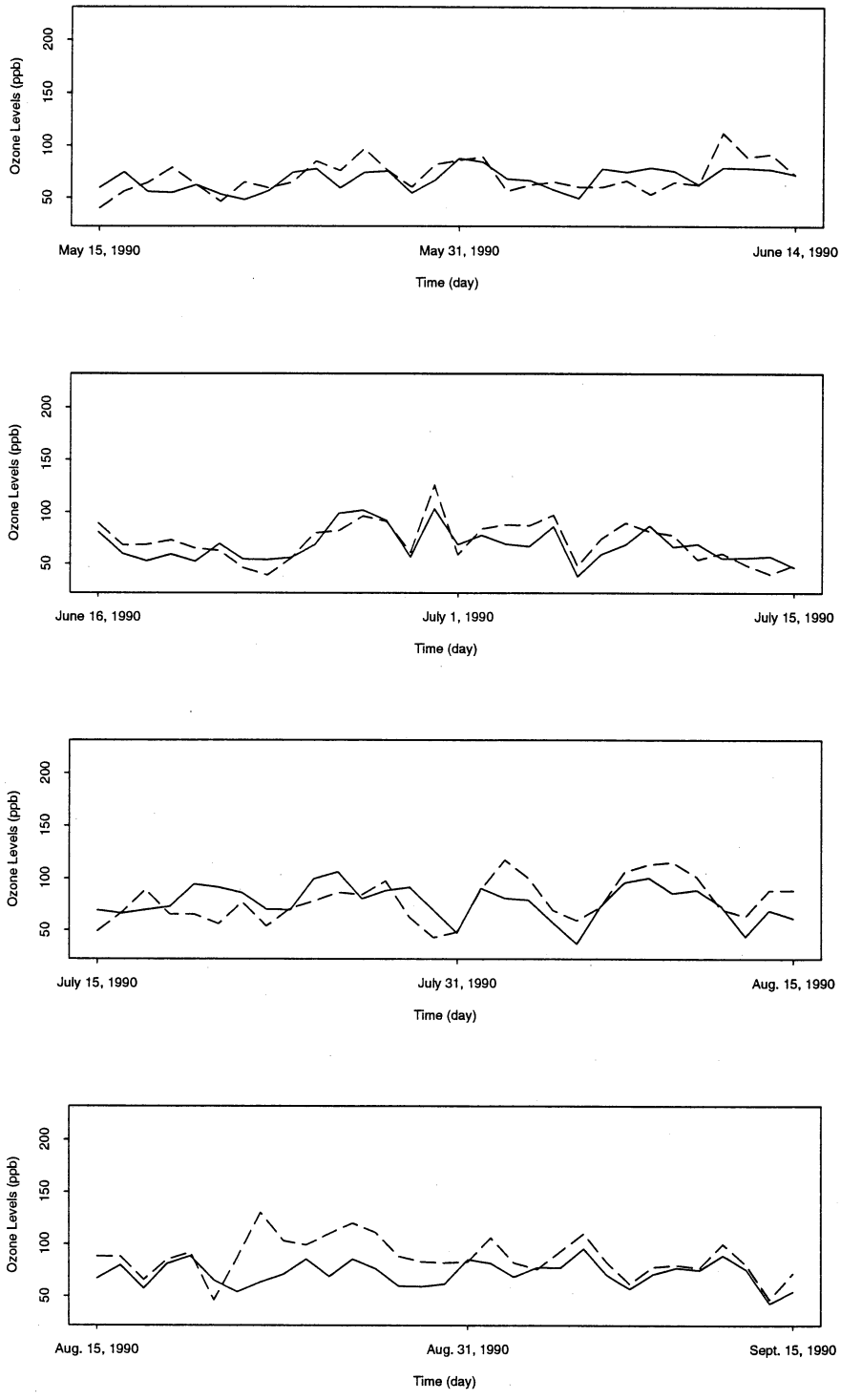



Figure 4.6: Predicted and actual network maximum for 1990. Predictions are made using models based on 1981 - 1989 data. Dashed lines are actual levels, solid lines are predicted levels.

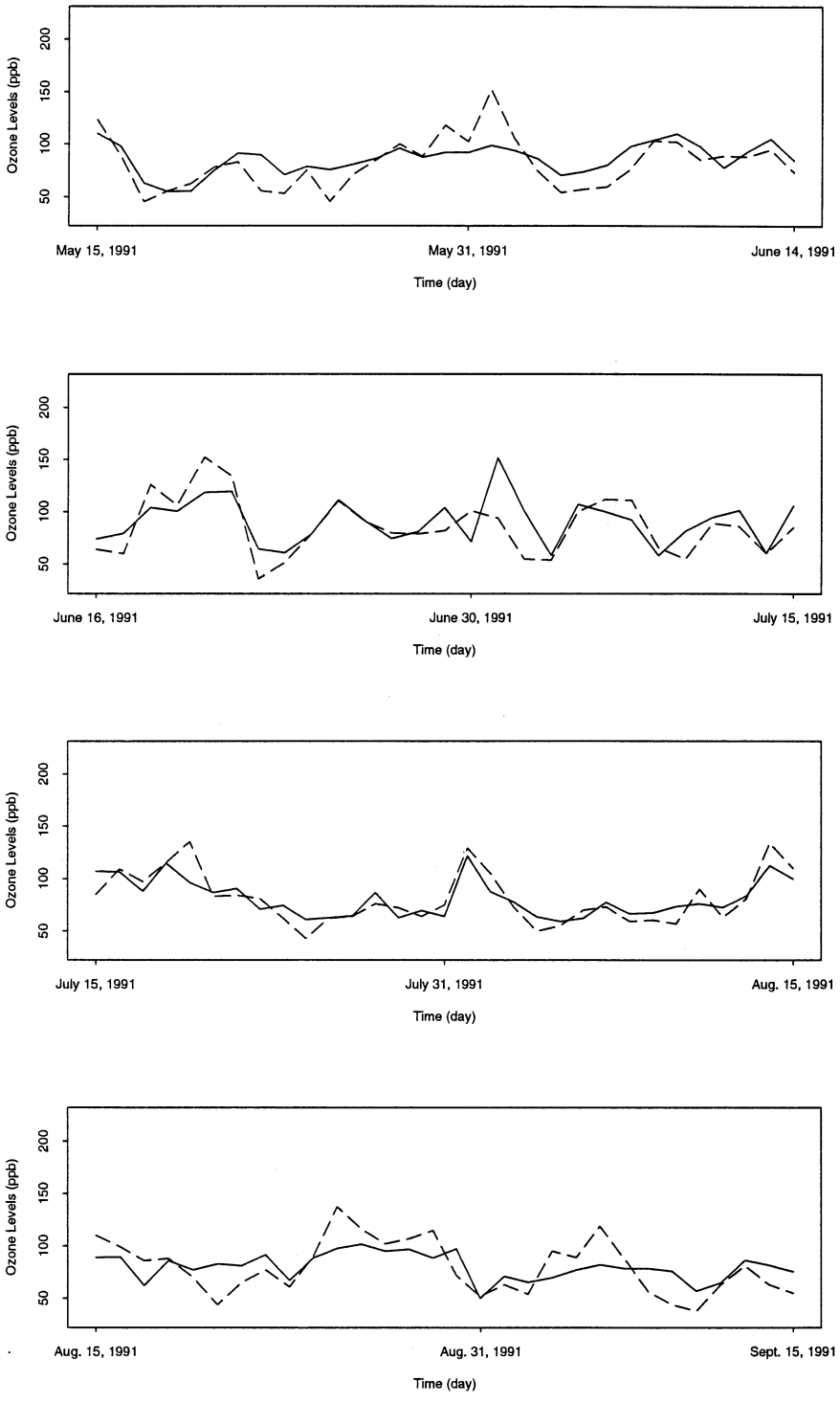



Figure 4.7: Predicted and actual network maximum for 1991. Predictions are made using models based on 1981 - 1990 data. Dashed lines are actual levels, solid lines are predicted levels.

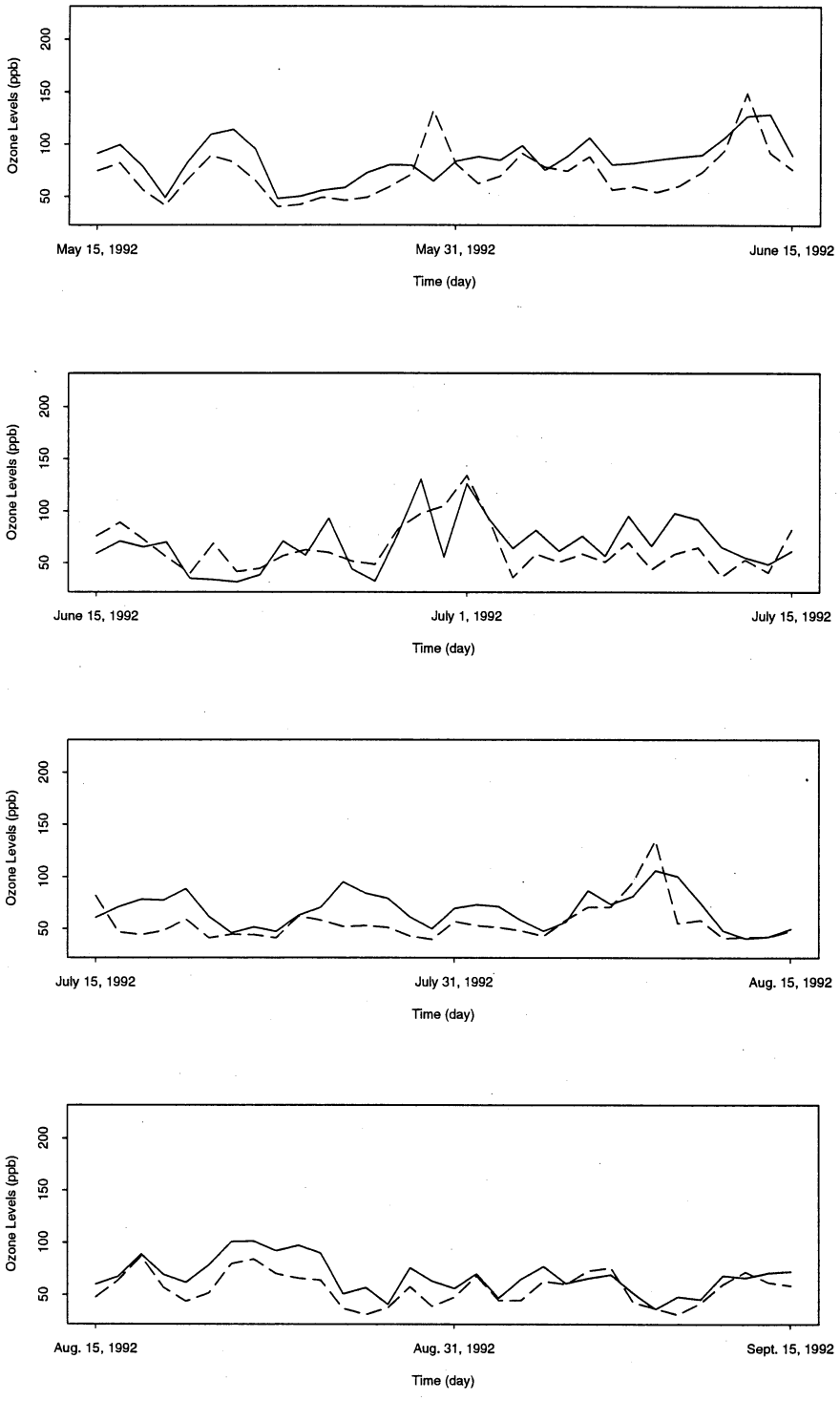



Figure 4.8: Predicted and actual network maximum for 1992. Predictions are made using models based on 1981 - 1991 data. Dashed lines are actual levels, solid lines are predicted levels.

Table 5.1: Estimates of $\theta$'s and $p$'s for models for the network secondary maximum with meteorological variables rescaled.

| | May 15 - June 15 | | June 15 - July 15 | | July 15 - Aug. 15 | | Aug. 15 - Sept. 15 | |
|---|---|---|---|---|---|---|---|---|
| variables | $\theta$ | $p$ | $\theta$ | $p$ | $\theta$ | $p$ | $\theta$ | $p$ |
| maxt | 1.5920 | 2 | 3.2421 | 2 | 1.1811 | 2 | 2.3049 | 2 |
| wspd | 0.3128 | 2 | 2.4552 | 2 | 0.0000 | 2 | 0.0000 | 2 |
| meanu | 0.0000 | 2 | 1.6906 | 2 | 2.4506 | 2 | 2.5651 | 2 |
| meanv | 0.5443 | 2 | 0.3722 | 1.193 | 4.8482 | 2 | 3.7159 | 2 |
| rh | 0.8922 | 2 | 1.0255 | 2 | 0.0000 | 2 | 1.7155 | 2 |
| vis | 0.1545 | 2 | 0.4597 | 1.334 | 0.1225 | 2 | 0.1563 | 2 |
| opcov | 0.0000 | 2 | 0.0440 | 2 | 0.0548 | 2 | 0.0829 | 2 |
| wspd700 | 0.5249 | 2 | 0.0000 | 2 | 0.1591 | 2 | 0.1392 | 2 |
| tlag1 | 0.0000 | 2 | 0.0000 | 2 | 0.0000 | 2 | 0.0000 | 2 |
| tlag2 | 0.0220 | 2 | 0.0000 | 2 | 0.0000 | 2 | 0.2247 | 2 |
| wlag | 0.4508 | 2 | 0.9189 | 2 | 0.0000 | 2 | 0.2095 | 1.337 |
| rhlag | 0.0000 | 2 | 0.0000 | 2 | 0.0000 | 2 | 0.1838 | 2 |
| day | 0.0000 | 2 | 0.1121 | 2 | 0.0355 | 2 | 0.0759 | 2 |

# 5 Modeling the Network Secondary Maximum

Similarly, the same type of model is fitted to the network secondary maximum value with the same meteorology data.

## 5.1 Important Variables

The maximum likelihood estimates of the $\theta$'s and $p$'s with the rescaled meteorological variables are given in Table 5.1.

From the table, it can be seen that temperature, relative humidity and wind (through wspd, wlag, meanu, meanv and/or wspd700) are consistently important across the months (relative humidity is unimportant for the third month).

## 5.2 Quality of the Fitted Models

Figure 5.1 shows the scatter plots of CV predicted network secondary maximum values vs. the actual values. The plots show that the model fits are generally good.

The Q-Q plots in Figure 5.2 show that standardized CV residuals are reasonably close to standard normal.

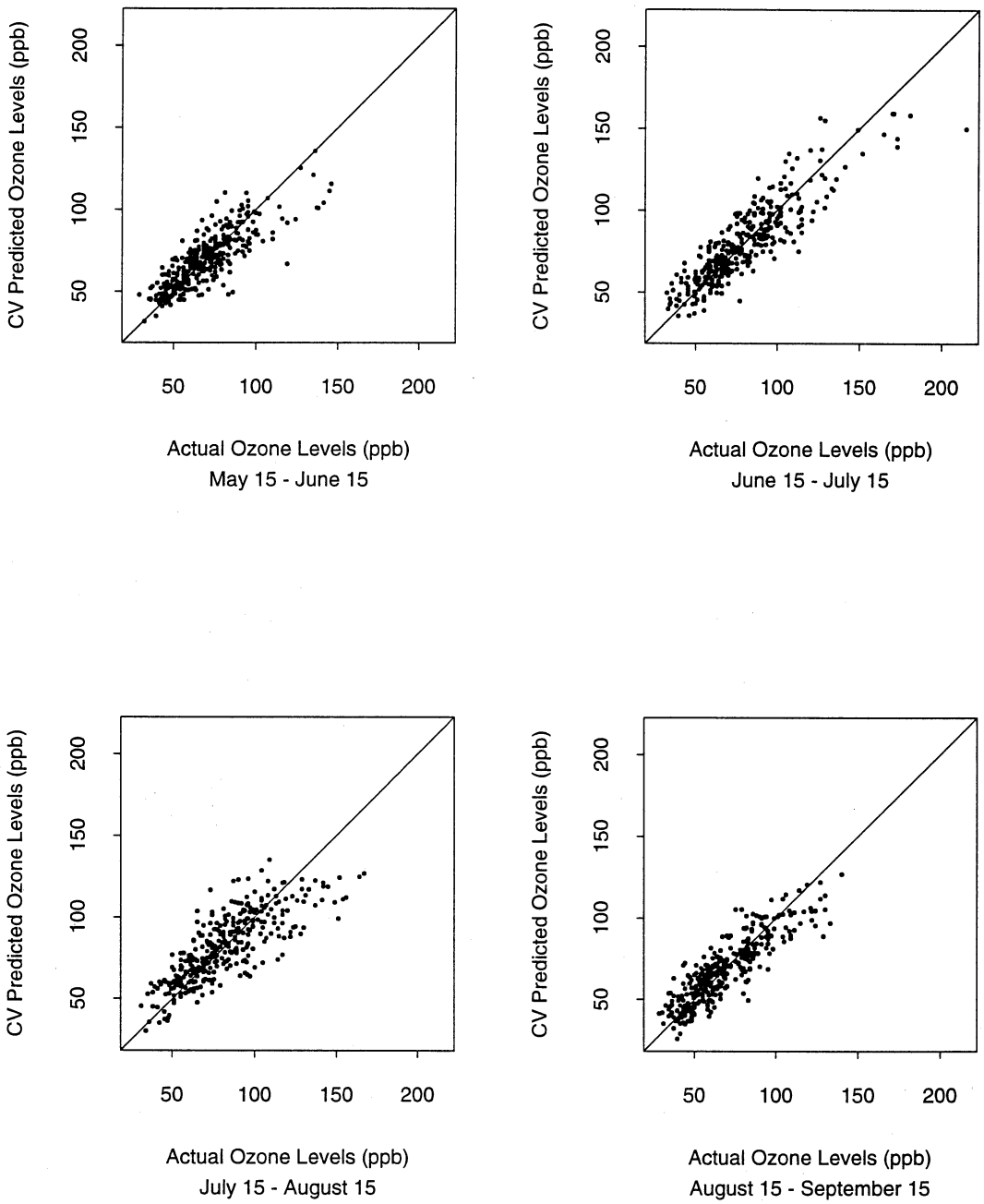



Figure 5.1: CV predicted network secondary maximum values vs. actual values.

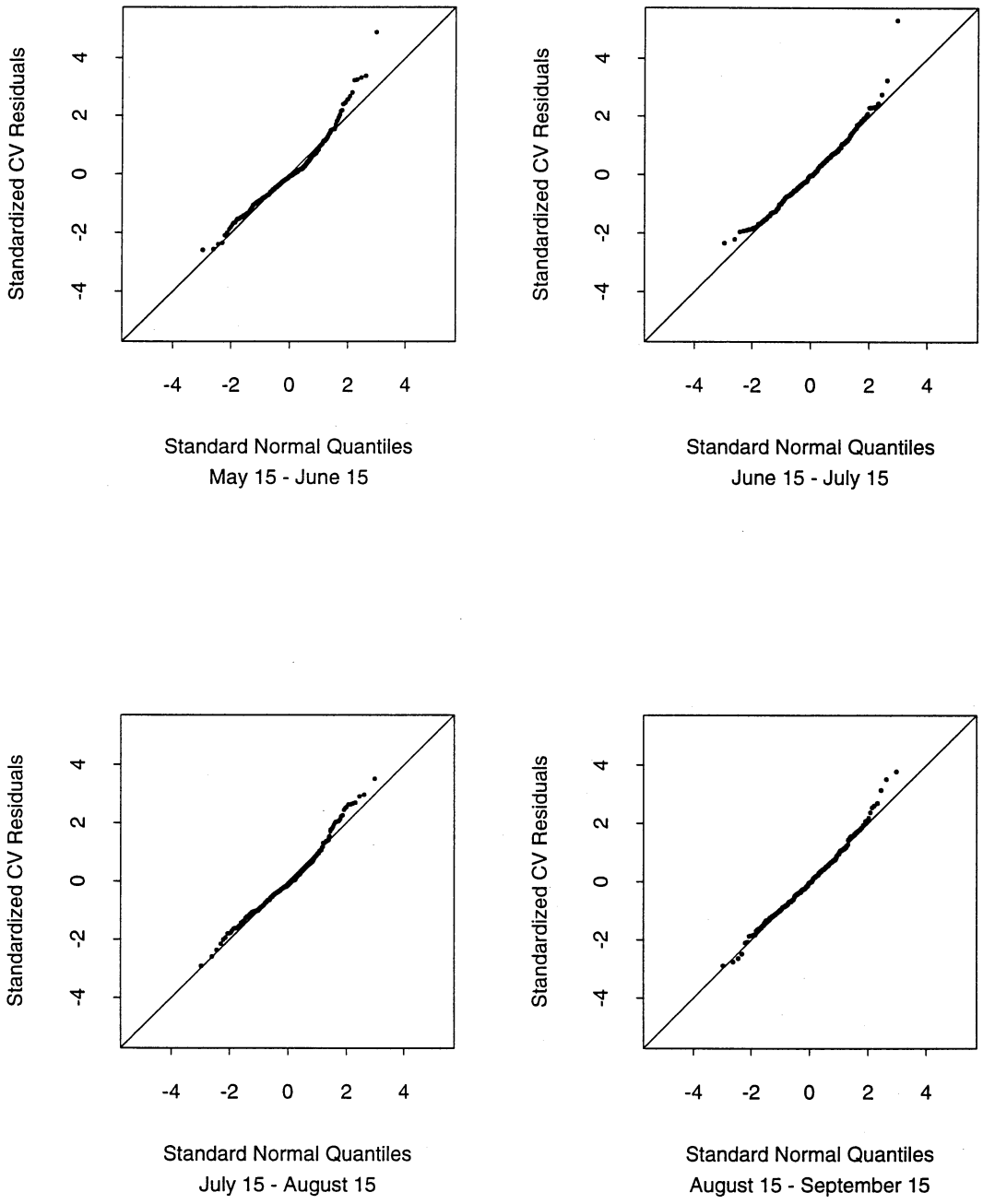



Figure 5.2: Q-Q plots of the standardized CV residuals for the network secondary maximum values vs. standard normal distribution.

Table 5.2: Estimates of $\sigma_Z$ and $\sigma_\varepsilon$, and CVRMSE for models for the network secondary maximum.

| Models | $\hat{\sigma}_Z$ | $\hat{\sigma}_\varepsilon$ | CVRMSE |
|---|---|---|---|
| May 15 - June 15 | 27.629 | 10.696 | 11.592 |
| June 15 - July 15 | 26.512 | 11.179 | 13.048 |
| July 15 - Aug. 15 | 26.059 | 14.424 | 15.387 |
| Aug. 15 - Sept. 15 | 20.774 | 9.479 | 10.889 |

Table 5.3: Estimates of slopes and their standard errors for the adjusted averages for the network secondary maximum.

| | Model Estimates | | | Jackknifed Estimates | | |
|---|---|---|---|---|---|---|
| Models | Slope | Standard Error | $t$ Value | Slope | Standard Error | $t$ Value |
| May 15 - June 15 | −0.0381 | 0.2213 | 0.1722 | −0.0782 | 0.2399 | 0.3260 |
| June 15 - July 15 | −1.0868 | 0.2584 | 4.2059 | −1.0222 | 0.2735 | 3.7375 |
| July 15 - Aug. 15 | −0.3938 | 0.2650 | 1.4860 | −0.3886 | 0.2733 | 1.4219 |
| Aug. 15 - Sept. 15 | −0.8923 | 0.1973 | 4.5226 | −0.9418 | 0.2096 | 4.4933 |

The MLEs of $\sigma_Z$ and $\sigma_\varepsilon$ and the CVRMSEs are listed in Table 5.2. The table shows that the CVRMSEs and the MLEs of $\sigma_\varepsilon$ for the fitted models are fairly close.

## 5.3 Trend Estimation

The same methods are used to assess the adjusted trends for the network secondary maximum. Again, Figure 5.3 shows that a linear trend would be reasonable.

The estimates of slopes and trends and their standard errors are listed in Table 5.3 and Table 5.4.

The tables show that for the periods of June 15 - July 15 and August 15 - September 15, there are significant decreasing trends in the adjusted averages. According to the jackknifed estimates, the adjusted network secondary maximum decreased by about 10.2 ppb over the decade with a standard error of 2.7 ppb for the period of June 15 - July 15, and by about 9.4 ppb over the decade with a

Table 5.4: Estimates of trends and their standard errors for the adjusted averages for the network secondary maximum.

| | Model Estimates | | | Jackknifed Estimates | | |
|---|---|---|---|---|---|---|
| Models | Trend | Standard Error | $t$ Value | Trend | Standard Error | $t$ Value |
| May 15 - June 15 | −0.0055 | 0.0320 | 0.1719 | −0.0119 | 0.0346 | 0.3439 |
| June 15 - July 15 | −0.1275 | 0.0393 | 3.2443 | −0.1209 | 0.0302 | 4.0033 |
| July 15 - Aug. 15 | −0.0469 | 0.0322 | 1.4565 | −0.0469 | 0.0317 | 1.4795 |
| Aug. 15 - Sept. 15 | −0.1233 | 0.0346 | 3.5636 | −0.1301 | 0.0270 | 4.8185 |

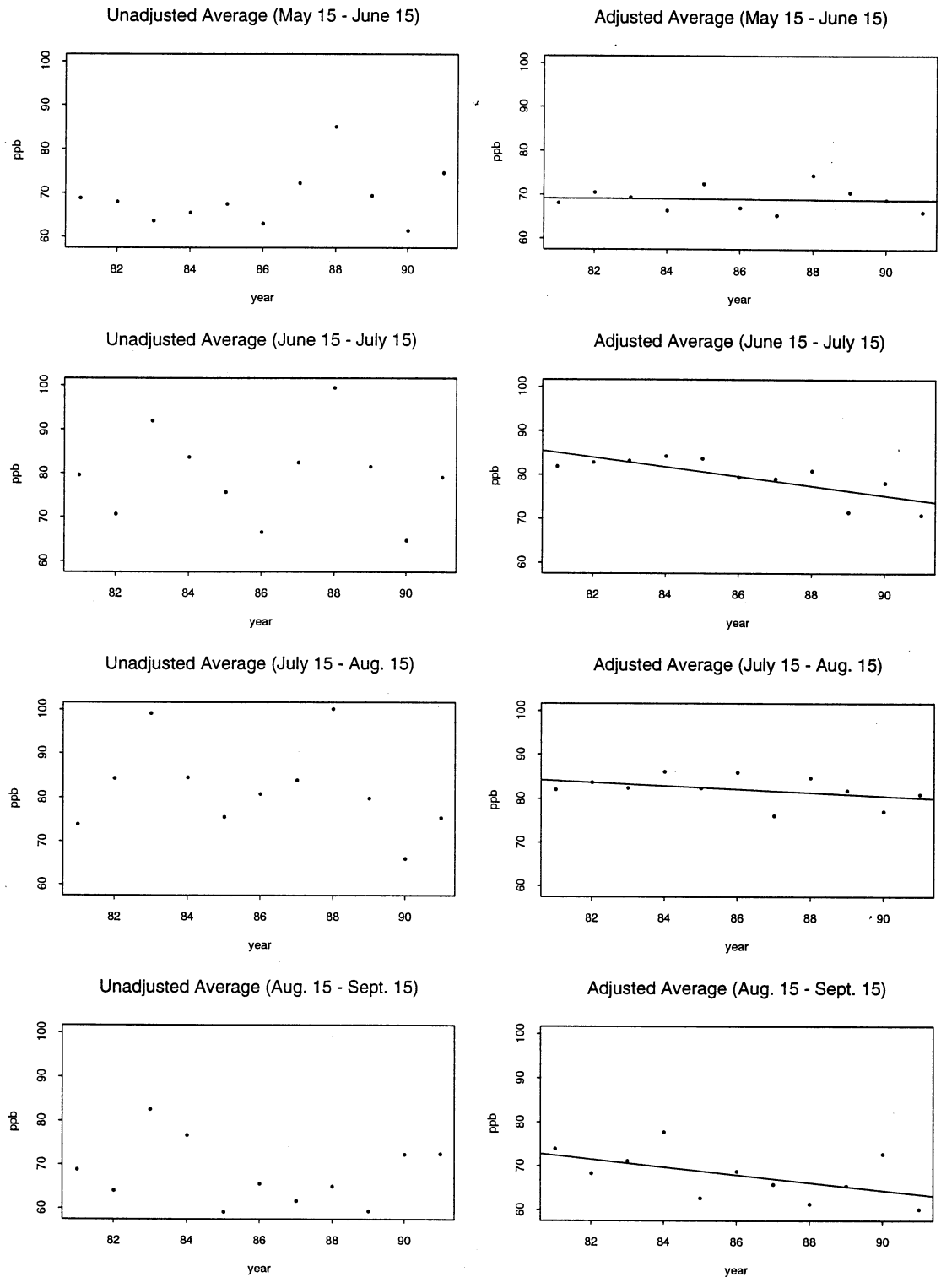



Figure 5.3: Adjusted and unadjusted averages of the network secondary maximum.

standard error of 2.1 ppb for the period of August 15 - September 15. The jackknifed estimates of trend are $-12.1\%$/decade with a standard error of 3.0%/decade for the period of June 15 - July 15, and $-13.0\%$/decade with a standard error of 2.7%/decade for the period of August 15 - September 15.

### 5.4 Predictions

The models can be used for prediction. Figures 5.4 shows the results.

## 6 Conclusions

The semiparametric modeling technique described in this report is shown to provide a good way to model the ozone concentration as a function of meteorology. This method can be used to assess that part of the trend in ozone levels that cannot be accounted for by meteorology. The models can also be used to predict ozone levels from meteorology.

It is found that for the periods of May 15 - June 15 and July 15 - August 15, there is no significant trend in ozone concentration after adjusting for meteorology for Chicago. However, there are significant downward trends for ozone after adjusting for meteorology for the periods of June 15 - July 15 and August 15 - September 15. For the period of June 15 - July 15, the adjusted trend for the network typical/average values is $-6.5\%$/decade with a standard error of 2.9%/decade, for the network maximum values is $-18.9\%$/decade with a standard error of 3.4%/decade and for the network second maximum values is $-12.1\%$/decade with a standard error of 3.0%/decade. For the period of August 15 - September 15, the adjusted trend for the network typical values is $-11.5\%$/decade with a standard error of 2.9%/decade, for the network maximum values is $-9.2\%$/decade with a standard error of 3.5%/decade and for the network secondary maximum values is $-13.0\%$/decade with a standard error of 2.7%/decade.

In Bloomfield et al. (1993), for the period of April 1 - October 31, the adjusted trend for the network typical values is found to be $-2.7\%$/decade with a (jackknife) standard error of 3.4%/decade, and for the network maximum values the adjusted trend is found to be $-9.5\%$/decade with an (adjusted) standard error of 1.8%/decade. Jackknife standard error for the adjusted trend for the network maximum values was not computed in Bloomfield et al. (1993).

If the trend estimates reported here are averaged over the four periods, they are broadly consistent with the parametric analysis in Bloomfield et al. (1993). By dividing the ozone season into periods, however, we are able to detect significant trend in the network typical value within two of the four periods. Similarly, our analysis suggests that the significant trend in network maximum value reported by Bloomfield et al. (1993) is largely due to a reduction in two of the four periods. The dependence of trend on period suggested by our analysis is intriguing and should be investigated further.

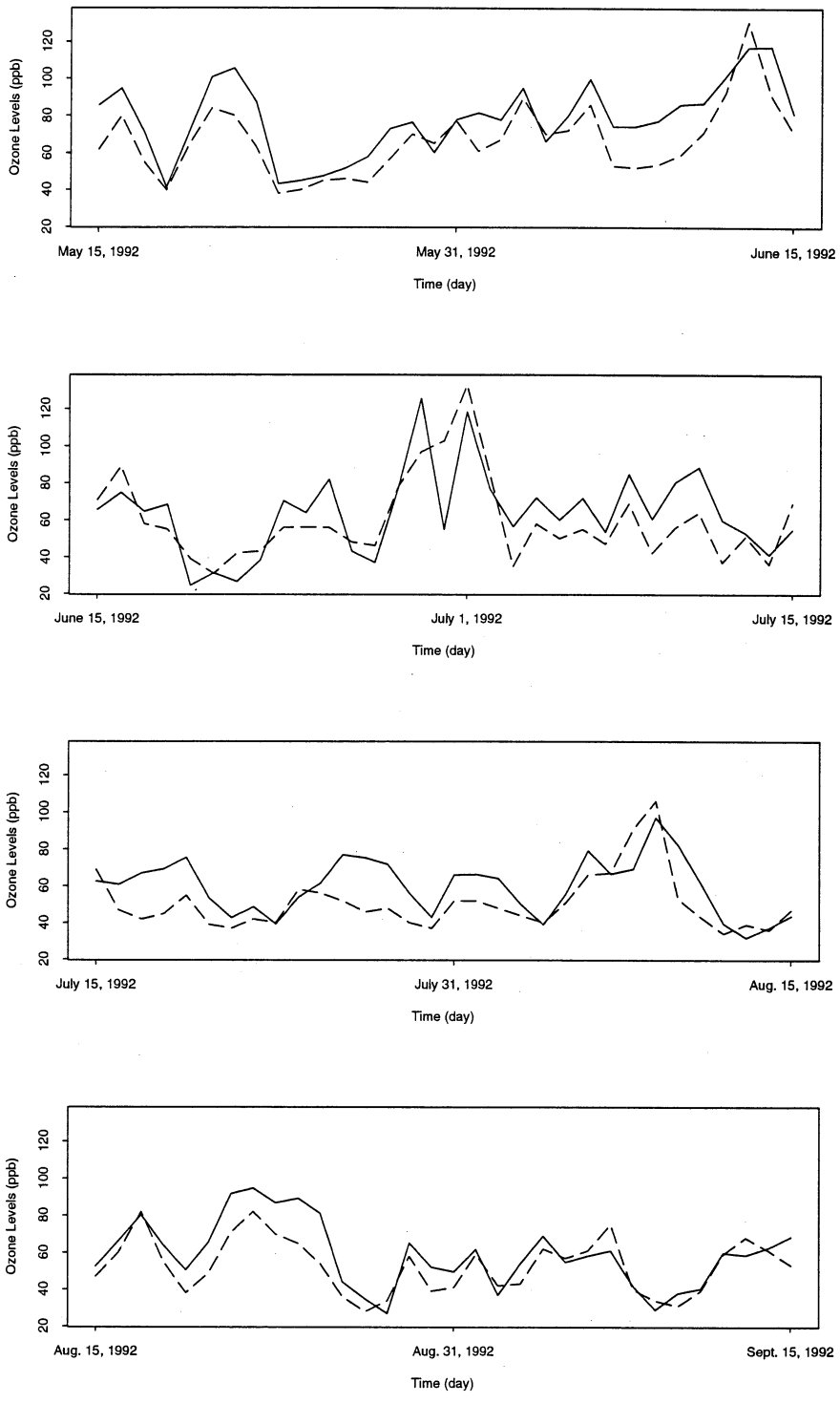



Figure 5.4: Predicted and actual network secondary maximum for 1992. Predictions are made using models based on 1981 - 1991 data. Dashed lines are actual levels, solid lines are predicted levels.

## References


Bloomfield, P., Royle, A. and Yang, Q. (1993). Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends, *Technical Report No. 1*. National Institute of Statistical Sciences, P.O.Box 14162, Research Triangle Park, NC 27709-4162.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to The Theory of Statistics*, McGraw-Hill Book Company, New York.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*, Addison-Wesley, Reading, Massachussetts.

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

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and Analysis of Computer Experiments, *Statistical Science* **4**: 409–435.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachussetts.