5-2004

Development of ozone forecast models for selected Kentucky metropolitan areas.

Yiqiu Lin 1971-
University of Louisville

DEVELOPMENT OF OZONE FORECAST MODELS FOR SELECTED KENTUCKY METROPOLITAN AREAS

By

Yiqiu Lin
B.S., University of Tianjin, China, 1993

A Thesis
Submitted to the Faculty of the
Speed Scientific School of the University of Louisville
In Partial Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Department of Mechanical Engineering
University of Louisville
Louisville, KY

May 2004

DEVELOPMENT OF OZONE FORECAST MODELS FOR SELECTED KENTUCKY METROPOLITAN AREAS

Submitted by: ______________________

Yiqiu Lin

A Thesis Approved on

January 29, 2004

by the following Reading and Examination Committee:

____

Thesis Director, Dr. W. Geoffrey Cobourn

____

Dr. John Lilly

____

Dr. Christopher Richards

ACKNOWLEDGEMENTS

First of all, special thanks go to my advisor, Dr. W. Geoffrey Cobourn, for his time used in guiding me carrying out the research and finishing this thesis. I would like to thank Leonel Martinez, and all those helped me collect the data and set up the databases. Also thanks to Tom Priddy at University of Kentucky, and the other individuals who provided the air quality data and meteorological data. Thanks to the other members of my committee, Dr. John Lilly and Dr. Christopher Richards. Finally, I would like to thank my parents Shipei Lin and Caiyun Huang, the rest of my family, and the most important person in my life, my wife Dongmei Zhang. They supported me with the greatest love. I dedicate this document to them.

ABSTRACT

DEVELOPMENT OF OZONE FORECAST MODELS FOR SELECTED KENTUCKY METROPOLITAN AREAS

Yiqiu Lin

May 14, 2004

Ground-level ozone forecast models were developed for the following middle and small metropolitan areas in Kentucky: Ashland, Bowling Green, Owensboro, and Paducah. These models were nonlinear regression models, based on models previously developed for Louisville and Lexington. For each of the four cities, the mean absolute errors (MAE) for the model estimates, based on the 1998-2002 model-fitted data sets, were less than 7.7 ppb; the MAE/$\overline{O_3}$ were less than 12.7%. The models could explain at least 66% of the variance of the daily peak ozone. On average, the errors of the model were within ±15.0 ppb on 88% of days, and were within ±10.0 ppb on 73% of days. Using an alarm threshold 80 ppb, the detection rates for National Ambient Air Quality Standard (NAAQS) Exceedences ranged from 0.48 to 0.67 for the four cities. The corresponding false alarm rates ranged from 0.29 to 0.44. The results of this study demonstrate that the ozone forecast models for each of the four cities can be expected to be useful tools for making next-day forecasts of local ground-level O_3 in those areas. Similar models, updated using 2003 data, will be used during the 2004 O_3 season for providing daily automated forecasts for these metropolitan areas.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Ozone is an odorless, colorless gas composed of three atoms of oxygen. While the "good ozone" in the stratosphere forms a layer that protects life on earth from the sun's harmful ultraviolet rays, in the earth's lower atmosphere, ground-level ozone is an air pollutant that causes human health problems, damages crops and other vegetation.

In the United States, high levels of ground-level ozone are currently responsible for most violations of the National Ambient Air Quality Standards (NAAQS). Since 1984, several major pollution control strategies toward the reduction of ozone precursor emissions have been enacted. As part of Acid Rain Program, the NO_x Program enacted by Environmental Protection Agency (EPA) was implemented in two phases, beginning in 1996 and 2000 (EPA, 2002). The first stage of the program was designed to reduce annual nitrogen oxide compounds (NO_x) emissions in the United States by over 400,000 tons per year between 1996 and 1999 (Phase I), and by approximately 1.17 million tons per year beginning in the year 2000 (Phase II). The second stage aimed to increase the reductions to 2.1 million tons per year, beginning in 2000. In October, 1998, EPA finalized the "Finding of Significant Contribution and Rulemaking for Certain States in the Ozone Transport Assessment Group Region for Purposes of Reducing Regional Transport of Ozone", commonly called the NO_x SIP Call. The NO_x SIP Call was designed to mitigate significant transport of NO_x, one of the precursors of ozone (EPA, 2003a). In many urban areas, motor vehicle inspection and maintenance programs designed to reduce the emissions of NO_x and volatile organic compounds (VOC) were instituted. Nationally, the Clean Air Act Amendments of 1990 mandated increasingly stringent rules to reduce car and truck tailpipe emissions.

The successful implementations of the ozone reduction strategies have reduced the concentrations of NO_x, VOC and ozone in recent years. During the period 1982 through 2001, nitrogen dioxide concentration decreased 24 percent; ozone dropped 11 percent based on daily 4^{th} maximum 8-hr average concentration (EPA, 2003b). To improve all the nonattainment areas to attain compliance with the NAAQS for ozone is still a tough problem, especially since EPA announced a much more stringent 8-hr ozone standard in 1997.

Before 1997, the sole NAAQS for ozone was based on 1-hr average concentration, not to exceed 0.12 parts per million by volume (ppm). In July 1997, based on scientific studies showing that prolonged exposure to ozone levels at concentrations well below the 0.12 ppm standard causes adverse health effects in children and in healthy adults engaged in outdoor activities, EPA promoted a new NAAQS for ozone, based on 8-hr average concentration, not to exceed 0.08 ppm. The new 8-hr standard is much stronger and more protective than the 1-hr standard. As of 1998, in the United States 51 million people lived in ozone nonattainment areas based on the 1-hr standard, while 130 million people lived in ozone nonattainment areas based on the 8-hr standard (Lin, 2001). Based on 2003 data, there were approximately 110 million persons living in the areas where the 8-hr standard had been exceeded (EPA, 2003c).

The ozone air pollution problem in the Louisville area has improved substantially over the past 20 years. On June 18, 2001, EPA announced that the Louisville area had met the health-based 1-hr ozone standard (EPA, 2003d). The 8-hr standard is expected to be applied in 2004. Based on data from 2001 to 2003, Louisville will not meet the new 8-hr ozone standard. To issue alerts in anticipation of high ozone levels so that community action can be taken on a voluntary basis to reduce the emissions of ozone forming compounds, also to provide the notices for the sensitive individuals to make plans to reduce outside activities at ozone action days, in 1997, a hybrid nonlinear regression (NLR) ozone forecast model for Louisville was developed at the U of L. In 2000, based on the Louisville model, a NLR ozone forecast model for Lexington was developed at U of L. These models have been updated each year. The model fit for Lexington is approximately as good as the fit for the Louisville model. For example, for the period 1998-2002, the mean absolute error (MAE) of the model fit was 6.02 ppb and the overall correlation coefficient (R^2) was 0.925. The model fit for the Lexington model over the same period, resulted in an MAE of 8.19 ppb and an overall R^2 of 0.855. For each of these models, the MAE for the forecasts was about 15% of the corresponding summertime mean daily O_3 peak.

Due to the successful implementation of the NLR model for Louisville and Lexington, developing models to be used for other cities in Kentucky is possible. But each community is unique: local ozone pollution is affected in part by the local and regional emissions, climate, and land-use. Louisville and Lexington are respectively the largest and second largest metropolitan areas of Kentucky. Compared to the other cities, there is much more population, traffic and economic activity.

The objective of this thesis was to develop models to provide 24-hr forecasts of the daily peak 8-hr average ozone concentration in additional metropolitan areas in Kentucky where ozone concentrations are of concern. The areas are Ashland-Huntington, Bowling Green (Mammoth Cave area), Owensboro-Evansville and Paducah. In 2003, no areas in KY were denoted as ozone nonattainment by the 1-hr ozone standard (EPA, 2003e). But based on the data of 2000, 2001, 2002, all of the areas above would be designated as nonattainment areas by the 8-hr ozone standard. The 8-hr standard is scheduled to be imposed in 2004, and it is expected that several of these areas will be designated as nonattainment areas. Accurate ozone forecast models can provide these areas a better chance to meet the new standard and provide advanced warning of potentially unhealthful air quality for the people living in these areas.

CHAPTER II

LITERATURE REVIEW

The ground-level ozone pollution problem has been noticed for a long time. Many ozone forecast models are described in the literature, such as multiple linear regression models, artificial neural network models, linear stochastic models, photochemical models, etc. There are certain advantages and disadvantages associated with each type, but there are only a few direct comparisons of accuracy in forecasting ozone concentration reported in the literature.

A. Multiple Linear Regression and Nonlinear Regression Models

Multiple regression is a mathematical technique commonly used in air pollution forecasting. The general purpose of multiple regression (the term was first used by Pearson, 1908) is to learn about the relationship between several independent or predictor variables X_n and a dependent or criterion variable Y. In linear regression, the regression procedure yields a linear equation of the form:

$$Y = a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n \tag{1}$$

The standard method for determining the coefficients is ordinary least squares. With this method, the regression procedure will compute a line so that the squared deviations of the observed scattered points from that line are minimized.

In order to compare neural network models with linear regression models for prediction of peak ozone in Houston, Prybutok et al. (2000) built a simple multiple regression model and a NN model. The preliminary regression model included 9 meteorological and ozone precursor parameters, but the final model only used four of them, which were ozone concentration at 9:00 a.m., maximum daily temperature, average nitrogen dioxide concentration and average surface wind speed between 6:00 a.m. and 9:00 a.m. The correlation coefficient R^2 of this model was 0.47. The NN model contained one input layer with 9 input variables, 1 hidden layer and 1 output layer with one variable. The Root mean square (RMS) errors were given as 31.2 and 16.4 ppb for the regression and NN models respectively.

When the regression equations of the models include some nonlinear functional forms, such as exponential, logarithmic, and power functions, the regression model is nonlinear. Bloomfield et al. (1996) developed a nonlinear regression model to predict the long-term ozone trend in the Chicago area. Up to 12 meteorological variables were used in this model, viz. maximum temperature, wind speed, wind direction, relative humidity, specific humidity, dew point temperature, total cloud cover, opaque cloud cover, ceiling height, barometric pressure, visibility, and height of pressure layer. The final R^2 of this model was 0.8042. Based on the 1981-1991 ozone and meteorological data, the model predictions were within ±5 ppb about half the time, and within ±16 ppb about 95% of the time. The overall RMS of the cross-validated prediction errors was 8.3 ppb. Bloomfield's model revealed that the meteorological data accounted for at least 50% of the variance of the ozone concentration. This model demonstrated the validity of nonlinear regression for predicting ozone.

B. Neural Network Model.

Artificial neural networks are collections of mathematical models that emulate some of the observed properties of biological nervous systems and draw on the analogies of adaptive biological learning. Artificial neural network models are computer programs that are designed to emulate human information processing capabilities such as knowledge processing, speech, prediction, classifications, and control. These models have the potential to describe highly non-linear relationships such as the relationship between ozone concentration and meteorological parameters. So several neural network models for ozone forecasting have been developed and have been proved to be useful and cost-effective.

The neural network (NN) model developed by Spellman (1999) had only three predictive parameters (meteorological and air quality parameters) and two hidden layers, used to predict the ozone concentrations of five selected cities of the United Kingdom, with typical summertime mean ozone concentrations about 30-40 ppb. Model results for the five selected cities were given as the mean absolute error (MAE) which ranged from 4.74 ppb to 9.30 ppb, the average error, or bias (-2.02 to 0.76 ppb), the ratio of MAE to mean ozone (0.12-0.24), and R^2 (0.28-0.60).

The finite impulse response (FIR) NN model developed by Balaguer et al. (2002) was applied to make 1-day ahead predictions of ozone concentrations in eastern Spain. Comparing the model predictions with the actual observed ozone concentrations for the 120 days from 1996 to 1999, the results for three sampling sites ranged from 6.39 to 8.8 ppb for MAE and from 0.73 to 0.79 for R^2. The NN model developed by Elkamel et al. (2001) was applied to predict the ozone levels around a heavily industrialized area in Kuwait; the MAE was 12.5% of the mean observed O_3; the model developed by

Narasimhan et al. (2000) was applied to Tulsa, Oklahoma. The only statistic reported was R^2 equal to 0.88.

An innovative neural network model was developed by Wang et al., City University of Hong Kong (2003). This model combines the adaptive radial basis function network with statistical characteristics of ozone to predict the daily maximum ozone concentrations in selected specific areas. In predicting ozone concentrations of Tsuen Wan, Hong Kong for the entire year 2000, the MAE was equal to 11.8 ppb, which was about 47% of the annual mean observed O_3 of 25.1 ppb.

C. Photochemical Model.

Photochemical air quality models play an important role in both scientific investigation of how pollutants evolve in the atmosphere as well as developing policies to manage air quality (Russell and Dennis, 2000). Two main types of photochemical models are Eulerian models and Lagrangian models. Early in 1973, Reynolds et al. created the Urban Airshed Model, which was an Eulerian model, evaluating episodes and air pollution control measures. In 1982, the European monitoring and evaluation programme sulfur model (EMEP sulfur model) was applied by Eliassen et al. to analyze ozone concentration. This model is a Lagrangian type. After that, many more photochemical models were developed to provide ozone trend analysis and ozone prediction. Most of them were of the Eulerian type, such as the Long Term Ozone Simulation Model, Regional Eulerian Model with 3 chemistry schemes, SARMAP air quality model, and Community Multi-scale Air Quality Model (Russell and Dennis, 2000).

The regional Eulerian model with 3 chemistry mechanisms (REM3) is a photochemical transport model. It has been applied operationally to forecast ozone since

1997 at the Freie University, Berlin (Flemming et al., 2001). The vertical resolution of the model is based on three dynamically changing layers. The chemical mechanism CBM4 is used in the model. It consists of 36 species and includes 83 reactions (Gery et al., 1989). In the evaluation made by Flemming et al. (2001) by inputting ozone data over Germany from 1997 to 1999, the correlation coefficient spread from about 0.9 at the first day of forecast period to about 0.77 at the third day of the forecast period. The REM3 model can forecast the large scale ozone situation successfully. The disadvantage is that it underestimates the ozone concentration when the ozone level is low.

Another Eulerian model was a photochemical grid model that was employed to analyze two ozone episodes in autumn (2000) and winter (2001) seasons in the Kaohsiung, Taiwan (Chen et al., 2003). CAMx-2.0 was used in this model, which is a three-dimensional, Eulerian photochemical-transport grid model. Meteorological conditions, such as wind field, temperature, pressure, relative humidity, and period of sunshine, were collected as input data. The results for the autumn episodes (R^2=0.865) and for the winter episodes (R^2=0.886) were reasonably good.

An example of a Lagrangian model is the Lagrangian photochemical box model. It was developed by Wotawa et al. (1998) to provide ozone forecasts. This model consists of up to 8 vertical and up to 5 horizontal boxes. It simulates emission, chemical reactions, horizontal diffusion, vertical diffusion, dry deposition, wet deposition and synoptic scale vertical exchange. Model input data include trajectory and other meteorological data. The model was applied in Vienna, Austria. The results from 1995 showed that the model always underestimates the ozone. The overall median bias was -12.3 ppb. But because it predicted the highest and lowest ozone concentration successfully, this photochemical box model was used to analyze the transport of ozone towards the Alps (Wotawa et al., 2000).

D. Other Ozone Forecast Models.

Slini et al. (2002) described a stochastic autoregressive integrated moving average (ARIMA) model for predicting the maximum ozone concentration in Athens, Greece. The Box-Jenkins approach was applied in this model. The first-order autoregressive, first-order differences moving average models ARIMA (1, 1, 1) and ARIMA (1, 1, 0) were constructed based on different mathematical model. Based on model fit on 1999 data, ARIMA (1, 1, 1) model gave the MAE as 5.78 ppb and R^2 as 0.94; ARIMA (1, 1, 0) model gave the MAE as 10.42 ppb and R^2 as 0.83. The corresponding mean observed O_3 in 1999 was 69.4 ppb.

Linear Stochastic models utilize the time series of ozone concentration to form an equation representing the trend of ozone concentration. In this model, variation in ozone concentration is decomposed into a trend, seasonal cycle, and stochastic elements. It has been shown to be a better model than simple persistence model (McCollister, 1975), but worse than univariate ARIMA model (Robeson and Steyn, 1990).

A generalized additive model (GAM) was created for ozone forecasts in Houston, TX (Davis et al., 1999) because other models, such as nonlinear regression model have been proved to be inappropriate for Houston. GAMs are statistical models based on the loess smoothing functions (Cleveland and Devlin, 1988) and generalized additive models (Hastie and Tibshirani, 1990), which are the data smoothing methods used to find the relationship between the dependent variable and the explanatory variables in a linear regression. In the examined years 1988 and 1991 in Houston, the root-mean-square prediction error for the 8h average forecasts ozone ranged from 13.2 to 16.3 ppb with R^2 ranging from 0.66 to 0.73 for the individual stations and from 18.5 to 22.0 ppb with R^2 ranging from 0.61 to 0.68 for daily domain peak ozone (Davis et al., 1999). The GAM models are effective when the relationship between the variable is expected to be complex and non-linear, and the data should express an appropriate functional form (Greenwell, 2000).

The classification regression tree (CART) algorithm was utilized in a pilot program to forecast ozone in Baltimore, Maryland (Ryan, 1994). It demonstrated skill at distinguishing strong and weak ozone cases but could not accurately predict high ozone events. Compared to the regression analysis in a same case, the CART analysis was characterized by poor correlations with observations and high standard error (23 ppb). Gardner and Dorling (2000) suggested that although linear regression and NN models are better at predicting ozone accurately, CART models are more readily physically interpretable.

The statistical model of ground-level short term ozone pollution (SMOGSTOP) software program was used in Belgium (Lissens et al., 2000). SMOGSTOP was constructed as an empirical model, applying a methodology called stratified pattern matching to link meteorological and precursor information into ozone forecasts. Input data were wind vector, temperature, pressure, humidity and precipitation.

The Simplified Ozone Modeling System (SOMS) created by Venkatram et al. (1994) was used in Baltimore, Maryland to generate long-term ozone predictions (Vukovich et al., 2001). SOMS is a semi-empirical model that can estimate quantitative effects of precursor emission control on ozone. It is based on the concept that ozone can be represented as a function of essentially three variables: concentrations of NOx and VOC, and the time over which the chemical species are exposed to sunlight to produce ozone. The result of the model using 3 years raw data had the bias as 1.9 ppb, MAE as 12.5 ppb and R^2 as 0.81.

In recent years, fuzzy logic has begun to be applied to the ozone forecast problem. In the simplified ozone model developed by Ryoke et al. (2000), fuzzy rule generation methodology was used to represent numerous results of the European Monitoring and Evaluation Program (EMEP) model. The results in this paper showed that the fuzzy model provided better predictions of ozone than the linear regression model. The correlation coefficient between predictions by the fuzzy model and the EMEP ozone model was 0.811, greater than the correlation coefficient between linear regression model and EMEP model, 0.6708. Jorquera et al. compared the performance of several ozone forecast models in Santiago, Chile (Jorquera et al, 1998). Compared to the time series model and NN model, the fuzzy model had the least root mean square error (RMSE), which ranged from 18.7 to 33.3 for different data sets. It can be expected that fuzzy logic approach will play a more important role in ozone forecasting in the future.

E. Model Comparison Studies.

Comrie (1997) developed NN models and multiple regression (MR) models to compare their performance in eight selected cities. The meteorological input data were daily maximum temperature, average daily dew point temperature, average daily wind speed, and daily total sunshine. The models were compared by using the unlagged data and lagged data respectively. In the lagged data, a lagged ozone concentration (typically a value from the previous day) was used as an additional predictor variable. Unlagged data did not include this variable. A total of 690 observations were used for each of the eight cities. The subset of 440 observations was used to develop ozone forecast models and the other subset of 250 observations was used as a quasi-independent subset for model testing. Comparison statistics for the 250 observations were given for the eight cities. The average observed ozone concentrations ranged from around 40 to 66 ppb for the eight cities. When using unlagged data, the MR models have the MAE from 8.24 to 13.46 ppb and R^2 from 0.15 to 0.59. The ratio of MAE to average ozone concentrations ranged from 0.16 to 0.27. The NN models have the MAE from 7.01 to 12.41 ppb and R^2 from 0.27 to 0.70. The ratio of MAE to average ozone concentrations ranged from 0.15 to 0.24. Using lagged data may improve the model performance for both of the MR and NN models in a similar degree.

Cobourn et al. (2000) developed a hybrid nonlinear regression (NLR) model and a neural network (NN) model, each designed to forecast next-day maximum 1-hr average ground-level O_3 concentrations in Louisville, KY, to compare the performance of these two models for two O_3 seasons-1998 and 1999. In the NLR model, the multiple linear regression and nonlinear regression were combined to produce a model with an interactive nonlinear term plus additional linear terms. The NN model was a three-layer perception network with six input parameters. The model predictions were compared for the forecast mode (forecast meteorological parameters as input) and hindcast mode (observed meteorological parameters as input). For the hindcast mode, the NLR model had the MAE of 11.0 and 11.2 ppb for 1998 and 1999 ozone seasons, which were 15%

and 16% of the corresponding average observed ozone concentrations. The NN model has the MAE of 12.9 ppb for both of the two years, which was 18% and 17% for 1998 and 1999 average ozone. The model forecasts of the NLR and NN model were close to each other. Both of them have the MAE of around 13.0 and 11.8 ppb for 1998 and 1999, which were 18% and 15% of the corresponding average observed ozone concentrations. During the 1998 and 1999 ozone seasons, the forecast detection rate of 120 ppb threshold exceedances was 42% for each model. The hindcast detection rate was 92% for the NLR model and 75% for the NN model.

Jorquera et al. compared three forecasting models - time series model, NN model, and the fuzzy model, for daily maximum ozone levels at Santiago, Chile. The time series model used a simple linear model that considers only surface air temperature as the exogenous variable. The NN model is a classical three-layer, feed-forward model. In the fuzzy model, the fuzzy C-means algorithm (Rousseeuw et al, 1996) was applied for parameter identification. The data from different stations were used for model calibration and evaluation. The statistical indices root mean square (RMSE) and index of agreement (IA, 0 showing no agreement and 1 showing perfect agreement) were used to test the model performance. Comparing the hindcasting results on different data set, the NN model has the best statistical indices, then the fuzzy model and the time series model. For example, testing on data set E6, the NN model has the RMSE of 21.8 ppb and IA of 0.894, the fuzzy model has the RMSE of 22.9 ppb and IA of 0.885, whereas the time series model has the RMSE of 24.2 ppb and IA of 0.872.

There are several viable methods have been used for forecasting domain peak ozone concentrations. As shown above, the multiple regression models, neural network models, photochemical models, and fuzzy logic models can forecast ozone reasonably well. The best of these models can forecast high ozone threshold exceedences about 50% of the time, with seasonal MAEs as low as about 15% of the corresponding average observed ozone concentrations.

CHAPTER III

ENVIRONMENTAL DATA AND MODEL PERFORMANCE METRICS

The databases used for developing the ozone forecast models for the five metropolitan areas consist of air quality data and meteorological data during the ozone season (May to September), over the five year period 1998-2002, plus a group of deterministic parameters. The air quality data were the observed domain daily peak 8-hr ozone concentration used as the dependent variable in the regression. The meteorological data include daily instantaneous maximum and minimum temperature, average meteorological, and derived meteorological products. The deterministic parameters relate to factors that play important roles in ozone formation, include ozone trend, atmospheric transmittance, length of day. The values of these parameters could be generated automatically in the databases.

A. Air Quality Data

The air quality data were provided by Kentucky Division of Air Quality (KDAQ). Ozone monitors in the air quality control regions and in the counties in close vicinity were used for this study (Figure 1). In detail, three monitors were used for the Ashland area, Bowling Green area, and Paducah area, five monitors were used for the Owensboro area (Table 1).

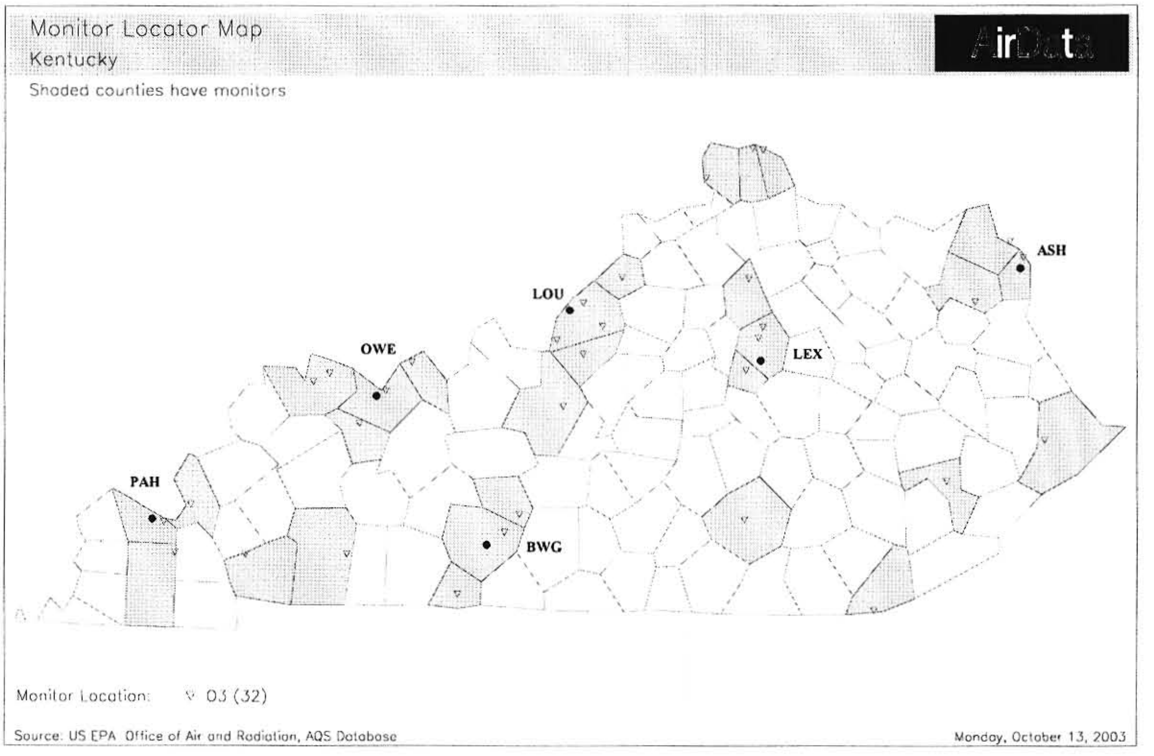

Figure 1. Ozone Monitor Locations in Kentucky (EPA, 2003f)

TABLE 1

Monitors in the Air Quality Control Regions.

Region	County	Monitor Location/ID
Ashland	Greenup	Worthington
	Boyd	Ashland
	Carter	Grayson Lake
Bowling Green	Simpson	Franklin DOT
	Edmonson	Mammoth Cave
	Warren	Oakland Elementary School
Owensboro	Hancock	Lewisport
	McClean	Guffie
	Henderson	Baskett Fire
	Henderson	Henderson
	Davies	Owensboro
Paducah	McCracken	JPRECC
	Livingston	Smithland
	Graves	Symsonia

Since a single forecast for the entire Air Quality Control region is issued by the local pollution control authorities, the appropriate ozone concentration for fitting the forecast models is the daily maximum value of all the monitors ("domain peak"). Hourly 8-hr average ozone concentrations for each monitor for the months of May-September, over the five year period 1998-2002, were obtained from KDAQ. Daily peak 8-hr average ozone concentrations were computed for each monitor. To obtain the most reliable domain peak ozone concentration, a requirement was that at least two of three monitors (three of five for Owensboro) had to be in operation. If on a particular day, the domain peak ozone concentration could not be determined for the minimum number of monitors, the day was excluded from the database.

B. Meteorological Data and Candidate Predictor Variables

In the regression process, observed meteorological data and certain deterministic parameters were used as the independent variables. These parameters can be divided into three classes: ①surface observed meteorological data, ② deterministic parameters, ③ derived meteorological products. From previous studies done for the Louisville and Lexington areas, and additional exploratory research done for this study, a list of 22 candidate parameters have been correlated with peak ozone concentrations. A new meteorological parameter thunderstorm (TS) was tested in this study. These candidate predictor variables were included in the ozone forecast model database for possible use in the multiple regression models (Table 2).

TABLE 2

Parameters Used in the Ozone Forecast Model

Parameter class	Parameter name	Symbol	Units	timing
Surface observed meteorological data	maximum temperature	tmax	°F	daily instantaneous
	minimum temperature	tmin	°F	daily instantaneous
	average temperature	tvg	°F	10 a.m. to 4 p.m. avg.
	dew point temperature	dewpt	°F	10 a.m. to 4 p.m. avg.
	cloud cover	cc		10 a.m. to 4 p.m. avg.
	Relative humidity	rh		10 a.m. to 4 p.m. avg.
	wind speed	mdwind	mph	10 a.m. to 4 p.m. avg.
	thunder storm	TS		5 a.m. to 5 p.m.
Deterministic parameters	length of day	lod	hours	daily
	atmospheric transmittance	xmitt		noon
	ozone trend	trend	Yr^{-1}	annual
	holiday	hol		--------
	Saturday	sat		--------
	Friday	fri		--------
Derived meteorological products	maximum temperature departure	tmx.dep	°F	daily
	minimum temperature departure	tmn.dep	°F	daily
	normal maximum temperature	Tmx.nrm	°F	daily
	normal minimum temperature	tmn.nrm	°F	daily
	special relative humidity 1	rhx1		10 a.m. to 4 p.m. avg.
	special relative humidity 2	rhx2		10 a.m. to 4 p.m. avg.
	special relative humidity 3	rhx3		10 a.m. to 4 p.m. avg.
	special relative humidity 4	rhx4		10 a.m. to 4 p.m. avg.

1. Surface observed meteorological data

The observed meteorological data used in this study were obtained from Local Climatological Data Reports (NCDC, 1998-2002). Location of the weather stations is shown in Table 3.

TABLE 3

Weather Stations in the Air Quality Control Regions.

Air Quality Control Region	Weather station name	Station location	Latitude	Longitude	Elevation (feet)
Ashland	Huntington, WV	Tri-state Airport	38° 22' N	82° 33' W	822
Bowling Green	Bowling Green, KY	Bowling Green	36° 59' N	86° 26' W	536
Owensboro	Evansville, IN	Dress Regional Airport	38° 02' N	87° 32' W	418
Paducah	Paducah, KY	Barkley Regional Airport	37° 03' N	88° 46' W	391

The effective data of the datasets consisted of daily maximum temperature, daily minimum temperature, hourly surface observations of sky description, precipitation, temperature, dew point, relative humidity, wind speed. To reduce the random fluctuations of the hourly observed data, some of the variables were averaged over several hours. Average temperature, dew point, wind speed, cloud cover were averaged from 10 A.M. to 4 P.M. Thunderstorm parameter was assigned a value of 1.0 when thunderstorm occurred in the time period 6 A.M. to 5 P.M., otherwise it was given a value of 0. The temperature extremes were instantaneous values from the datasets, not extremes of the hourly data.

Daily maximum temperature is the strongest predictor of ozone concentration. The standard explanations offered by Robeson and Steyn (Robeson and Steyn, 1990) are ① the rates of photochemical reactions are highly sensitive to temperature, and ② high air temperatures are associated with strong solar radiation, sunny skies, stagnant circulation, and subsiding upper air. The regressions of O_3 on maximum temperature for the four ozone control regions were shown in Figure 2. The second-order polynomial was found to be the best transformation function. For Ashland, Bowling Green, Owensboro, and Paducah, the values of coefficient of determination (R^2) were 0.426, 0.323, 0.363, and

0.325 respectively. On average, nearly forty percent variance in peak O_3 could be explained by maximum temperature.

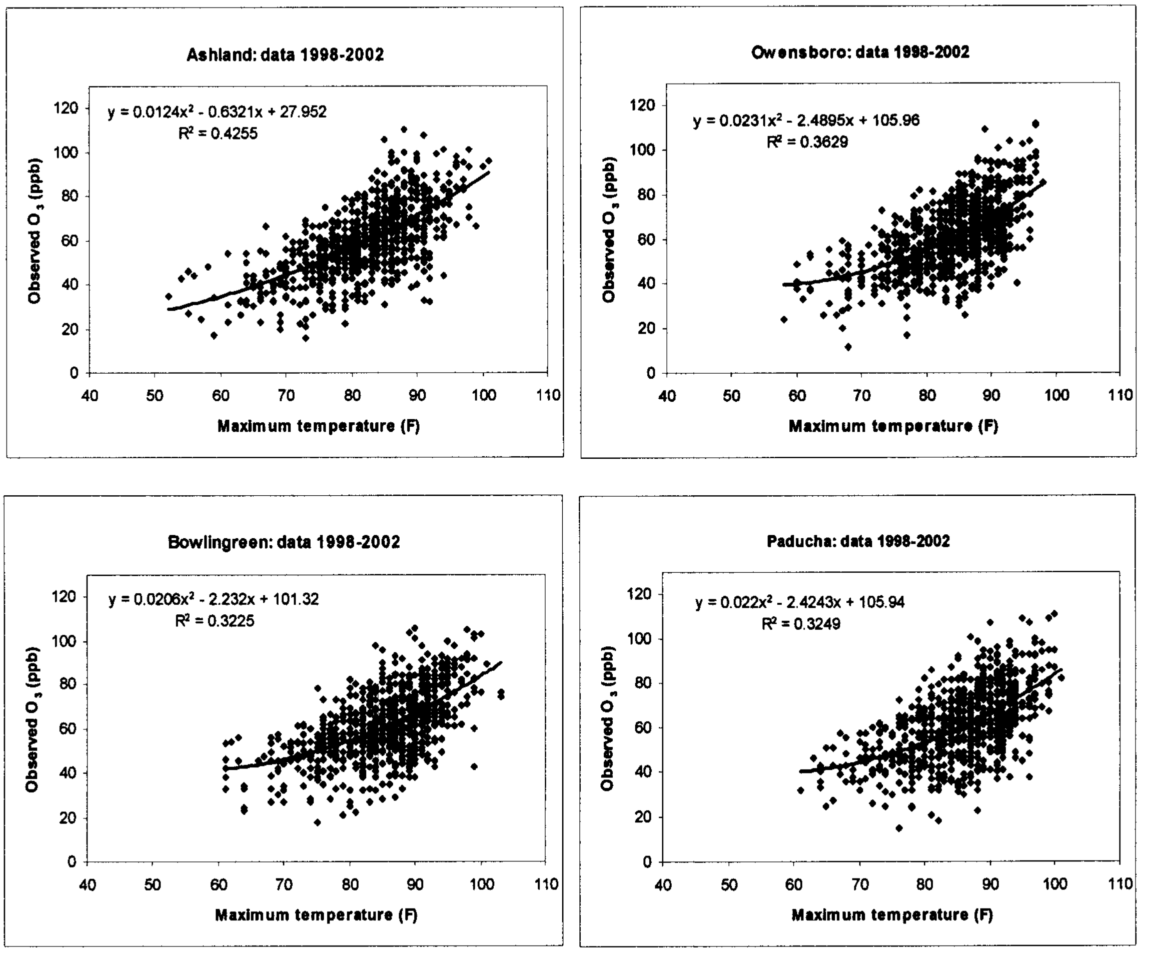

Figure 2. Second-order Polynomial Regression of Peak O_3 on Maximum Temperature.

Minimum temperature was used to calculate the derived meteorological parameter "rhx3" and "tmn.dep" in this study. In exploratory investigations, peak O_3 correlated positively with minimum temperature, but minimum temperature was not used as a direct ozone predictor because it is highly correlated with some other parameters, such as dew point and the maximum temperature. Average temperature was used to calculate the special relative humidity “rhx3”, one of three parameters contained in nonlinear term in the model.

Cloud cover is negatively correlated with the ozone concentration because clouds directly reduce solar radiation. The NCDC data set provides the sky condition descriptions such as clear, overcast, etc. These description terms were converted to equivalent tenths of cloud cover for the regression analysis (Table 4.) in the ozone forecast models.

TABLE 4

Sky Condition Descriptions Converted to Tenth of Cloud Cover

sky condition description	symbol	equivalent cloud cover value (tenth)
clear	CLR	0.5
few cloud	FEW	1.5
scattered	SCT	3
broken	BKN	7
overcast	OVC	9.5

Dew point provides a lower limit value on the minimum temperature, because of the latent heat of condensation of water. So the dew point displays similar relationship to ozone as minimum temperature. It is negatively correlated with ozone. Dew point was also used to calculate special relative humidity “rhx1” and “rhx2”.

Simple air pollutant concentration models give the theoretical relationship between pollutant concentration (C) and the wind speed (U). For example, the fixed-box model (De Nevers, 1995):

$$C = b + \frac{q \cdot l}{U \cdot h} \tag{2}$$

Or the Gaussian plume diffusion model (De Nevers, 1995):

$$C = \frac{q}{2\pi \cdot U \cdot \sigma_y \sigma_z} \exp(-\frac{y^2}{2\sigma_y^2}) \exp(\frac{-(z-h)^2}{2\sigma_z^2}) \tag{3}$$

These models show that pollutant concentration varies roughly inversely with wind speed. A variety of functional forms were fitted to the wind speed data. The best model was found to be a nonlinear exponential curve of the form

$$Y = \beta \exp(\theta \cdot X) \tag{4}$$

where X is the wind speed and Y is the concentration of pollutant. This form was used as part of the nonlinear term in the model. Investigation for the four ozone control regions gave the determination coefficients R^2 as 0.025, 0.039, 0.029, and 0.036 for Ashland, Bowling Green, Owensboro, and Paducah area respectively (Figure 3)

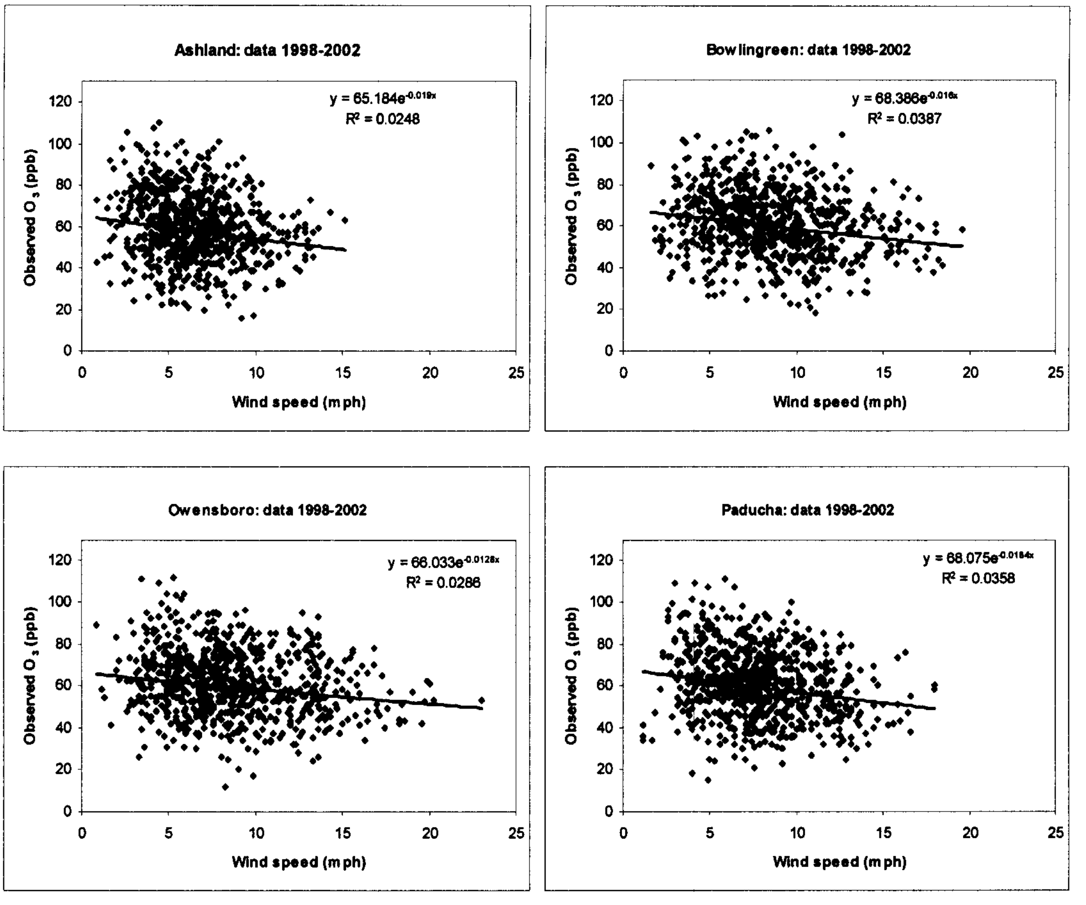

Figure 3. Nonlinear Regression of Peak O_3 on Wind Speed

Thunderstorm activity (TS) was selected as a new parameter based on two reasons. ① Thunderstorms are usually accompanied with heavy rain and unstable atmospheric conditions. The rain may reduce ozone through the process of "wet scavenging", and the vertical instability could vent pollutants. ② A forecasted thunderstorm probability can be obtained 24 hours in advance. To use the TS in the regression analysis, the value "1" was assigned when a thunderstorm occurred during the period 5am-5pm and "0" when it did not. In the Louisville ozone forecast model, TS has been found to be a statistically significant parameter in the model. But adding TS to the models developed in this study was not useful, since TS was not statistically significant in any of the multiple regressions.

2. Deterministic parameters.

This class of predictors consisted of parameters that are found useful in ozone forecasting. Length of day and atmospheric transmittance are two candidate regressors that account for the solar radiation, which drives the photochemical ozone formation process. These two parameters are calculated by day of year, zenith angle, and altitude angle of the ozone control area location. They are strongly correlated with each other. Therefore, the one that performed better in the multiple regression was selected as the independent parameter in the forecast model.

The trend parameter was included in the ozone forecast model based on the fact that in the previous decade, ozone concentrations have dropped gradually. From 1991 to 2000, the nationwide 1-hr ozone concentration reduced 10%; 8-hr ozone declined 7% (EPA, 2000). Emissions of NO_x and VOC compounds have also declined during this period. Ozone concentrations in the four ozone control regions are consistent with a negative trend (Figure 4). However, the ozone concentration trends were affected by meteorology, and 2000 and 2001 were relatively cool years for the 1998-2002 period. By including the trend term in the model, we can assess whether there was a downward trend independent of meteorology.

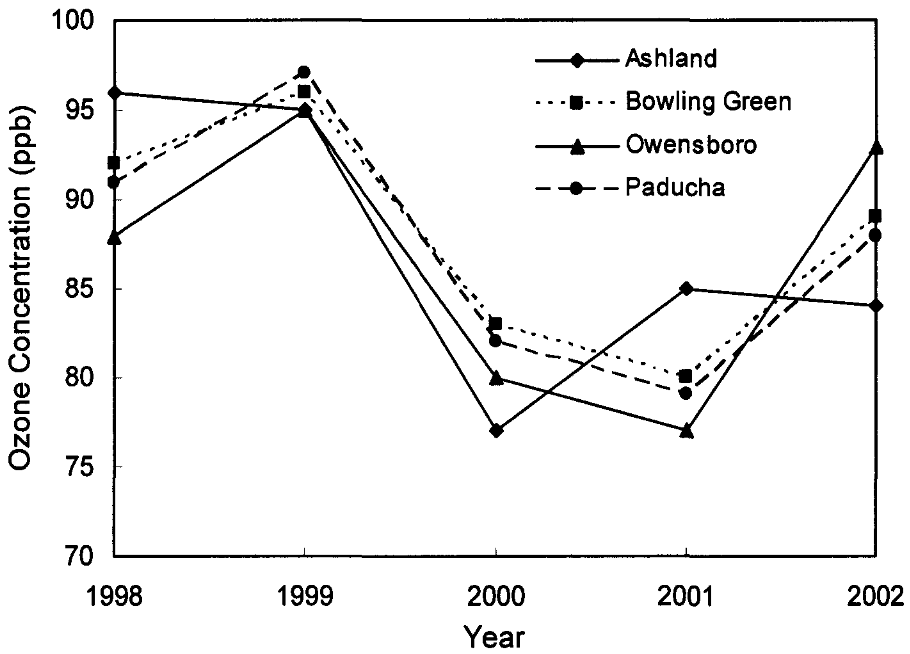

Figure 4. The 8th Maximum Ozone Concentration Pattern for the Four Cities, 1998-2002 (The 8th max ozone is the 8th highest value for the ozone data of certain city, which is a better statistic to represent the upper end of the distributions than the highest value.)

Saturday, Friday, and Holiday (4^{th} of July) were considered as parameters because on a holiday, the reduction of traffic and manufacturing could reduce the emissions of VOC and NOx, which are the precursors of ozone. Each of the three parameters has been statistically significant in some previous models.

3. Derived meteorological parameters.

Derived meteorological parameters consist of normal maximum and minimum temperature (tmx_nrm and tmn_nrm), maximum temperature departure (tmx_dep), minimum temperature departure (tmn_dep), and special relative humidity (rhx). The climatological normal daily temperatures are the 30-year average values computed from the data recorded during the period 1971-2000 (NCDC, 2003). The departure temperatures were obtained by calculating the differences between the actual maximum (or minimum) temperature and the normal maximum (or minimum) temperature. The departure temperatures were statistically significant in the Louisville and Lexington models (either one of the maximum and minimum departure temperature). It has been found that the seasonal patterns of 8-hr ozone and the normal temperatures are similar (Greenwell, 2000). This may explain why the departure temperatures were significantly correlated with ozone concentrations.

The relative humidity correlated negatively with peak ozone. Four special relative humidity terms (rhx1, rhx2, rhx3, rhx4) were used as the candidate regressors for the relative humidity in the model. They are computed by maximum daily, minimum daily temperature, average temperature, and average dew point temperature (Equation 5-8). The value of rhx2 is close to the average relative humidity for the 10am-4pm period.

$$rhx1 = \frac{Psat(dewpt)}{Psat(t\max)} \tag{5}$$

$$rhx2 = \frac{Psat(dewpt)}{Psat(tavg)} \tag{6}$$

$$rhx3 = \frac{Psat(t\min)}{Psat(t\max)} \tag{7}$$

$$rhx4 = \frac{Psat(t\min)}{Psat(tavg)} \tag{8}$$

where *Psat* is the saturation vapor pressure of water. Following standard practice, a polynomial function was used to calculate *Psat(T)*. Each special relative humidity is "surrogate" for relative humidity in the nonlinear and linear regressions. For the previous Louisville and Lexington models, the rhx3 were used to compute the nonlinear terms and the rhx2 is more statistically significant than the average relative humidity in the linear model.

C. Model Performance Metrics

In order to evaluate the performance of the forecasting models, several statistical indices were used, including correlation coefficient, statistical significance test value, mean error, mean absolute error, root mean squared error, detection rate, false alarm rate, and critical success index.

1. Correlation Coefficient (R^2).

Pearson's Correlation Coefficient R is the usual measure of correlation between the dependent variable and the independent variable in a linear regression, sometimes called product-moment correlation. The Square of Correlation Coefficient (R^2) is a measure of association which varies from 0 to 1, with 0.0 indicating no relationship and 1.0 indicating perfect relationship, defined as

$$R^2 = \frac{\sum_{i=1}^{N}(p_i - \bar{o}_i)^2}{\sum_{i=1}^{N}(o_i - \bar{o}_i)^2} \tag{9}$$

where p_i and o_i are the predicted and observed peak ozone, $\bar{o}_i$ is the average observed peak ozone. The R^2 is usually interpreted as the fraction of the variance of the dependent variable explained by the model.

2. Statistical significance test value (t-value)

Statistical significance brings into focus the possible uncertainty in the regression results due to sample size. The test statistic t-value reflects the statistical significance of each regression coefficient for multiple linear regressions. The t-value is formed by the ratio of a parameter coefficient divided by its respective estimated standard error, formed as

$$t = \frac{b_k}{s(b_k)} \tag{10}$$

where b_k is the estimate parameter coefficient, $s(b_k)$ is the standard error of b_k, defined as

$$s(b_k) = \frac{s_{res}}{\sqrt{(n-1)\cdot s_k^2 \cdot (1-R_k^2)}} \tag{11}$$

where n is the sample size, s_k^2 is the sample variance for the *k*th estimate parameter, R_k^2 is the squared multiple correlation between the *k*th estimate parameter and the remaining estimate parameters, s_{res} is the variance error of estimate.

The t-value is compared to the critical values of t at the designated level of significance (the probability of the t-value outside the critical value) with degrees of freedom. If the t-value of a regression coefficient is greater than the critical value, we can infer that the regression parameter is statistically significantly and there is correlation between the corresponding independent variables and the dependent variables. For the multiple linear regressions in this study, at the 0.05 level of significance with the degrees of freedom more than 700, the critical value of t is about 2.0 (Lomax, 2001).

3. Mean error (Bias) and Mean absolute error (MAE).

Bias is the arithmetic mean of the errors. The bias for the fitted data in a regression model should be zero. The bias for forecasted data using a regression model should be near zero. The Bias is given by

$$Bias = \frac{\sum_{i=1}^{n}(p_i - o_i)}{n} \tag{12}$$

MAE is the average absolute value of the forecast errors. It is used to evaluate the average absolute deviation of the predicted values from the observed values. The MAE is given by

$$MAE = \frac{\sum_{i=1}^{n} | p_i - o_i |}{n} \tag{13}$$

4. Root mean squared error (RMSE)

RMSE is the square root of the mean of the squares of all the forecast errors, given by

$$RMSE = \sqrt{\frac{\sum (p_i - o_i)^2}{n}} \tag{14}$$

RMSE is also called standard deviation used to evaluate the deviation of the predicted values from the observed values. Compared to MAE, RMSE is more sensitive to outliers. RMSE is also more widely used than MAE and can be employed in further statistical analysis (Wilmott, 1981).

5. Detection Rate (DR).

The DR is the fraction of the observed exceedences detected by the model. It is calculated by

$$DR = \frac{DE}{EX} \tag{15}$$

where DE is the number of detected exceedences, and EX is the number of total observed exceedences. The model "detects" an exceedence based on the model prediction exceeding pre-determined alarm level. The alarm level may be set at the air quality exceedence level, or slightly below, to provide a margin of safety. The DR generally decreases with increasing alarm threshold (Hubbard, 1997). The recommended alarm threshold for the Louisville and Lexington ozone forecast models is 75 ppb (Cobourn, 2001), which is 10 ppb below the nominal NAAQS exceedence threshold (85 ppb).

6. False alarm rate (FAR)

The FAR is the fraction of alarms that were false alarms. It is defined as the ratio of false alarms (FA) to total alarms (AL) predicted by the model.

$$FAR = \frac{FA}{AL} \tag{16}$$

A false alarm occurred when the observed value is below and predicted value above the alarm threshold. Increasing the alarm threshold will reduce both the AL and FA, but the FAR tends to increase. Lowering the alarm threshold would tend to improve the DR, but increase the number of alarms and false alarms. Based on the justification that too many alarms would lead to a loss of public confidence, public officials in Louisville and Lexington apply the 85 ppb as the alarm threshold.

7. Critical success index (CSI)

The CSI is the ratio of valid alarms (alarms minus false alarms: $AL - FA$) to critical events. Critical events include alarms and undetected exceedences (exceedences minus detected exceedences ($EX - DE$). The CSI can be calculated by

$$CSI = \frac{AL - FA}{AL + EX - DE} \tag{17}$$

CHAPTER IV

PRELIMINARY STUDY

A. The Standard and Hybrid Models.

Two basic regression equations were used in each of the ozone forecast models. These were named the standard model and the Hi-lo model. The standard model was fitted to all days in the database, with equal weighting, so as to predict ozone levels with equal probability of success on all days. The Hi-lo model was developed to improve the detection rate on days conducive to high ozone. This was done by fitting the Hi-lo model to the days on which the ozone concentrations were in the upper or lower 10% of the ozone distribution. In this way, the middle part of the ozone distribution was removed, in order to increase the influence of the high ozone days on the model fit. Compared to standard model, the Hi-lo model had greater success in detecting on the high ozone days.

The ozone forecast models are thus hybrid models, which combine the standard model and Hi-lo model equations to improve forecast performance. An independent set of criteria was used to switch between standard model and Hi-lo model, as follows:

- Maximum temperature greater than 87 °F;
- Wind speed less than 6.0 mph;
- Cloud cover less than 2.5 tenths.

The criteria are called the 3S criteria in recognition of three important weather characteristics associated with high ozone level: sunny, sultry, and stagnant. This switching strategy increased the detection rate and increased the explained variance, without significantly changing the bias or MAE error for the model (Cobourn, 1999).

B. Trajectory Models and Non-trajectory Models

Trajectory-based models include a derived meteorological parameter: the trajectory parameter. Ozone and its precursors, particularly NOx, can be transported over distances of several hundred kilometers or more. Air mass trajectory analysis could be used to identify the direction and location of sources of ozone or NOx. By parameterizing the trajectory information, a trajectory parameter was formed and included into the regression model.

The NOAA Transport and Dispersion website provided a Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model for calculating forward or backward trajectories at various levels for continental US locations (NOAA, 2003). Both the past trajectories and trajectory forecasts can be calculated by HYSPLIT. The archives of the trajectories were available for studying the relationship between the trajectories and the peak ozone concentration, also forecasts of the trajectory can be used for predicting ozone concentrations. The trajectories were compared to a map (Figure 5.) that displays an envelope encompassing most of the recorded high ozone trajectories and large NOx emission sources. The trajectory parameter was assigned as a value of 1.0 if the backward trajectory was fully inside the envelope, a value of zero if outside the envelope, a value of 0.5 if on the margin.

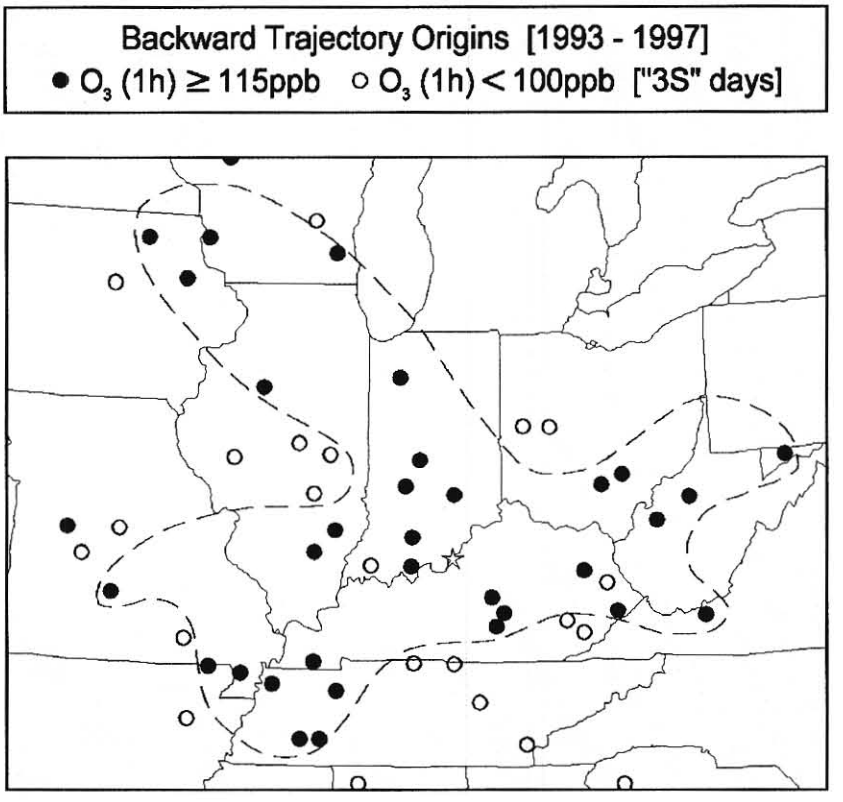

Figure 5. Origins of the 36-hr Backward Trajectories at 750m Elevation on High-ozone Days during the Period 1993-1997 (Cobourn, 1999)

Applying the trajectory parameter in Louisville model and Lexington model resulted in an improvement in the model accuracy. For example, the model fit for the Louisville model over the period of 1993-1997, when the trajectory parameters were included, the MAE improved from 11.1 to 9.4 ppb for the Hi-Lo regression; the MAE improved from 8.7 to 8.3 ppb for the standard model (Cobourn, 1999). Currently both the Louisville model and the Lexington model are trajectory-based models.

Non-trajectory models are the models that do not include the trajectory parameter. The non-trajectory models were necessary when the models were designed to run automatedly, such as the ozone forecast models applied on the internet ozone calculator. For these automated models, a requirement is that all the input parameters could be computed automatically. Since the values of the trajectory parameter need to be determined manually by professional, the trajectory parameter could not appear in the automated ozone forecast models. Results from the non-trajectory models were reasonably closed to those of the trajectory models. The ozone forecast models for the four ozone control regions in this study are non-trajectory models.

C. Louisville and Lexington Model Comparison

The ozone forecast models for the four ozone control regions in this study were developed based on the Louisville and Lexington models. The Louisville and Lexington ozone forecast 2001 models were fitted to the data from 1997 to 2001. For the Louisville model, nine and five explanatory parameters were used in the standard and Hi-lo model respectively. For the Lexington model, ten and eight explanatory parameters were used in the standard and Hi-lo model respectively. Trajectories were used in both of the two models.

The predicted daily peak ozone concentrations were called model forecasts if they were computed using forecast meteorological data, and called model hindcasts if they were the retrospective predictions using observed meteorological data. Using Louisville and Lexington 2001 models to predict the ozone concentrations during the 2002 ozone season, the model forecasts tracked the day-today ozone variation reasonably well (Figure 6 and Figure 7). The ozone forecasts were next day forecasts, with 30 hours forecast meteorological input data.

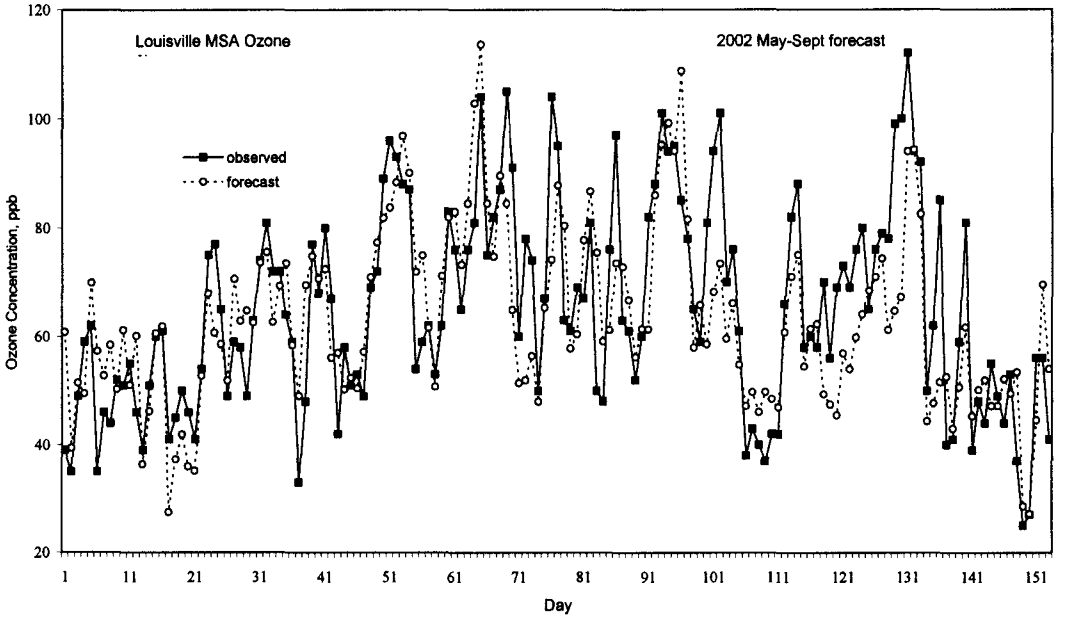

Figure 6. Time Series of Observed and Predicted Daily Maximum Ozone Concentration for the Forecasts during 2002 Ozone Season (Louisville)

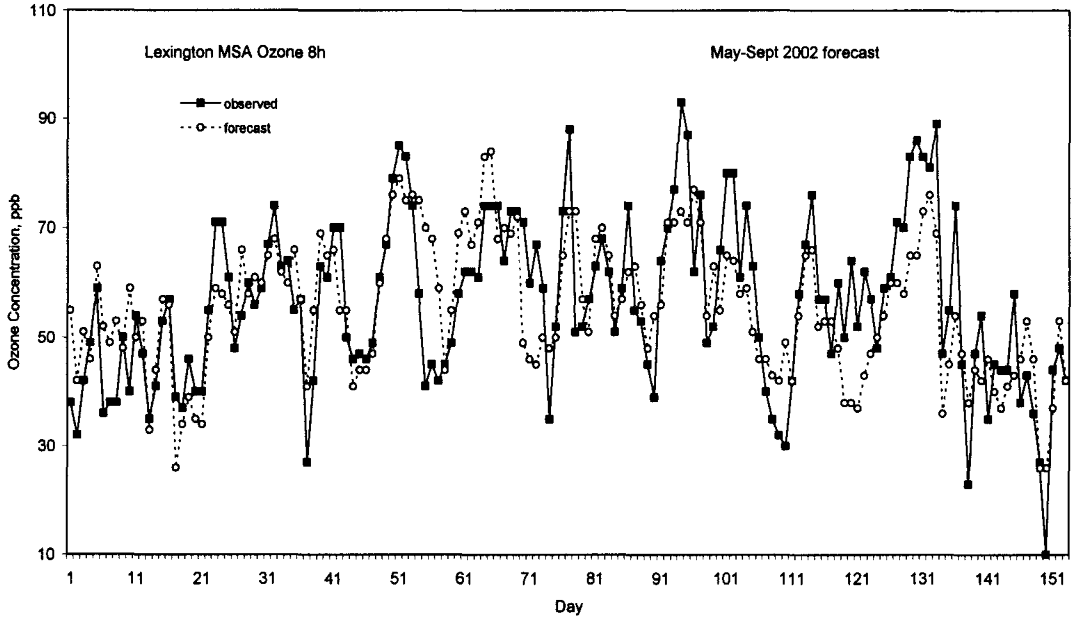

Figure 7. Time Series of Observed and Predicted Daily Maximum Ozone Concentration for the Forecasts during 2002 Ozone Season (Lexington)

The overall statistical comparison of the 2002 daily maximum ozone forecasts, hindcasts by 2001 ozone forecast model, and the model estimates by 2002 model is presented in Table 5. The model estimates were the retrospective predictions on the model calibration period, using observed meteorological data. The statistical results showed that the model estimates were more accurate than those of the hindcasts. Also the hindcasts were more accurate than those of the forecasts. The degradation from model estimates to hindcasts, then to forecasts, is caused by two reasons. First, compared to the model estimates, in the hindcast mode the model is being tested against new data. The effect of the unexplained factors contributing to ozone may vary from year to year. Second, in the forecast mode, the model input is meteorological forecast data. There are always some errors in these forecasts, i.e. the observed meteorological data are different from the forecasted data. So the model forecasts had higher errors than the model hindcasts.

TABLE 5

Statistics for 2002 Daily Maximum Ozone Model Estimates, Hindcasts, and Forecasts (Louisville and Lexington, threshold=75ppb)

Index	Model estimate (by 2002 model)	Hindcasts (by 2001 model)	Forecasts (by 2001 model)
		Louisville	
MAE (ppb)	6.7	8.8	9.8
Bias (ppb)	0.4	1.1	-1.8
DR	1.00	0.92	0.69
FAR	0.02	0.12	0.17
AL	42	42	30
		Lexington	
MAE (ppb)	5.8	7.5	8.5
Bias (ppb)	0.1	1.1	-0.8
DR	0.83	0.83	0.17
FAR	0.06	0.26	0.44
AL	17	23	9

D. Special Studies

1. Averaging period

The average temperature (tavg), dew point (dewpt), and midday wind (mdwind) are important ozone predictors. In the previous Louisville and Lexington models, the averaging period of these parameters was 10am to 4pm. In this special study, other averaging periods for tavg, dewpt, and mdwind were applied to examine whether other averaging periods can improve the regression. The averaging period tested in this study is 10am-6pm, 12pm-4pm, 4pm-8pm, and 8am-8pm.

The NCDC data were selected to do this study, because the archived NCDC data files contain complete meteorological data and the files can be imported into a spreadsheet. The Louisville 2001 ozone forecast model was used for the study. We first re-computed average values of the parameters in the new averaging periods; then re-fitted the models using parameters with various averaging times. The comparison of correlation coefficients (R^2) and residual standard error (RSE) based on different averaging periods are shown in Table 6.

TABLE 6

Statistics for the Louisville 2001 Models Based on Various Averaging Periods: 10am-4pm, 10am-6pm, 12pm-4pm, 4pm-8pm, and 8am-8pm.

Index	10am-4pm	10am-6pm	12am-4pm	4pm-8pm	8am-8pm
R^2	0.68	0.69	0.68	0.69	0.69
RSE (ppb)	9.732	9.617	9.696	9.524	9.539

Using the different averaging period, the statistical results were closed to each other. The largest difference for RSE was 0.2 ppb (9.732 vs. 9.539), which was not really significant. The largest variation of the R^2 was less than 1.8%. The averaging period in later afternoon (4pm-8pm), or a longer average period including later afternoon (8am-8pm) has little better values of R^2 and RSE, but the time period 4pm-8pm goes against what we know about period of ozone formation, which should be during the midday (Hubbard, 1997). The average period 10am-4pm still used in the models for Louisville, Lexington and models for other cities in this study.

2. Rainfall

Rainfall is associated with reducing ozone levels for several reasons. First, the rainfall may directly remove part of the O_3 and O_3-precursors from the air, called wet scavenging. Second, rainfall is associated with increased cloud cover and increased convective activity. Increased cloud cover lead to reduced level of the ultraviolet, increased convective activity may increase mixing and dilution of atmospheric pollutant. All these factors would reduce ozone levels. In this special study, we use 1996-2001 ozone and rainfall data in Louisville to investigate how the rainfall affects ozone levels.

The hourly precipitation data were extracted from the NCDC data files. The rainfalls in different time period in a day were calculated separately. The morning rain, mid-day rain 1, mid-day rain 2, and evening rain were defined as the total rainfall in the time period of 1 am-8 am, 9 am-8 pm, 6 am-5 pm, and 9 pm-12 pm respectively. The statistics of the linear regressions of peak ozone on rainfall in different time period were shown in Table 7. Since peak ozone usually occurs at afternoon, the mid-day rain, especially the 6am-5pm rainfall accumulation most strongly correlated with the peak ozone as expected, with a R^2 of 0.0709 and t-value of -7.12. A comparatively weak correlation between the morning rain and the peak ozone was obtained, with a R^2 of 0.0302 and t-value of -4.39.

If the morning rain ended several hours prior to the diurnal peak ozone, not much ozone in air at that time, anyway. Also by the time of maximum O_3 production and concentration, new air will have been convected into city. In this air, may not have rained in the morning. The results also showed that the evening rain have very small effect on the diurnal ozone peak.

TABLE 7

Statistics of the Linear Regressions of Peak Ozone on Rainfall.

Index	morning rain (1am-8am)	mid-day rain 1 (9am-8pm)	mid-day rain 2 (6am-5pm)	evening rain (9pm-12pm)
R^2	0.0302	0.0476	0.0709	0.0016
Coefficient	-23.39	-17.67	-23.52	-6.66
t-value	-4.39	-5.77	-7.12	-1.00

For the Louisville 2001 model, the parameter rainfall was not included in the final model. Because the parameter rainfall correlates with some other independent parameters, such as cloud cover, midday wind speed, and humidity, the rainfall parameter was not statistically significant in the multiple linear regression. However, the mid-day rain (6am-5pm) has been proved to be strongly correlated with peak ozone. If the hourly rainfall is available for the ozone forecast model, the rainfall should be a candidate parameter tested in the regression.

3. Thunderstorm (TS) and thunderstorm probability (TSP)

Thunderstorm is characterized by the heavy rain and strong wind. Thunderstorm would reduce ozone levels with the same mechanism that rainfall affects the ozone. The observed TS were used in the multiple linear regression to set up the ozone forecast models. Based on the investigation on rainfall, the TS occurred in time period 6am to

5pm was assigned value "1", otherwise it was assigned value "0". One and two days ahead predictions of thunderstorm probabilities can be obtained from the meteorological predictions made by nested grid model (NGM), issued by NOAA. The TSP predictions give out the probabilities value of the thunderstorm occurring in the following days between 0 and 1. Actually thunderstorm didn't occur in most of the days with predicted TSP less than 0.7. But it is still reasonable to use TSP to predict the ozone level. Because the days with high TSP usually accompany overcast and turbulent air flow, these factors would reduce ozone level, even though no rainfalls in part of those days.

The 1998-2002 ozone concentration and TS data for Louisville, Ashland, Bowling Green, Owensboro, and Paducah were used to investigate the effects of the TS on ozone levels. Correlating TS with ozone concentrations by a linear regression, the t-values of the TS for all the regressions were greater than 2.0 in absolute value (Table 8). That indicated the TS significantly correlated with ozone concentrations. The coefficients of the TS parameters have the negative values between -7.52 and -12.06. That verified that the TS could help to reduce the ozone concentration.

TABLE 8

Statistics of the Linear Regressions of Peak Ozone on TS (6am-5pm) for Five Cities (Data: 1998-2002)

Index	Louisville	Ashland	Bowling Green	Owensboro	Paducah
R^2	0.0225	0.0156	0.0316	0.0348	0.0510
Coefficient	-12.06	-7.52	-8.87	-11.27	-11.66
t-value	-3.51	-3.45	-4.96	-5.22	-6.39

The thunderstorm parameter was successfully applied on Louisville 2002 model. In this model, the TS parameter was statistically significant in the multiple linear regressions for the standard model, with the t-value of -3.27, the coefficient of -3.56. By adding TS parameter, the R2 for the standard linear regression was improved from 0.7649 to 0.7686. The TS was also used as the candidate parameter for constructing the ozone forecast model for Ashland, Bowling Green, Owensboro, and Paducah. However, the TS parameter was not used in the final models for those cities since it was not statistically significant in the multiple linear regressions for those models. The reason is same as that of the rainfall: the TS parameter correlates with some other independent parameters, such as cloud cover, midday wind speed, and humidity.

4. Error correction concept

The error correction conception derived from the serial correlation of the daily ozone data. Based on the phenomenon that daily maximum ozone concentrations are partially dependent on the previous day's concentrations, some investigators have used previous day ozone as a predictor variable in their models (Comrie, 1997). Also a 24-hr parameter, which intended to represent the previous day ozone concentration along a 24-hr backward air trajectory, has been developed and tried to improve the Louisville ozone forecast model (Greenwell, 2000). However, since the current day observed ozone are usually not available in time when making the next-day ozone forecast, both the previous day ozone and 24-hr parameter were not used as input parameter in the Louisville and Lexington ozone forecast models.

In this special study, the error was defined as the difference between the model forecast and observed maximum ozone concentration. The observed and model forecast ozone data came from Louisville 2001 database, covered 774 days from 1995 to 2000. The errors of the previous one day (e_1), two day (e_2), and three day (e_3) were correlated with the current day error (error) using linear, second order and third order nonlinear regression equations. For the linear regression, the multiple R^2 was 0.234; the t-value of e_1, e_2, and e_3 were greater than 2.0, indicating all of them were statistically significant in the regression. The e_1 most strongly correlated with the current day error with a coefficient of 0.326 and t-value of 9.949. The e_3 had the weakest correlation with current day error. Several linear regressions that include second order and third order of the errors were also investigated. The best one was the following equation

$$error = c_1 \cdot e_1 + c_2 \cdot e_2 + c_3 \cdot e_3 + c_4 \cdot e_1^2 + c_5 \cdot e_2^2 + c_6 \cdot e_3^2 \tag{18}$$

By adding the second order items, the R^2 was improved from 0.234 to 0.281. Except the $(e_2)^2$, the other parameters were statistically significant in the regression. Still the e_1 most strongly correlated with the current day error with the largest coefficient and t-value (Table 9).

TABLE 9

Statistics of the Multiple Regressions of Current Day Errors on Previous Day Errors

(Data: Louisville 1995-2000)

	Index	e_1	e_2	e_3	$(e_1)^2$	$(e_2)^2$	$(e_3)^2$
	Coef.	0.326	0.137	0.088			
Regression1	t-value	9.949	3.986	2.701			
	Multiple R^2=0.234						
	Coef.	0.359	0.096	0.076	0.006	0.001	0.002
Regression2	t-value	10.673	2.673	2.295	6.271	0.154	2.069
	Multiple R^2=0.281						

This investigation showed that the previous one day, two day, and three day errors did strongly correlate with current day error. The previous day forecast errors can be used to correct the next day forecast error, so as to improve the model forecast accuracy. When making prediction, the availability problem for the previous day data still exist. Since the previous one day error strongest affects the current day error, the error correction didn't applied in our ozone forecast models. If the data can be obtained in the future, the error correction could be used in the updated models.

5. Common model

Stepwise regression procedure was used in constructing the Louisville and Lexington model. The stepwise regression procedure is an automated selection technique that used to filter out the variables that provide non-significant contributions and eventually obtain the optimal combination of the variables for the regression. The stepwise regression procedure was conducted when constructing the earlier Louisville and Lexington models (Hubbard, 1997; Greenwell, 2000). Different parameter groups were selected by the stepwise regression procedure for the earlier Louisville and Lexington models. For example, parameters used for Louisville 2000 model were nonlin, LOD, traj, trend, rhx2, tmx_dep, sat, holiday, and cc; parameters used for Lexington 2000 model were nonlin, xmitt, traj, trend, rhx2, tmx_dep, tmn_dep, sat, fri, and mdwind. Some of the parameters, include nonlin, traj, trend, and rhx2, were used in both of the two models. These parameters strongly correlated with the peak ozone with high t-value in the multiple linear regressions. Further investigation showed that these parameters could explain most of variation of the peak ozone. For instance, based on Louisville 2000 database, the linear regression for standard model using only nonlin, traj, trend, and rhx2 had a R^2 of 0.6321, which was close to the R^2 value of the final model, 0.6787.

The analysis above indicated that a group of common parameters could exist, which are strong ozone predictors and could be commonly used in the model for different cities. A model including only these common parameters is called common model. The common model could be used as the ozone forecast model for the cities with similar size and location. It can make roughly peak ozone predictions with an acceptable accuracy. Also for a particular city, based on the common model, other parameter significantly correlated with peak ozone could be added to improve the model accuracy. These parameters call additional parameters. It's reasonable to divide the model parameters into two groups: common parameters and additional parameters. The common parameters include the strongest factors that affect the ozone concentrations. The additional parameters are still significant, but improve the model predictive power only slightly. They depend on the characteristics of the ozone control region. The common model concept was applied to develop the ozone forecast models for Ashland, Bowling Green, Owensboro, and Paducah. The preliminary study showed that when the common parameters for Louisville and Lexington model were also apply to the models in this study, these parameters are still significantly correlated with peak ozone.

CHAPTER V

MODEL DEVELOPMENT

The ozone forecast models for the four ozone control regions– Ashland, Owensboro, Bowling Green, and Paducah are non-trajectory, hybrid nonlinear regression models. The development of these models was based on the Louisville and Lexington models. Since the models are designed to run automatedly, the trajectory parameter was excluded from the models. The hybrid model consists of two separate regressions, known as the standard model and the Hi-Lo model, and the 3S criteria used to switch between standard model and Hi-lo model. The multiple linear regressions for both the standard and Hi-lo models contain a nonlinear term and several other parameters. The nonlinear term was constructed by three key parameters: temperature, wind speed, and relative humidity. The model building processes for the four cities are similar. As an example, the process for Ashland was described in detail. The final model parameters and coefficients for the other ozone control regions are given in the following sections.

A. Ashland Ozone Forecast Model

1. Data preparation

The Ashland air quality data used in this study consist of the daily maximum 8-hr average ozone concentrations from three ozone monitor sites, named by county location: Greenup, Boyd, and Carter. Some of the data from these sites were missing. To insure consistence in calculating the daily domain peak ozone concentration from these data, only days for which data from at least two of three sites (three of five for Owensboro) was available were retained in the database. This criterion eliminated 11 days from the Ashland database, leaving 754 days representing the five ozone seasons.

For the Hi-lo model, an additional calibration dataset was obtained by removing days in the middle of the ozone distribution. To standardize the models, uniform cutoff thresholds were applied to all the models: 82.1 ppb and 42.1 ppb, corresponding roughly to the 90^{th} and 10^{th} percentiles. The Ashland Hi-lo model database contained 218 days from the five ozone seasons.

2. Description of the model construction process

The two-step model building approach used in Louisville and Lexington model was also used in building the ozone forecast models for the four ozone control regions. In the first step, a nonlinear term was developed that accounts for the nonlinear behavior of ozone with regard to maximum temperature, wind speed, and relative humidity. In two-way regressions, a second-order polynomial in maximum temperature correlated well with peak O_3, an exponential function in wind speed correlated well with peak O_3, and an exponential function in relative humidity correlated well with peak O_3. Through explanatory analysis (Cobourn and Hubbard, 1999), it has been determined that the following interactive function well describes the combined effects of temperature, wind speed and relative humidity on ozone:

$$nonlin = (p_1 + (p_2 \cdot t\max + p_3 \cdot t\max^2) \cdot \exp(p_4 \cdot mdwind)) \cdot \exp(p_5 \cdot rh) \qquad (19)$$

Since the special relative humidity term rhx3 was more statistically significant than relative humidity in the regression, it was used in place of relative humidity in equation (19). So the functional form of the nonlinear term used in this study was

$$nonlin = (p_1 + (p_2 \cdot t\max + p_3 \cdot t\max^2) \cdot \exp(p_4 \cdot mdwind)) \cdot \exp(p_5 \cdot rhx3) \quad (20)$$

The coefficients p_1 to p_5 were determined by the nonlinear regression. The nonlinear term can be regarded as a rough nonlinear ozone forecast model that could forecast ozone concentration based on the three parameters: tmax, ws, and rhx3. The nonlinear term was then used as an independent variable in the multiple linear regression determined in the second model constructing step.

In the second step, a multiple linear regression was fitted using the ordinary least squares (OLS) method. The general form of a multiple linear regression model is expressed by

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \quad (21)$$

where Y is the dependent variable, X_j represents the p independent variables, β_0 is the intercept, β_j are the regression coefficients and ε is the errors. The independent variables consisted of the nonlinear term plus part of the parameters listed in Table 2 in Chapter III. Whether a parameter correlate to ozone concentration and can be used as an independent variable mainly depends on the t-statistic value of this parameter in the multiple linear regression. As described in the previous chapter, the threshold of t-value was 2.0 in this study. That means if the t-value of a parameter is greater than 2.0, this parameter can be considered as an independent variable.

To determine which variables would be chosen in the models, an initial linear model formulation was required as the start point. The Louisville model (2002) was used as the initial model respectively to develop the standard model. Then the standard model parameter set was used as the initial model to develop the Hi-lo model. A combination of the stepwise regression method provided by the statistical software package and trial-and-error substitution was used to fit the models in this study. The stepwise method does not guarantee that an optimum model will be formed. In addition, physical reasoning and previous model building experience has led to certain guideline, as follows:

• The xmitt and LOD are correlated each other. Both of them account for the effect of the solar radiation to the ozone formation. They were not used in the model together.

• The various special relative humidity terms was cross-correlated, so they were test one by one.

• The various temperature parameters – tmax, tmin, dewp, tvg, tmx.dep, and tmn.dep are correlated with each other. They were tested in the linear regression separately but not used in the model together.

The two-step model building approach was applied in developing both the standard and Hi-lo models. Finally the hybrid model was developed by combining the standard and Hi-lo model, using the 3S criteria for model selection..

3. Final model parameters and coefficients

The final form of the standard model consisted of an intercept, six regression coefficients, and six explanatory variables (See Table 2, Page 19):

$$O_3 = b_0 + b_1 \cdot nonlin + b_2 \cdot xmitt + b_3 \cdot trend + b_4 \cdot rhx2 + b_6 \cdot cc + b_7 \cdot mdwind \quad (22)$$

The fitted coefficients, standard errors, and t-statistic values for the standard model are shown in Table 10.

TABLE 10

Regression Coefficients for the Standard Model (Ashland)

Variable	Coef. symbol	Coef. Value	Std. error	t-statistic
Constant	b_0	-172.09	20.6	-8.35
nonlin	b_1	0.793	0.039	20.2
xmitt	b_2	307.46	32.16	9.56
trend	b_3	-1.226	0.237	-5.17
rhx2	b_4	-0.115	0.03	-3.81
cc	b_6	-0.684	0.197	-3.47
mdwind	b_7	-0.333	0.146	-2.28

The characteristics of the Ashland standard model were as follows:

- The t-values of all the variables exceed 2.28 in absolute value. That indicates all the explanatory variables contributed significantly to the linear regression.
- The nonlinear term was the strongest contributor with a t-value of 20.2.
- The transmittance term had higher t-value than the length of day term, so the transmittance term was kept in the model instead of length of day.
- Among the four specific relative humidity terms, the rhx2 term was kept in the model because it has better performance in the linear regression.
- The cloud cover term was included in the mode, even though the effect of clouds may be partially accounted for by the rhx2 term.
- The wind speed term still contributed to the linear regression even though it was used in the nonlinear term.
- The thunder storm term was evaluated in the multiple linear regression. It was not kept in the Ashland model because of its t-value was much smaller than 2.

The nonlinear term accounted for 63.0% of the variation in ozone concentration. The first four terms (nonlin, xmitt, trend and rhx2) accounted for 68.8%, and the complete model accounted for 69.5% of the variation. The nonlinear term was calculated by the nonlinear regression equation (Equation. 20). In the nonlinear regression, the initial values of the coefficients were determined based on the Louisville model. The fitted coefficients for the nonlinear parameters are shown in Table 11 along with the standard errors and t-statistic.

Table 11

Regression Coefficients for the Nonlinear Regression Used in the Standard Model (Ashland)

Coefficient	Fitted value	Std. error	t-statistic
p1	80.41	5.79	13.89
p2	-2.45	0.495	-4.95
p3	0.0347	0.00536	6.47
p4	-0.0871	0.0190	-4.59
p5	-0.00924	0.00053	-17.50

The final form of the Hi-lo model consisted of an intercept, five regression coefficients, and five explanatory variables:

$$O_3 = b_0 + b_1 \cdot nonlin + b_2 \cdot xmitt + b_3 \cdot rhx2 + b_4 \cdot trend + b_5 \cdot cc \qquad (23)$$

The fitted coefficients, standard errors, and t-statistic values for the standard model are shown in Table 12.

TABLE 12

Regression Coefficients for the Hi-lo Model (Ashland)

Variable	Coef. symbol	Coef. Value	Std. error	t-statistic
Constant	b_0	-166.74	38.11	-4.56
nonlin	b_1	0.761	0.066	11.56
xmitt	b_2	316.25	62.73	5.04
rhx2	b_3	-0.129	0.071	-1.8
trend	b_4	-1.898	0.523	-3.63
cc	b_5	-0.756	0.452	-1.67

Comparing to the standard model, the Hi-lo model of Ashland has the following characteristics:

- The parameter set of the Hi-lo model consisted of nonlinear, transmittance, trend, and cloud cover term. These parameters were also used in the standard model. The wind speed term was excluded from the Hi-lo model because of its low t-value in the regression.
- The nonlinear term still was the largest contributor to the regression.
- The t-values of the rhx2 and cloud cover term were lower than but close to 2. These two terms were kept in the model because relative humidity and cloud cover are two important factors that affect the ozone concentrations.

The coefficients for the nonlinear term used in the Hi-lo model along with the standard errors and t-statistics are given in Table 13.

Table 13

Regression Coefficients for the Nonlinear Regression Used in the Hi-lo Model (Ashland)

Coefficient	Fitted value	Std. error	t-statistic
p1	69.96	31.20	2.42
p2	-2.49	1.567	-1.59
p3	0.0428	0.01460	2.93
p4	-0.0629	0.0314	-2.00
p5	-0.01511	0.00153	-9.85

B. Bowling Green Ozone Forecast Model

The data preparation and the model development process for Bowling Green, Owensboro, and Paducah model were the same as the Ashland model. The standard model for Bowling Green consisted of an intercept, seven regression coefficients, and seven explanatory variables:

$$O_3 = b_0 + b_1 \cdot nonlin + b_2 \cdot xmitt + b_3 \cdot trend + b_4 \cdot rhx2 + b_5 \cdot tmn\_dep + b_6 \cdot cc + b_7 \cdot mdwind \quad (24)$$

The fitted coefficients for the standard model were different from, but somewhat close to those of the Ashland model (Appendix A). For the standard model, except for the cc variable, the t-values of all the other variables exceeded 3.83 in absolute value, indicating that all the explanatory variables contributed significantly to the linear regression. The nonlinear term was the strongest contributor with a t-value of 17.25. The minimum temperature departure term was significant in the regression for Bowling Green model with a positive coefficient.

The final form of the Hi-lo model utilized four explanatory variables, as follows:

$$O_3 = b_0 + b_1 \cdot nonlin + b_2 \cdot xmitt + b_3 \cdot trend + b_4 \cdot rhx2 \quad (25)$$

The fitted coefficients for the Hi-lo model were different from, but in the general vicinity of those of the Ashland model (Appendix A). The wind speed, minimum temperature departure, and cloud cover terms were excluded from the model because of their low t-value. The trend term was kept in the model even though its t-value was slightly lower than 2.0 in absolute value. The nonlinear term was still the most significant parameter in the regression.

The coefficients for the nonlinear regression used in the Hi-lo model were the same as for the standard model. When a separate Hi-lo nonlinear regression was done, some of the terms were not significant when fitting the nonlinear equation (Equation. 20) to the Hi-lo model data set. This meant that there was some uncertainty in the values of the regression coefficients. Also, the values of these coefficients were much different from those of nonlinear regressions for the other cities. Therefore, for Bowling Green, the nonlinear term of the standard model was used for the Hi-lo model nonlinear term.

C. Owensboro Ozone Forecast Model

The standard model of Owensboro consisted of an intercept, eight regression coefficients, and eight explanatory variables:

$$\begin{aligned} O_3 = b_0 + b_1 \cdot nonlin + b_2 \cdot xmitt + b_3 \cdot trend + b_4 \cdot rhx2 + b_5 \cdot tmn\_dep + b_7 \cdot mdwind \\ + b_8 \cdot dewpt + b_9 \cdot hol \end{aligned} \tag{26}$$

The values of the fitted coefficients for the standard model were unique, but in the vicinity of those of the other models (Appendix A). The t-values of all variables were greater than 2.0 in absolute value, except dew point and holiday term, which were just under 2.0. The nonlinear term was the strongest contributor with a t-value of 13.43. For the nonlinear term, the t-values of all coefficients were greater than 2.0 in absolute value.

The final form of the Hi-lo model utilized six explanatory variables: nonlin, xmitt, trend, tmn_dep, mdwind, and dewpt.

$$\begin{aligned} O_3 = b_0 + b_1 \cdot nonlin + b_2 \cdot xmitt + b_3 \cdot trend + b_4 \cdot rhx2 + b_5 \cdot tmn\_dep + b_7 \cdot mdwind \\ + b_8 \cdot dewpt \end{aligned} \tag{27}$$

The fitted coefficients for the Hi-lo model were of comparable magnitudes to those of Ashland and Bowling Green model (Appendix A). The t-values of all variables were greater than 3.78 in absolute value except the wind speed term. The nonlinear term was the most significant parameter with a high t-value of 26.85. The wind speed term was kept in the model even though its t-value was slightly lower than 2.0 in absolute value. For the nonlinear term, the t-values of all coefficients were greater than 2.0 in absolute value.

D. Paducah Ozone Forecast Model

The standard model of Paducah consisted of an intercept, seven regression coefficients, and seven explanatory variables:

$$O_3 = b_0 + b_1 \cdot nonlin + b_2 \cdot xmitt + b_3 \cdot trend + b_4 \cdot rhx2 + b_5 \cdot tmn\_dep + b_6 \cdot cc + b_7 \cdot mdwind \quad (28)$$

The fitted coefficients for the standard model were comparable in magnitude to those of the other city models (Appendix A). The t-value of all the variables exceeded 2.46 in absolute value, indicating that all the explanatory variables contributing significantly to the linear regression. The nonlinear term was the strongest contributor with a t-value of 16.74. All city models included the terms nonlin, xmitt, trend, and rhx2. The Paducah model was typical of all models in that these four terms accounted for about 65.0% of the variation in ozone concentration, compared to 66.2% for the complete model. For the nonlinear term, the t-values of all coefficients are greater than 2.0 in absolute value.

The final form of the Hi-lo model utilized six explanatory variables:

$$O_3 = b_0 + b_1 \cdot nonlin + b_2 \cdot xmitt + b_3 \cdot trend + b_4 \cdot rhx2 + b_7 \cdot mdwind + b_{10} \cdot tmn\_nrm \quad (29)$$

The fitted coefficients for the Hi-lo model are shown in Appendix A. The nonlinear term was the most significant parameter with a t-value of 10.36. The minimum normal temperature term was used in the model instead of minimum temperature departure term. The coefficients for the nonlinear regression were not significant when fitting the nonlinear equation (Equation. 20) to the Hi-lo model data set. So the coefficients for the nonlinear term that were used in the standard model were also used in the standard model.

CHAPTER VI

MODEL VALIDATION

The final hybrid models for the four ozone control regions were used to make predictions of 8-hr daily maximum ground-level ozone concentrations based on the 1998-2002 calibration data set (model estimates). Also to evaluate the model performance on independent data set, the models were recalibrated to 1998-2001 data sets and were used to predict ozone concentrations on 2002 ozone season with observed meteorological data (model hindcasts). The observed ozone concentrations were compared with the model estimates and hindcasts for the days with available ozone data. The statistics for the model estimates and hindcasts were compared for each ozone control regions. Also the statistics used to evaluate the models for the four ozone control regions were compared with each other. In addition, to evaluate model forecasts on a new data set, Lexington 2002 model, which calibrated to 1998-2002 data set, was used to predict O_3 on 2003 ozone season with meteorological forecasts data.

A. Ashland Ozone Forecast Model

1. Performance on calibration data set (1998-2002).

Performance of the final Ashland ozone forecast model on calibration data set was evaluated by comparing the model estimates with the observed ozone concentration in the calibration periods. For the standard model, the prediction errors were approximately normally distributed about an average error approach to zero (-0.002 ppb), with a standard deviation of 9.01 ppb (Table 14). The correlation coefficient (R^2) was 0.695. For the hybrid model, by combining the standard model and Hi-lo model using the 3S criteria, the R^2 was improved to 0.856. The average absolute error was 7.29 ppb. Approximate 88% of the absolute errors (674 of 765 days) were less than 15.0 ppb, 74% of the absolute errors (566 of 765 days) were less than 10 ppb.

TABLE 14

Model Performance Statistics of the Models on 1998-2002 Calibration Periods (Ashland)

Model	MAE (ppb)	RMSE (ppb)	Bias (ppb)	R^2
Standard	7.1	9.01	-0.002	0.695
Hi-lo	9.0	11.03	1.990	0.832
Hybrid	7.3	9.22	0.94	0.856

For the database period 1998-2002, the errors (MAE and RSME) of the hybrid model were slightly higher than the errors of the standard model. This may be because low ozone concentrations occurred on some of the 3S days and the standard model had already over predicted ozone for those days. The hybrid model predictions were usually larger, so in those cases the error was greater. For example, on the 3S day July 3rd 1999, the observed ozone concentration is only 44 ppb. The prediction of the standard model was 74.8 ppb, whereas the prediction of the hybrid model was even higher, 77.3 ppb. However, on the high ozone days, the hybrid model was more accurate. For the 10% highest ozone days (77 days) in the 1998-2002 calibration periods, the hybrid model had a bias of -6.6 ppb and MAE of 10.2 ppb, whereas the standard model had a bias of -10.8 ppb and MAE of 11.8 ppb.

Time series plots of observed and predicted daily maximum ozone concentrations using the hybrid model demonstrate the ability of the final hybrid model to track day-to-day ozone variation. As an example, Figure 8 shows the time series plot for the final hybrid model for September 2001. The predictions are seen to agree quite closely with the observed concentrations on most days. On a few days there were comparatively large errors. On some high ozone days, the ozone concentrations were under predicted by the model, such as the 13th September. Time series for the other months shows the similar situations.

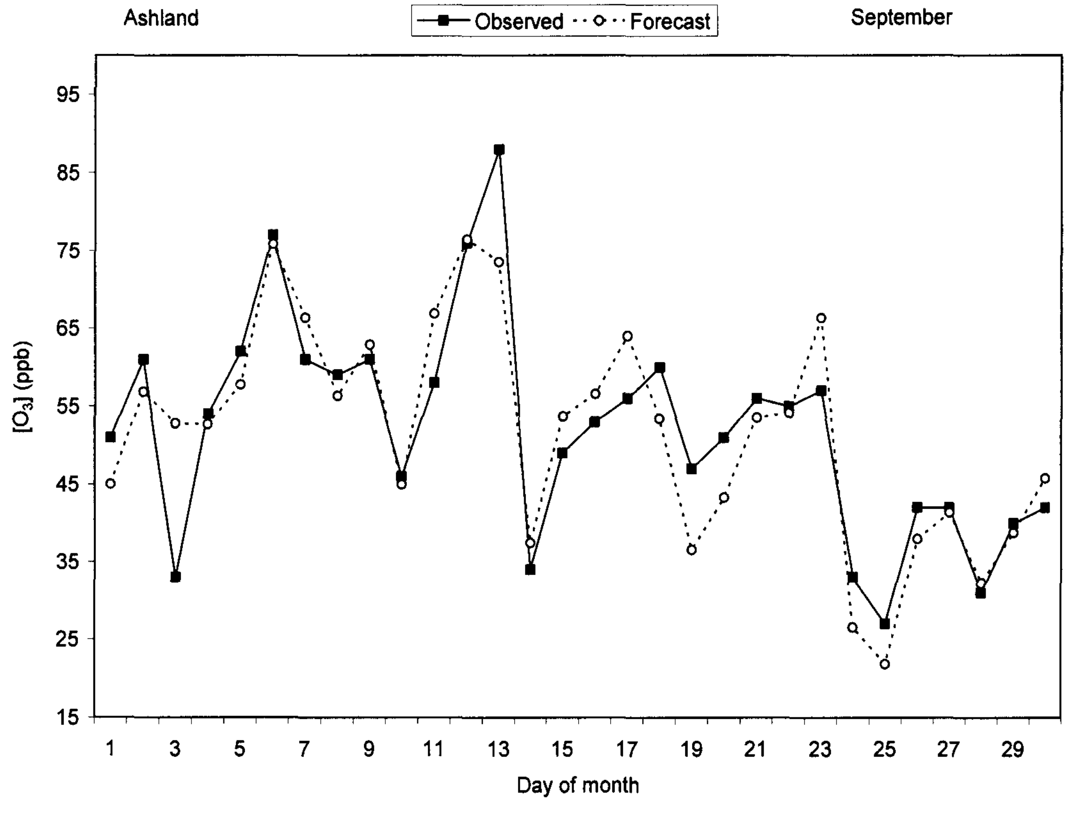

Figure 8. Time Series Comparison of Observed and Model Estimated Ozone Concentration for Ashland, KY, during September 2001 (Final model).

The indexes detection rate (DR), false alarm rate (FAR), and critical success index (CSI) indicate the effectiveness of a model in predicting high ozone concentrations. The values of these parameters were affected by the alarm threshold. The NAAQS unhealthy limit is 85 ppb. Alarm levels of 75 ppb and 80 ppb have been proposed for forecasting purposes. The DR, CSI, and FAR indicators based on the lower alarm thresholds were significantly better than those based on the 85 ppb threshold (Table 15). However, a lower threshold always results in more alarms and more false alarms. Based on the threshold of 85 ppb, the number of alarms and false alarms was 53 and 29, whereas based on the threshold of 75 ppb, the number of alarms and false alarms increased to 143 and 45. Too large number of alarms and false alarms might carry unnecessary limitations for individual and social activities so as to jeopardize popular support for the ozone action program. So for this Ashland ozone forecast model, the threshold of 80 ppb may be a good choice.

Table 15

Exceedance Detection Parameters for the Final Model Using the Alarm Threshold of 75, 80 and 85 ppb (Ashland)

parameter	symbol	thre.=75	thre.=80	thre.=85
detection rate	DR	0.87	0.62	0.45
false alarm rate	FAR	0.31	0.44	0.55
critical success index	CSI	0.65	0.46	0.29
events	EV	150	111	82
exceedences	EX	53	53	53
detected exceedences	DE	46	33	24
alarms	AL	143	91	53
false alarms	FA	45	40	29

The scatter plot of the predicted ozone concentrations versus the observed ozone concentrations for the calibration data set is shown in Figure 9. Approximately equal numbers of points lying on both sides of the diagonal line indicate the good correspondence between hindcasts and observations.

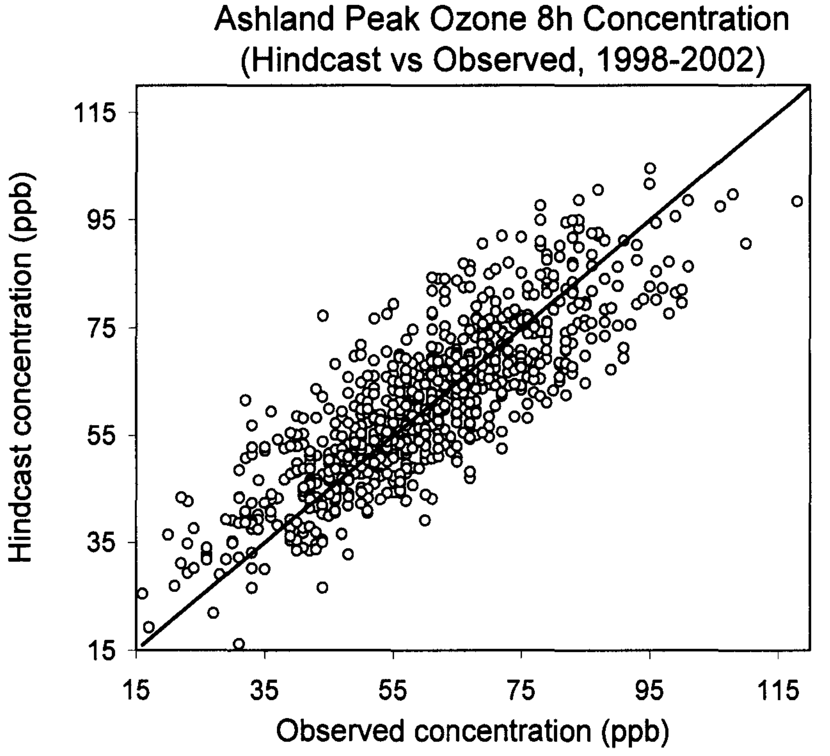

Figure 9. Scatter Plot of Hybrid Prediction against Observations for the Calibration Data Set (Ashland). (The diagonal indicates the line of perfect correspondence between hindcasts and observations.)

The scatter plot of residuals of the model estimates versus observed ozone concentrations shows that errors were mostly unbiased over the range of O_3 cone (Figure 10.). The residual is defined as the difference between the observed and predicted values,

$$\text{Re}\,sidual = [O_3]_{obs.} - [O_3]_{pred.}$$

Linear regression models typically exhibit negative bias at high concentrations; the nonlinear hybrid models usually do not.

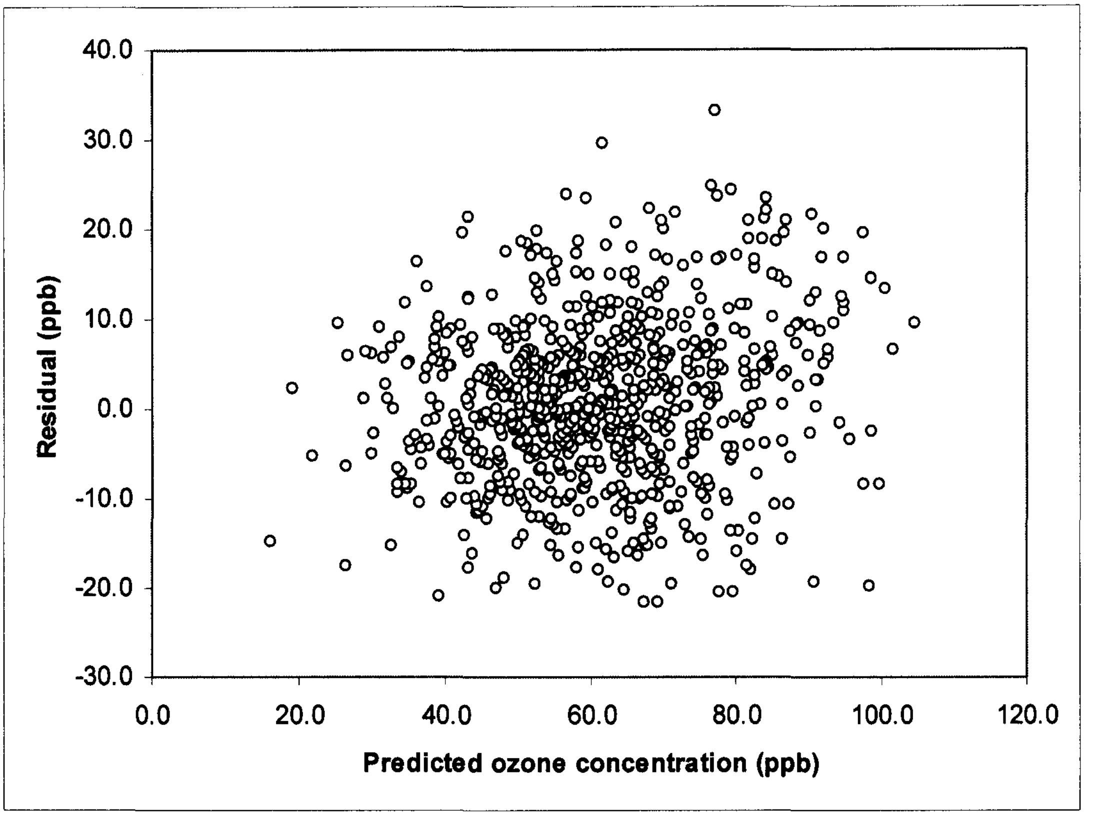

Figure 10. Residuals of the Hybrid Model versus the Predicted Ozone Concentrations (Ashland)

2. Validation with Independent Data Set

To test the final models on an independent data set, we recalibrated the models to the 1998-2001 dataset using exactly the same parameters group as the final model. The new model was then used to predict the peak ozone concentrations of 2002. The ultimate test will be with the 2003 and 2004 environmental data. Experience has shown that 5 years of calibration data is recommended for good, reliable model performance, so in that sense, this represented harsh testing conditions. Since these predictions were based on the observed meteorological data, they are actually hindcasts, here they are referred to as the "model hindcasts" to distinguish with the "model estimates" that were based on the 1998-

2002 calibration period. For the models recalibrated to 1998-2001 period, the regression coefficients for the standard and hybrid models are listed in Appendix B.

The model hindcasts tracked the day-to-day ozone variation reasonably well (Figure 11.). The time series of the errors of model hindcasts shows that the errors uniformly distribute on both side of the zero line (Figure 12.). The model over-predicted ozone concentrations for some of the high ozone days, such as September 6th, 7th, and 8th. It under predicted some of the low ozone days, such as May 4th, May 21st, and July 11th. However, the original model fitted to 1998-2002 showed the tendency of over-predictions on high ozone days and under-predictions on low ozone days. This anomaly was possibly caused by the uncharacteristic differences in meteorology and ozone climatology between 2002 and the 1998-2001 periods.

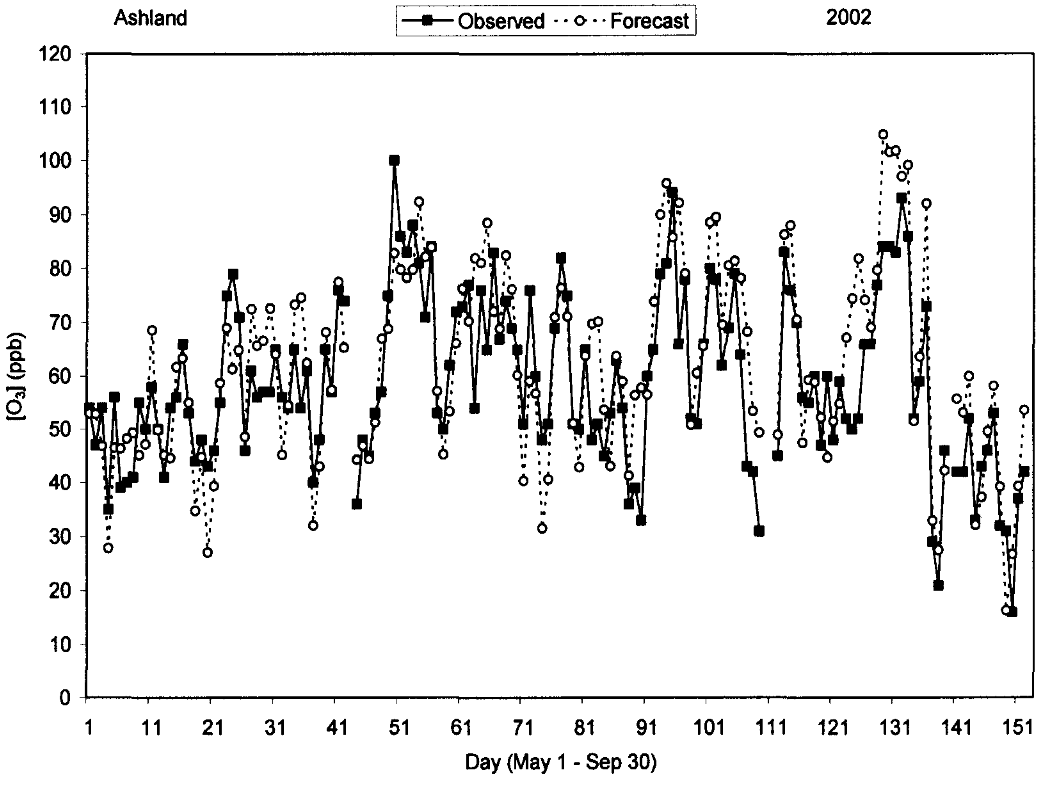

Figure 11. Time Series of Observed and the Re-calibration Model Hindcasts during the 2002 Ozone Season (Ashland).

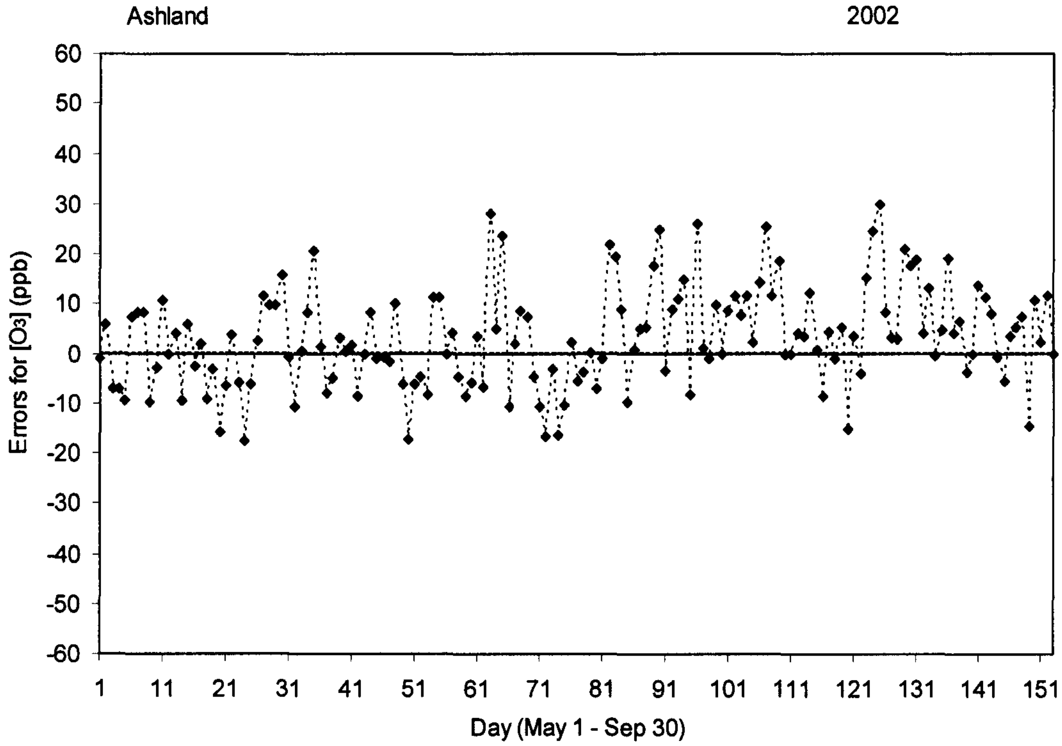

Figure 12. Time Series of Errors for the Re-calibration Model Hindcasts during the 2002 Ozone Season (Ashland).

The overall statistical comparison of the 2002 re-calibration model hindcasts and original model estimates is presented in Table 16, based upon the alarm threshold value of 80 ppb. The MAE of the hindcasts was 8.1 ppb, which is 0.9 ppb higher than the value of the model estimates. The hindcast Bias and RSME were greater than that of the model estimates by 1.9 and 1.4 ppb respectively. This degradation in error is typical for applying the model to new (unfitted) data (Cobourn, 2003). The FAR and CSI of the model hindcasts were slightly worse than those of the model estimates, as expected.

Table 16

Model Performance Statistics, 2002 Predictions (threshold = 80 ppb, Ashland)

Index	model hindcasts (model fit:1998-2001)	model estimates (model fit:1998-2002)
Bias (ppb)	3.0	1.1
MAE (ppb)	8.1	7.2
MAE/ $\overline{O_3}$ (%)	13.8%	12.3%
RSME (ppb)	10.6	9.2
DR	0.67	0.50
FAR	0.52	0.38
CSI	0.44	0.53

B. Bowling Green Ozone Forecast Model

1. Performance on calibration data set (1998-2002)

Based on the calibration data set, the overall correlation coefficient (R^2) for the final standard model was 0.68. The average error was close to zero (0.003 ppb), with a standard deviation of 8.77 ppb (Appendix C). Compared to the standard model, the hybrid model had a better R^2 of 0.836. For the database period 1998-2002, the errors (MAE and RSME) of the hybrid model were slightly higher than the errors of the standard model, for the reasons explained for the Ashland model. For the hybrid model, the average absolute error was 7.0 ppb. Approximate 87% of the absolute errors (668 of 765 days) were less than 15.0 ppb, 71% of the absolute errors (544 of 765 days) were less than 10 ppb. Among the models for the four ozone control regions, Bowling Green ozone forecast models have the lowest MAE and RSME for both the standard and hybrid model.

An example of time series plots for September 2001 was given in Figure 13. The predictions are seen to agree quite closely with the observed concentrations on most days. Time series for the other months shows the similar situations.

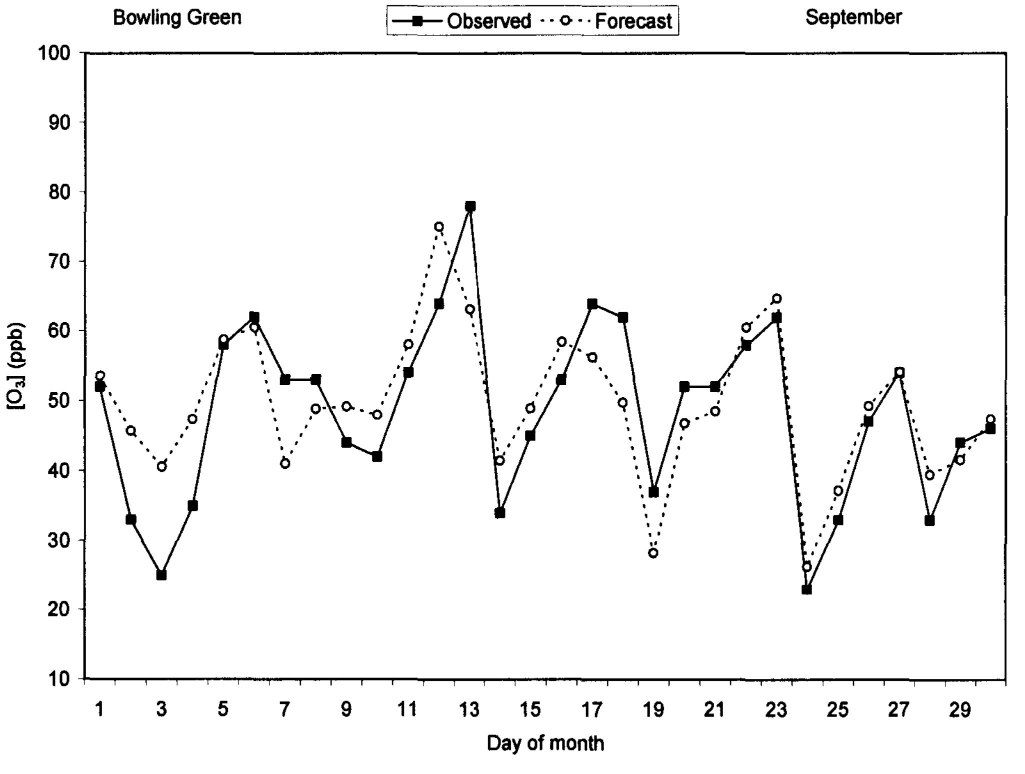

Figure 13. Time Series Comparison of Observed and Model Estimated Ozone Concentration in Bowling Green, KY, during September 2001 (Final model).

For the Bowling Green hybrid model, by using alarm threshold of 80 ppb, the detection rate (0.48), the critical success index (0.45), and the false alarm rate (0.38) were slightly lower than the DR, CIS, and FAR for the other cities. The reason is that this test was based on a four year instead of five year calibration period. These statistics (DR, CSI, FAR) depend on a small subset of the complete dataset; so it is to be expected that these would be more month to month variation. With small datasets, a few aberrant events can change statistics more easily than with large datasets.

The scatter plot of the predicted ozone concentrations versus the observed ozone concentrations for the calibration data set is shown in Figure 14. Approximately equal numbers of points lying on both sides of the diagonal line indicate the good correspondence between hindcasts and observations.

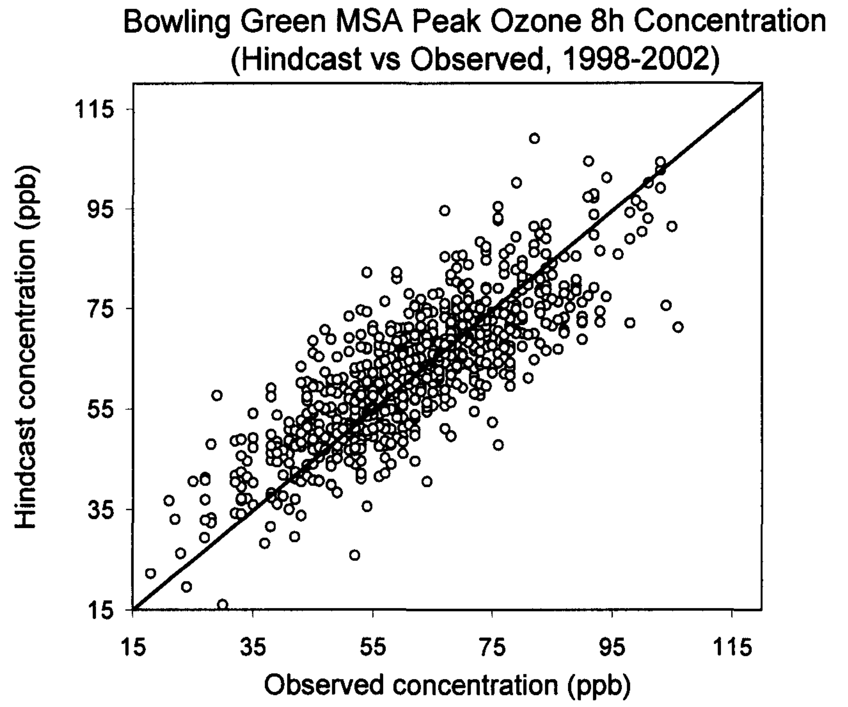

Figure 14. Scatter Plot of Hybrid Prediction against Observations for the Calibration Data set (Bowling Green). (The diagonal indicates the line of perfect correspondence between hindcasts and observations.)

2. Validation with Independent Data Set.

To test the model on an independent data set, the model was recalibrated to 1998-2001 data sets and used to predict the peak ozone concentrations of 2002. The regression coefficients for recalibrated models were listed in Appendix B. The time series plot during 2002 ozone season (Figure 15.) showed serious under-predictions for some days.

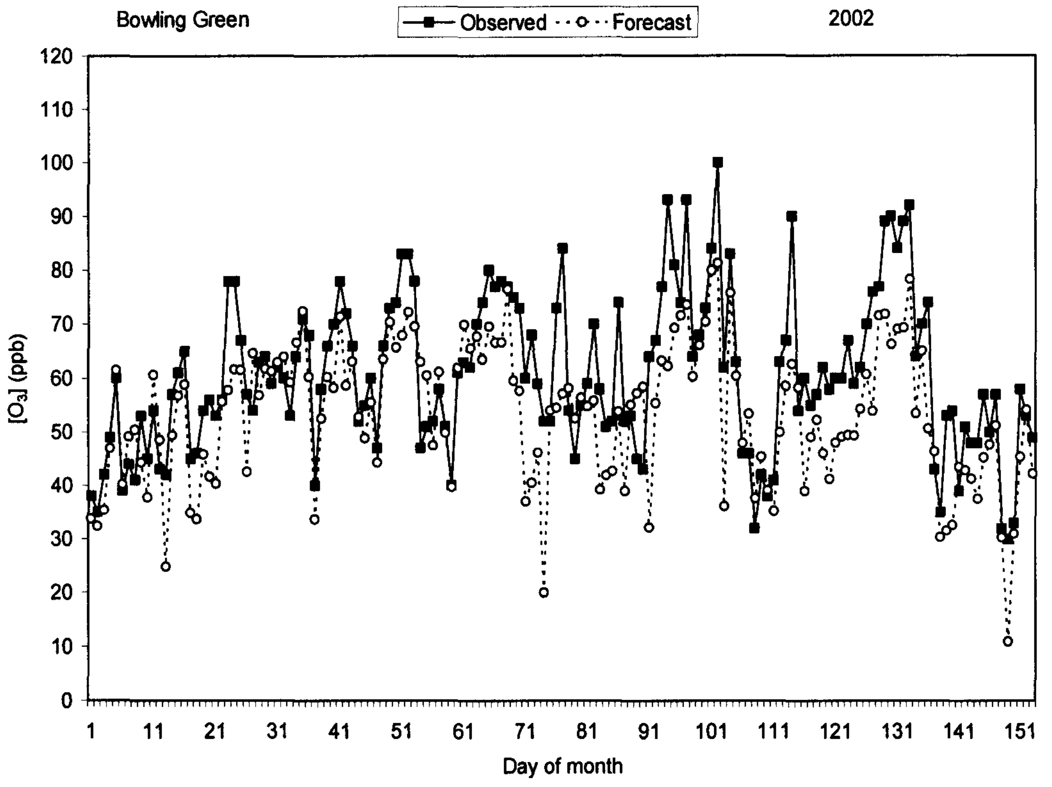

Figure 15. Time Series of Observed and the Re-calibration Model Hindcasts during the 2002 Ozone Season (Bowling Green).

The overall statistical comparison of the re-calibration model hindcasts and original model estimates for the 2002 ozone season is presented in Table 28, based upon the alarm threshold value of 80 ppb. The performance statistics of the model estimates had low errors and comparative high detection rate, critical success index, and false alarm rate (Table 17), whereas the model hindcasts had a large negative Bias of -7.2 ppb and low detection rate and critical success index. These results indicate the model hindcasts seriously under predicted the ozone concentration on some days. The reason for the under-predictions may be the abnormal temperature in summer 2002 in Bowling Green, which was much lower than the average temperature of the model calibration periods

1998-2001. This result reinforced the notion that a 5 year calibration period is necessary for reliable model performance.

Table 17

Model Performance Statistics, 2002 Predictions (threshold = 80 ppb, Bowling Green)

Index	model hindcasts (model fit:1998-2001)	model estimates (model fit:1998-2002)
Bias (ppb)	-7.2	-1.8
MAE (ppb)	9.5	6.6
MAE/ $\overline{O_3}$ (%)	15.7%	10.9%
RSME (ppb)	12.0	8.6
DR	0.13	0.63
FAR	0.00	0.25
CSI	0.13	0.55

C. Owensboro Ozone Forecast Model

1. Performance on calibration data set (1998-2002)

Based on the calibration data set, the R^2 for the final standard model was 0.68. The MAE, RSME, and Bias were 6.84, 8.72, and -0.003 ppb respectively. Compared to the standard model, the hybrid model had a better R^2 of 0.78, and slightly higher errors (Appendix C). For the final hybrid model, the average absolute error was 6.84 ppb. Approximate 90% of the absolute errors (687 of 765 days) were less than 15.0 ppb, 76% of the absolute errors (579 of 765 days) were less than 10 ppb.

An example of time series plots for May 2001 was given in Figure 16. The predictions are to agree with the observed concentrations reasonably well on most days. Time series for the other months shows the similar situations.

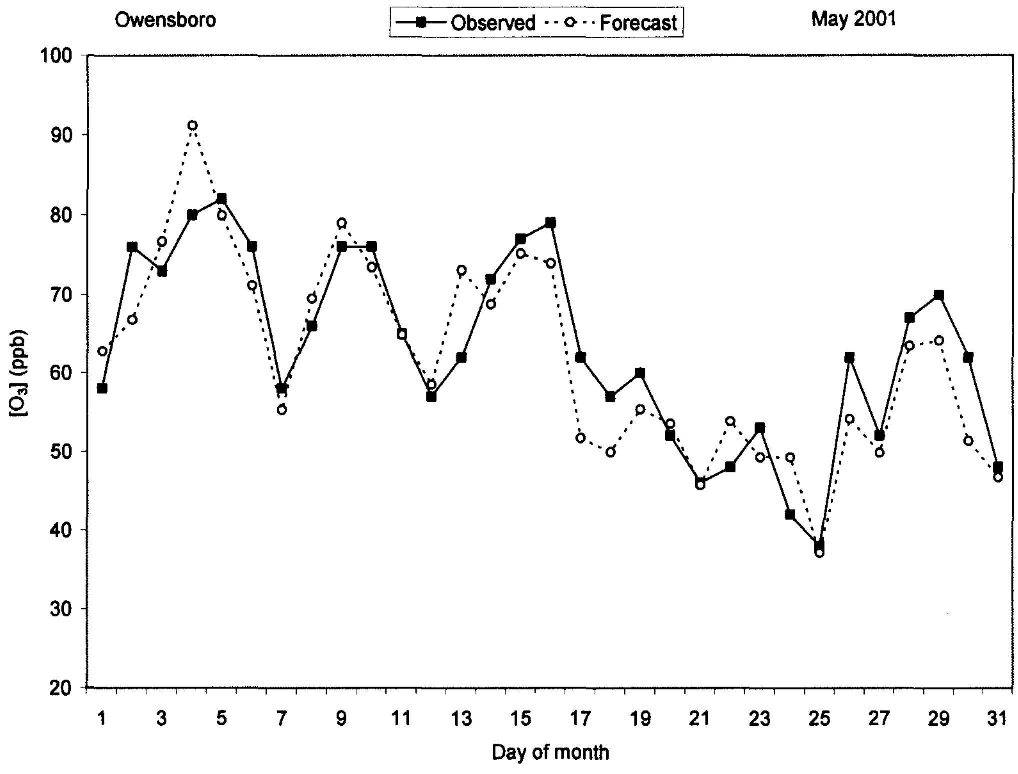

Figure 16. Time Series Comparison of Observed and Model Estimated Ozone Concentration in Owensboro, KY, during May 2001 (Final model).

For the final hybrid model, by using alarm threshold of 80 ppb, a comparatively high DR (0.61) and CSI (0.51), and low FAR (0.29) was obtained (Appendix XXX). The number of alarms (59) was closed to the number of exceedences (62). Among that, only 17 false alarms were issued by the model.

The scatter plot (Figure 17.) showed the good correspondence between the model estimates and the observations. In this figure, most of the dots were very close to the 45 degree diagonal. There were few dots were far from the diagonal, indicates the ozone concentrations on those days were seriously over-predicted or under-predicted.

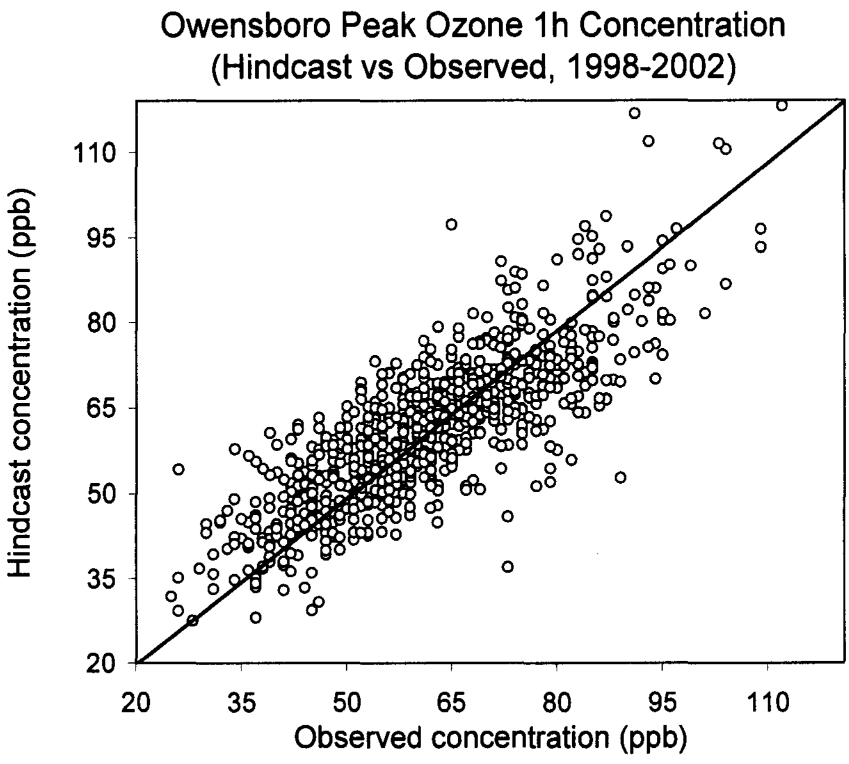

Figure 17. Scatter Plot of Hybrid Prediction against Observations for the Calibration Data Set (Owensboro). (The diagonal indicates the line of perfect correspondence between hindcasts and observations.)

2. Validation with Independent Data Set

The models was recalibrated to 1998-2001 data sets and used to predict the peak ozone concentrations of 2002. The regression coefficients for recalibrated models were listed in Appendix B. The model hindcasts tracked the day-to-day ozone variation reasonably well (Figure 18.).

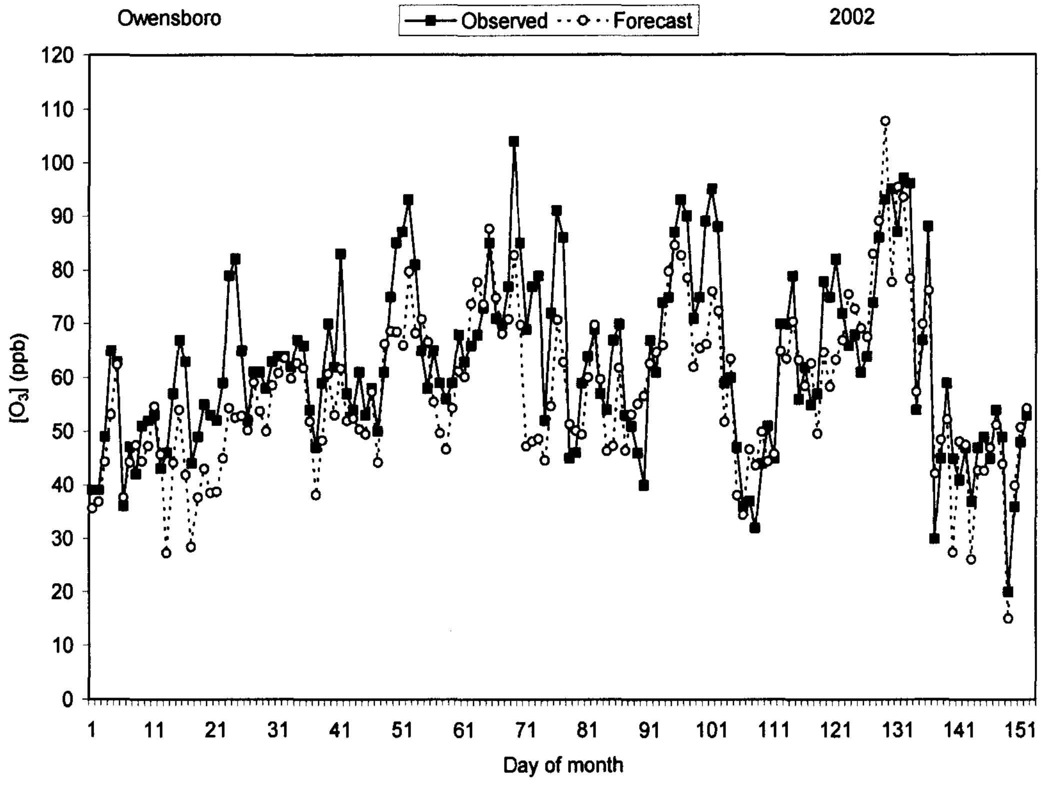

Figure 18. Time Series of Observed and the Re-calibration Model Hindcasts during the 2002 Ozone Season (Owensboro).

The overall statistical comparison of the 2002 re-calibration model hindcasts and original model estimates is presented in Table 18, based upon the alarm threshold value of 80 ppb. The errors for the model estimates were slightly smaller than the model hindcasts. Also the model estimates have better DR, FAR, and CSI than those for the model hindcasts. These results again showed the degradation from model estimates to model hindcasts. The Bias of the model hindcasts and model estimates are -5.3 and -2.2 respectively. The large negative Bias for both of the two models may be caused by some unexplained factors that bumped the average zone concentrations in 2002 in Owensboro.

Table 18

Model Performance Statistics, 2002 Predictions (threshold = 80 ppb, Owensboro)

Index	model hindcasts (model fit:1998-2001)	model estimates (model fit:1998-2002)
Bias (ppb)	-5.3	-2.2
MAE (ppb)	8.7	7.4
MAE/ $\overline{O_3}$ (%)	13.9%	11.8%
RSME (ppb)	10.8	9.4
DR	0.38	0.67
FAR	0.11	0.18
CSI	0.36	0.58

D. Paducah Ozone Forecast Model

1. Performance on calibration data set (1998-2002)

Based on the calibration data set, the R^2 for the final standard model was 0.66. The MAE, RSME, and Bias were 7.41, 9.40, and -0.003 ppb respectively. Compared to the standard model, the hybrid model had a better R^2 of 0.78, and slighter higher MAE, RSME, and Bias (Appendix C). For the final hybrid model, the MAE was 7.67 ppb. Approximate 87% of the absolute errors (667 of 765 days) were less than 15.0 ppb, 71% of the absolute errors (543 of 765 days) were less than 10 ppb.

An example of time series plots for September 2001 showed that the predictions agreed quite closely with the observed concentrations on most days (Figure 19.). On a few days there were comparatively large errors. Time series for the other months shows the similar situations.

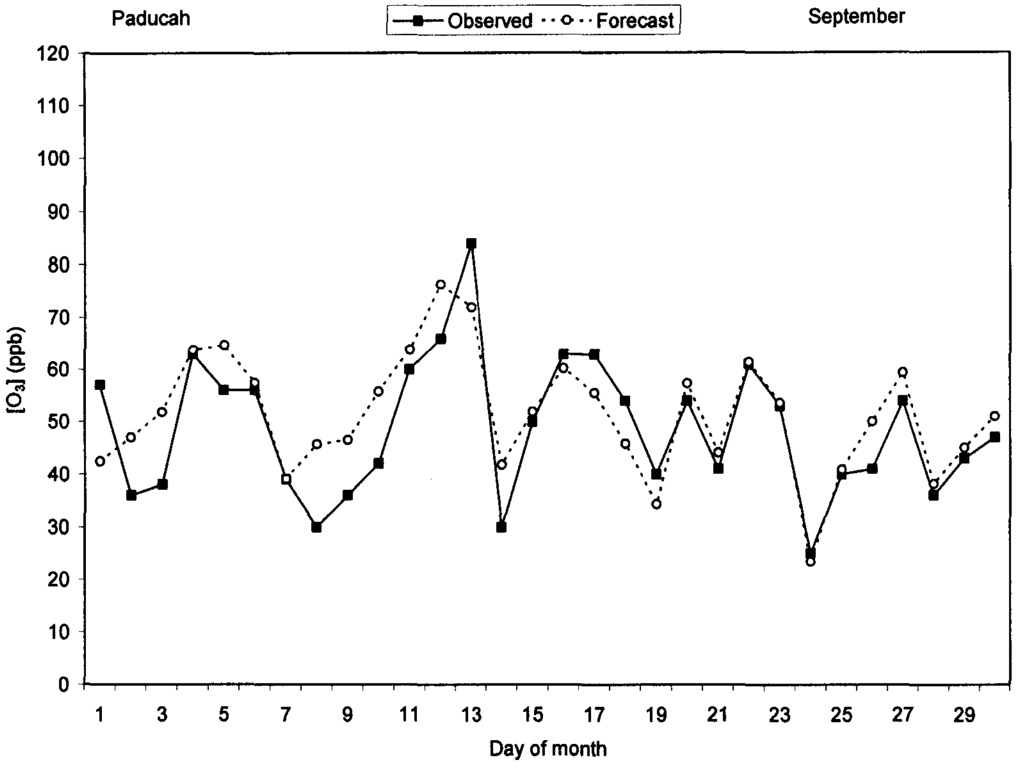

Figure 19. Time Series Comparison of Observed and Model Estimated Ozone Concentration in Paducah, KY, during September 2001 (Final model).

For the final hybrid model, by using alarm threshold of 80 ppb, a comparatively high DR (0.67), CSI (0.55), and low FAR (0.31) were obtained (Appendix XXX). The number of alarms (75) was closed to the number of exceedences (61) and only 23 false alarms were issued by the model.

The scatter plot of the predicted ozone concentrations versus the observed ozone concentrations for the calibration data set is shown in Figure 20. Approximately equal numbers of points lying on both sides of the diagonal line indicate the good correspondence between hindcasts and observations. It can be seen that most of the dots were very close to the 45 degree diagonal.

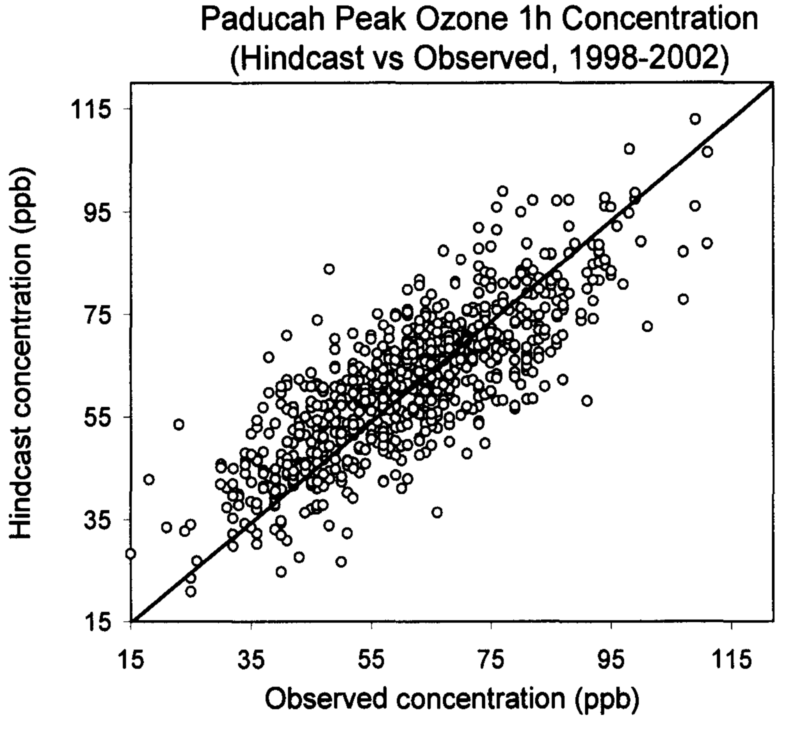

Figure 20. Scatter Plot of Hybrid Prediction against Observations for the Calibration Data Set (Paducah). (The diagonal indicates the line of perfect correspondence between hindcasts and observations.)

2. Validation with Independent Data Set.

The models was recalibrated to 1998-2001 data sets and used to predict the peak ozone concentrations of 2002. The regression coefficients for recalibrated models were listed in Appendix B. The model hindcasts tracked the day-to-day ozone variation reasonably well (Figure 21.).

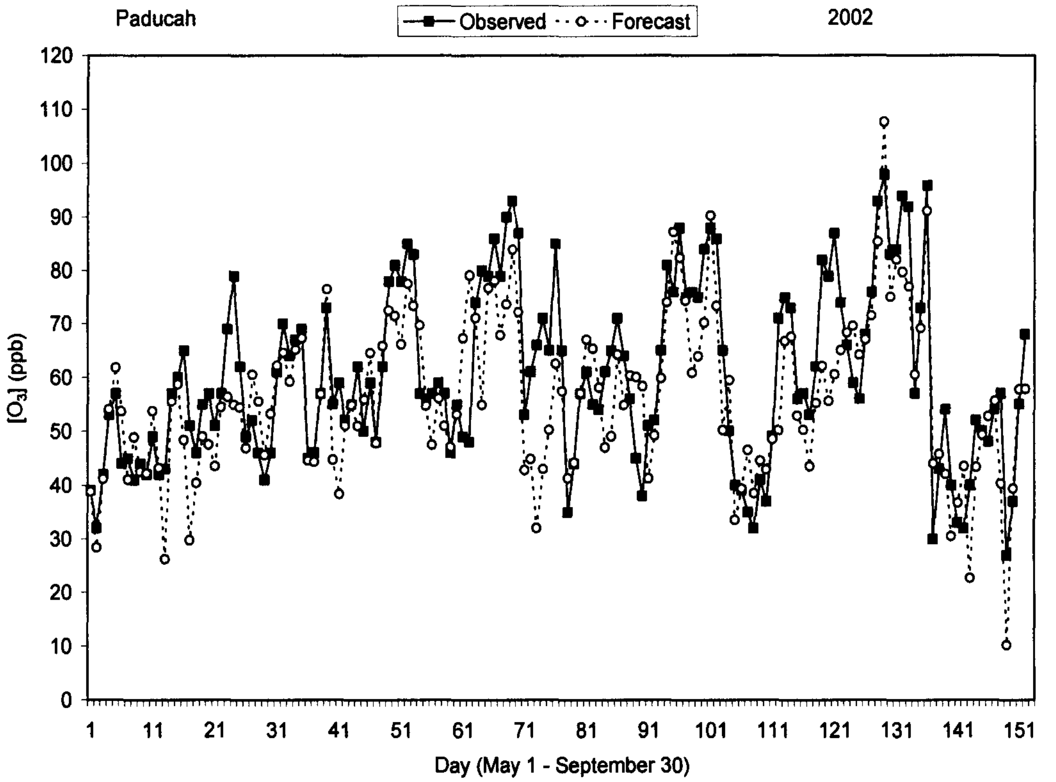

Figure 21. Time Series of Observed and the Re-calibration Model Hindcasts during the 2002 Ozone Season (Paducah).

The overall statistical comparison of the 2002 re-calibration model hindcasts and original model estimates is presented in Table 19, based upon the alarm threshold value of 80 ppb. The performance statistics of the model estimates were better than those of the model hindcasts. The MAE and Bias for the model estimates were 0.9 ppb and 2.1 ppb less than the model hindcasts respectively. Also the DR and FAR for the model estimates were much better than those for the model hindcasts.

Table 19

Model Performance Statistics, 2002 Predictions (threshold = 80 ppb, Paducah)

Index	model hindcasts (model fit:1998-2001)	model estimates (model fit:1998-2002)
Bias (ppb)	-3.9	-1.0
MAE (ppb)	8.4	7.5
MAE/ $\overline{O_3}$ (%)	13.9%	12.4%
RSME (ppb)	10.8	9.8
DR	0.40	0.67
FAR	0.13	0.27
CSI	0.41	0.55

E. Model Performance on 2003 Ozone Season (Lexington)

Our ozone forecast models were designed to predict daily peak ozone in the new ozone season. It's necessary to validate the models by entering the 2003 data. But for 2003 ozone season, the air quality data were not available for Ashland, Bowling Green, Owensboro, and Paducah. It's available only for Lexington. The ozone forecast models for the four ozone control regions in this study were developed from Lexington and Louisville models, all these models have the same model structure and similar parameters. So performance of Lexington 2002 model on 2003 ozone season was evaluated here as a reference for the other models.

Model forecasts were generated by entering the 2003 meteorological forecast data to Lexington model, which calibrated to 1998-2002 model calibration period. The meteorological forecast data obtained from the NGM numerical weather prediction model. The NGM MOS output is available twice daily from internet sites. Here use the second output which is available 30 hours ahead the predicted day. Model hindcasts were also obtained by entering the observed meteorological data from NCDC. The model forecast MAE was 7.30 ppb, which was 14% of the seasonal average O_3. The Bias was 1.80 ppb. The errors for the model forecasts were slightly higher than the errors for the model hindcasts (Table 20). Since the summer of 2003 is relatively cool compared with the previous decades, there was only one ozone action day (O_3>85ppb) in the 2003 ozone season. The DR, FAR, and CSI based on such a small sample size could not reflect the performance of the model. So these statistics are neglected here.

Table 20

Model Performance Statistics for Lexington, 2003 Predictions

Model	Bias (ppb)	MAE (ppb)	MAE/$\overline{O_3}$ (%)
Model forecasts	1.8	7.3	14%
Model hindcasts	1.1	6.0	12%

The time series plot for the model forecasts showed that the Lexington 2002 model predicted peak ozone on 2003 ozone season well (Figure 22). The model correctly predicted the only ozone action day (June 4, 2003). For most of the other days, the model forecasts tracked the observed ozone variation very well.

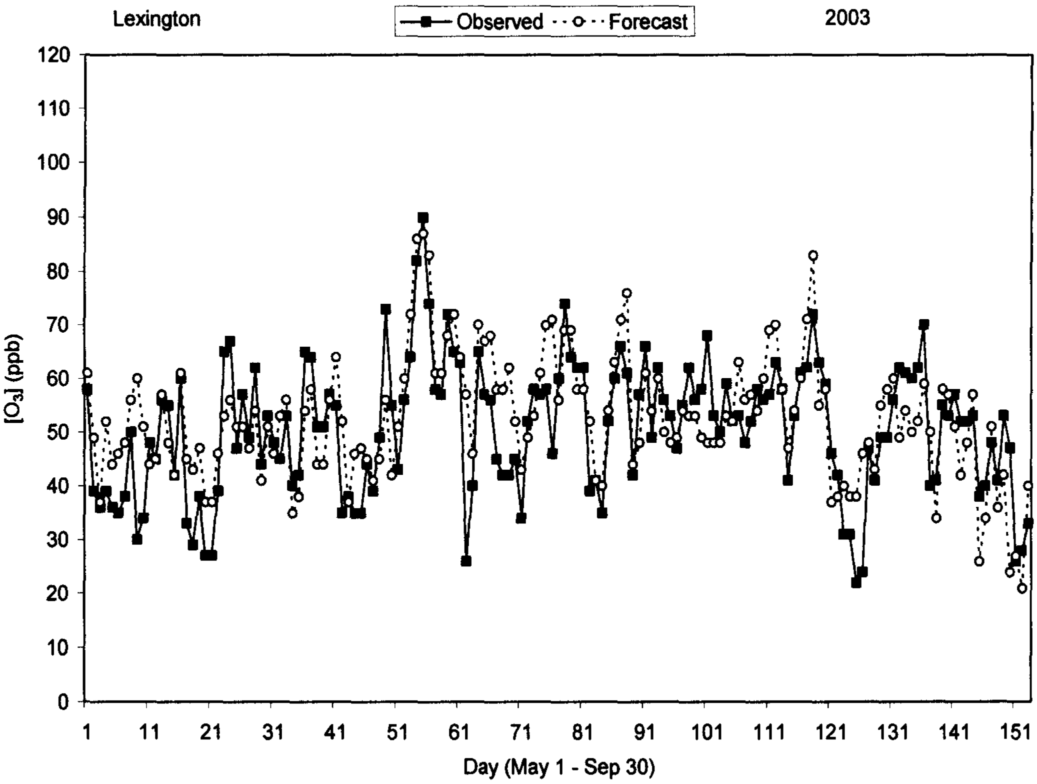

Figure 22. Time Series of Observed and the Model Forecasts during the 2003 Ozone Season (Lexington).

Study on Lexington 2003 ozone season showed that the errors for model forecasts were greater than those of model hindcasts. The reason is that the model predictions are sensitive to small changes in certain meteorological parameters such as temperature and wind speed. The errors contained in the meteorological forecast data increased the errors of the model predictions. The degradation in error is typical when the model using meteorological forecast, instead of observed data. This conclusion was also supported by the experience with using the previous Louisville and Lexington ozone forecast models (Cobourn, 1999).

CHAPTER VII

SUMMARY AND CONCLUSIONS

The multiple-linear regression ozone forecast models for Ashland, Bowling Green, Owensboro, and Paducah were developed, based on the Louisville and Lexington models. The datasets for these models consisted of air quality data and meteorological data during the 1998-2002 calibration period. To identify parameters that significantly correlated with daily peak ozone concentration, graphical and regression methods were used. A standard model and Hi-lo model were developed for each of the four cities. Then a hybrid model was obtained by combining the standard and Hi-lo model using 3S criteria. The hybrid model had better forecast performance beyond what could be achieved using either standard or Hi-lo model alone, especially in predicting the high ozone days. Model performances on calibration data set were evaluated. Also these models were recalibrated to 1998-2001 period and were used to predict peak ozone on 2002 ozone season, to test the model performance on independent data set.

Conclusions:

1. The multiple nonlinear regression models were successfully applied to daily ozone forecast for the middle metropolitan areas, Ashland, Bowling Green, Owensboro, and Paducah, as well as for the large metropolitan area Louisville and Lexington. For the model estimates that based on 1998-2002 calibration period, the MAE was less than 7.7 ppb; the MAE/$\overline{O_3}$ was less than 12.7% for each of the cities. The models

could explain at least 66% of the variance of the daily peak ozone. Recalibrating the models to 1998-2001 period, then using the models to predict the ozone concentrations in 2002 ozone season, the MAE ranged from 8.1 to 9.5 ppb; the MAE/$\overline{O_3}$ ranged from 13.8% to 15.7%. The values of these statistical parameters were close to those for Louisville and Lexington models.

2. With the 80 ppb threshold, the model estimates had relatively high detection rate (ranged from 0.48 to 0.67) and low false alarm number (ranged from 17 to 40) for the four ozone control regions. That means most of the ozone action days had been detected without issuing too many false alarms. A lower threshold (such 75 ppb) may lead to significant decrease of the detection rate; a higher threshold (such 85 ppb) usually cause too many false alarms issued. So an alarm threshold value of 80 ppb was recommended for the models developed in this study.

3. By developing the ozone forecast models for the four cities, a group of parameters was found to be strongly correlated with peak ozone in the regressions for all the cities. These parameters, called common parameters, included the nonlinear term, xmitt, trend, and rhx2. An ozone forecast model with the common parameters only, called the common model, could explain most (usually more than 90%) of the ozone variation explained by the complete model (includes both common parameters and additional parameters). The common model concept is a tool for simplifying the process in developing an ozone forecast model in a new area.

4. The trajectory parameter was not used in the ozone forecast models in this study, since these models were designed to be automatically operated by computer. Applying the trajectory parameter to the model usually can improve the MAE for the hybrid model by around 1.0 ppb.

REFERENCES

Ambroise C. and Grandvalet Y. (2001) Prediction of Ozone Peaks by Mixture Models. *Ecological Modeling*, 145, 275-289.

Balaguer Ballester E., Camps i Valls G., Carrasco-Rodriguez J.L., Soria Olivas E., and del Valle-Tascon S. (2002) Effective 1-day Ahead Prediction of Hourly Surface Ozone Concentrations in Eastern Spain Using Linear Models and Neural Networks. *Ecological Modeling*, 156, 27-41

Bloomfield P., Royle J. A., Steinberg L. J., and Yang Q. (1996) Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends. *Atmospheric Environment*, 30, 3067-3077

Chen K. S., Ho Y. T., Lai C. H., and Chou Y. M. (2003) Photochemical Modeling and Analysis of Meteorological Parameters During Ozone Episodes in Kaohsiung, Taiwan. *Atmospheric Environment*, 37, 1811-1823

Cobourn W. G. and Hubbard M. C. (1999) An Enhanced Ozone Forecasting Model Using Air Mass Trajectory Analysis. *Atmospheric Environment*, 33, 4663-4674

Cobourn W. G. (2000) A Comparison of Nonlinear Regression and Neural Network Models for Ground-Level Ozone Forecasting. *Air & Waste Manage*, 50, 1999-2009

Cobourn W. G. (2003) Personal Communication.

Comrie A. C. (1997) Comparing Neural Networks and Regression Models for Ozone Forecasting. *Air & Waste Manage*, 47, 653-663

Davis J. M. and Speckman P. (1999) A Model for Predicting Maximum and 8h Average Ozone in Houston. *Atmospheric Environment*, 33, 2487-2500

De Nevers N. (1995) Air Pollution Control Engineering. McGraw-Hill, Inc., New York

Diem J. E. and Comrie A. C. (2002) Predictive Mapping of Air Pollution Involving Sparse Spatial Observations. *Environmental Pollution*, 119, 99-117

Elkamel A., Abdul-Wahab S., Bouhamra W., and Alper E. (2001) Measurement and Prediction of Ozone Levels around a Heavily Industrialized Area: A Neural Network Approach. *Advances in Environmental Research*, 5, 47-59

EPA (2002) Clean Air Markets – Programs and Regulations: Nitrogen Oxides (NOx) Reduction under the Acid Rain Program. URL: http://www.epa.gov/air-markt/arp/nox/index.html

EPA (2003a) Clean Air Markets – Programs and Regulations: EPA's NOx Budget Trading Program - The NOx SIP Call. URL: http://www.epa.gov/airmarkt/fednox/in-dex.html

EPA (2003b) Air Trend: Six Principal Pollutants. URL: http://www.epa.gov/airtrends/six-poll.html

EPA (2003c) Green Book: Ozone Nonattainment Area Summary. URL: http://www.epa.-gov/oar/oaqps/greenbk/onsum.html

EPA (2003d) Public Affairs – 2001 Archive: Louisville Area Meets Ozone Standard. URL: http://www.epa.gov/region4/oeapages/01press/010618.htm

EPA (2003e) Green Book: Ozone Nonattainment Area/State/County Report. URL: http://www.epa.gov/oar/oaqps/greenbk/onca.html

EPA (2003e) AirData - Monitor Locator Map for Kentucky (Ozone). URL: http://www.epa.gov/air/data/monloc.html

Flemming J., Reimer E., and Stern R (2001) Long Term Evaluation of the Ozone Forecast by an Eulerian Model. *Physics and Chemistry of the Earth*, 26, 775-779

Gardner M. W. and Dorling S. R. (2000) Statistical Surface Ozone Models: An Improved Methodology to Account for Non-linear Behavior. *Atmospheric Environment*, 34, 21-34

Greenwell C. G. (2000) A Regression Model to Forecast the Daily Peak 8-h Ozone Concentration for the Louisville Metropolitan Statistical Area. M.S. thesis. University of Louisville

Hubbard M. C. and Cobourn W. G. (1997) Development of a Regression Model to Forecast Ground-Level Ozone Concentrations in Jefferson County, Kentucky. Atmospheric Environment, 32, 2637-2647

Hubbard M. C. (1997) Development of a Regression Model to Forecast Ground-Level Ozone Concentrations in Jefferson County, Kentucky. M.S. thesis. University of Louisville

Jorquera H., Perez R., Cipriano A, Espejo A., Letelier M. V. and Acuna G. (1998) Forecasting Ozone Daily Maximum Levels at Santiago, Chile. *Atmospheric Environment*, 32, 3415-3424

Lin C.Y.C., Jacob D. J., fiore A. M. (2001) Trends in Exceedences of the Ozone Air Quality Standard in the Continental United States, 1980-1998. *Atmospheric Environment*, 35, 3217-3228

Lissen G., Debruyn W., Mensink C. and Dumont G. (2000) SMOGSTOP: A Model for Forecasting Maximum Daily Ozone Concentration in Belgium. *Environmetrics*, 11, 511-521

Lomax R.G. (2001) Statistical Concepts: A Second Course for Education and the Behavioral Sciences. Lawrence Erlbaum Associates, Inc. Mahwah, NJ

Narasimhan R., Keller J., Subramaniam G., Raasch E., Croley B., Duncan K. and Potter W. T. (2000) Ozone Modeling Using Neural Networks, *Journal of Applied Meteorology*, 39, 291-296

NCDC (1998-2002) *Local Climatological Data-Huntington, WV*. National Climatic Data Center, Asheville, North Carolina 28801. (Ashland)

NCDC (1998-2002) *Local Climatological Data-Bowling Green, KY*. National Climatic Data Center, Asheville, North Carolina 28801.

NCDC (1998-2002) *Local Climatological Data-Evansville, IN*. National Climatic Data Center, Asheville, North Carolina 28801. (Owensboro)

NCDC (1998-2002) *Local Climatological Data-Paducah, KY*. National Climatic Data Center, Asheville, North Carolina 28801.

NCDC (2003) *1971-2000 Daily/Monthly Station Normal*. National Climatic Data Center, Asheville, North Carolina 28801.

NOAA (2003). HYSPLIT (Hybrid Single-Particle Lagrangian Integrated Trajectory) Model. URL: http://www.arl.noaa.gov/ready/sec/hysplit4.html, NOAA Air Resources Laboratory, Silver Spring, MD, USA

Pokrovsky O. M., Kwok R. and Ng C. N. (2002) Fuzzy Logic Approach for Description of Meteorological Impacts on Urban Air Pollution Species: A Hong Kong Case Study. *Computers & Geosciences*, 28, 119-127

Prybutok V. R., Yi J. and Mitchell D. (2000) Comparison of Neural Network Models with ARIMA and Regression Models for Prediction of Houston's Daily Maximum Ozone Concentrations. *European Journal of Operational Research*, 122, 31-40

Robeson S. M. and Steyn D. G. (1990) Evaluation and Comparison of Statistical Forecast Models for Daily Maximum Ozone Concentrations. *Atmospheric Environment*, 24B, 303-312

Russell A. and Dennis R. (2000) NARSTO Critical Review of Photochemical Models and Modeling. *Atmospheric Environment*, 34, 2283-2324

Ryoke M., Nakamori Y., Heyes C., Makowski M. and Schopp W. (2000) A Simplified Ozone Model Based on Fuzzy Rules Generation, *European Journal of Operational Research*, 122, 440-451

Slini Th., Karatzas K. and Moussiopoulos N. (2002) Statistical Analysis of Environmental Data as the Basis of Forecasting: An Air Quality Application. *The Science of the Total Environment*, 288, 227-237

Spellman G. (1999) An Application of Artificial Neural Networks to the Prediction of Surface Ozone Concentrations in the United Kingdom. *Applied Geography*, 19, 123-136

Vingarzan R. and Taylor B. (2003) Trend Analysis of Ground Level Ozone in the Greater Vancouver/Fraser Valley Area of British Columbia. *Atmospheric Environment*, 37, 2159-2171

Vukovich F. M., Gilliland A., Venkatram A. and Sherwell J. (2001) On Performing Long-term Predictions of Ozone Using the SOMS Model. *Atmospheric Environment*, 35, 569-578

Wang W., Lu W., Wang X. and Leung A. Y. T. (2003) Prediction of Maximum Daily Ozone Level Using Combined Neural Network and Statistical Characteristics. Environment International, 29, 555-562

Wotawa G., Stohl A. and Neininger B. (1998) The Urban Plume of Vienna: Comparisons Between Aircraft Measurements and Photochemical Model Results. Atmospheric Environment, 32, 2479-2489

APPENDIX A.

Table 21. Parameters and Coefficients for the Final Standard Models for Ashland, Bowling Green, Owensboro, and Paducah (Calibrated to 1998-2002 period)

	Variables	Coef.	ASH	BWG	OWE	PAH
	const	b_0	-167.09	-214.7	-235.7	-180.4
1	nonlin	b_1	0.793	0.73	0.92	0.77
2	xmitt	b_2	307.46	393.1	418	329.8
3	trend	b_3	-1.226	-0.898	-1.779	-1.436
4	rhx2	b_4	-0.115	-0.298	-0.185	-0.206
5	tmn_dep	b_5		0.26	0.395	0.174
6	cc	b_6	-0.684	-0.447		-0.845
7	mdwind	b_7	-0.333	-0.43	-0.32	-0.336
8	dewpt	b_8			-0.195	
9	hol	b_9			-6.62	

Table 22. Parameters and Coefficients for the Final Hi-lo Models for Ashland, Bowling Green, Owensboro, and Paducah (Calibrated to 1998-2002 period)

	Variables		ASH	BWG	OWE	PAH
	const	b_0	-166.74	-245.1	-184.2	-182.7
1	nonlin	b_1	0.761	0.839	1.004	1.06
2	xmitt	b_2	316.25	437.2	368.5	345.9
3	trend	b_3	-1.898	-1.201	-2.254	-1.865
4	rhx2	b_4	-0.129	-0.424		-0.2
5	tmn_dep	b_5			0.658	
6	cc	b_6	-0.756			
7	mdwind	b_7			-0.342	-0.581
8	dewpt	b_8			-0.68	
9	hol	b_9				
10	tmn_nrm	b_{10}				-0.421

APPENDIX B.

Table 23. Parameters and Coefficients for the Standard Models for Ashland, Bowling Green, Owensboro, and Paducah (Recalibrated to 1998-2001 period)

	Variables	ASH	BWG	OWE	PAH
	const	-146.9233	-222.8361	-201.4248	-184.8941
1	nonlin	0.7803	0.695	0.8705	0.7499
2	xmitt	268.3537	414.0221	369.7667	340.907
3	trend	-0.9152	-1.7816	-2.7137	-1.8821
4	rhx2	-0.1132	-0.3482	-0.2173	-0.23
5	tmn_dep		0.2901	0.3881	0.1986
6	cc	-0.8054	-0.3536		-0.8083
7	mdwind	-0.3068	-0.4227	-0.3189	-0.3597
8	dewpt			-0.1553	
9	hol			-7.75	

Table 24. Parameters and Coefficients for the Hi-lo Models for Ashland, Bowling Green, Owensboro, and Paducah (Recalibrated to 1998-2001 period)

	Variables	ASH	BWG	OWE	PAH
	const	-140.0745	-245.0595	-158.4158	-177.3784
1	nonlin	0.7786	0.5548	0.976	0.8692
2	xmitt	262.1739	479.1893	335.8867	337.636
3	trend	-0.799	-0.39871	-3.4289	-3.2263
4	rhx2	-0.1406	-0.4918		-0.1518
5	tmn_dep			0.6446	
6	cc	-0.7757			
7	mdwind			-0.4049	-0.1512
8	dewpt			-0.7113	
9	hol				
10	tmn_nrm				-0.2545

APPENDIX C.

Table 25. Statistics for the Models for Ashland, Bowling Green, Owensboro, and Paducah, Performance on Calibration Period 1998-2002 (1)

standard model statistics				
Index	Ashland	Bowling Green	Owensboro	Paducah
MAE (ppb)	7.06	6.79	6.84	7.41
RSME (ppb)	9.01	8.77	8.72	9.40
Bias (ppb)	-0.002	0.003	-0.003	-0.003
R^2	0.70	0.68	0.68	0.66

hybrid model statistics				
Index	Ashland	Bowling Green	Owensboro	Paducah
MAE (ppb)	7.29	6.96	7.07	7.67
MAE/ $\overline{O_3}$ (%)	12.0%	11.5%	11.3%	12.7%
RSME (ppb)	9.22	9.06	9.03	9.72
Bias (ppb)	0.94	0.63	-0.12	-0.04
R^2	0.86	0.84	0.78	0.78

Table 26. Statistics for the Models for Ashland, Bowling Green, Owensboro, and Paducah, Performance on Calibration Period 1998-2002 (2)

threshold = 80 ppb

parameter	symbol	Ashland	Bowling Green	Owensboro	Paducah
detection rate	DR	0.62	0.48	0.61	0.67
false alarm rate	FAR	0.44	0.38	0.29	0.31
critical success index	CSI	0.46	0.45	0.51	0.55
events	EV	111	102	83	95
exceedences	EX	53	54	62	61
detected exeedences	DE	33	26	38	41
alarms	AL	91	74	59	75
false alarms	FA	40	28	17	23

APPENDIX D.

Sample Airnow Map for National Ozone

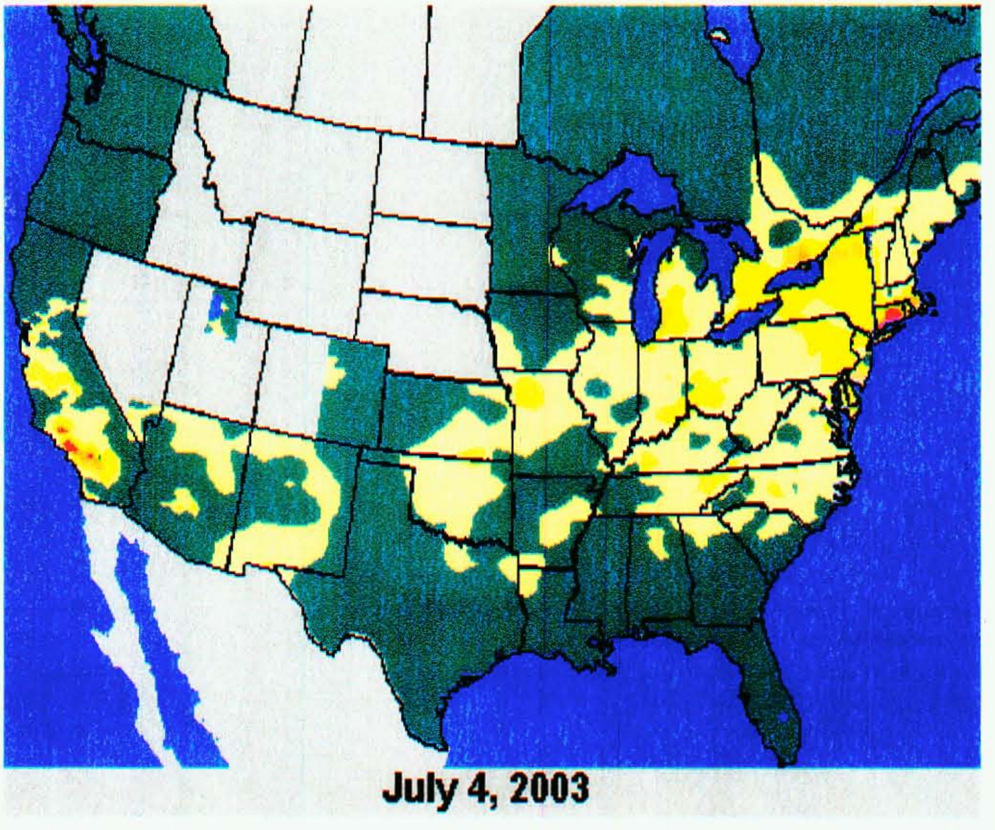

Figure 23. Airnow 8-h Ozone Concentration Contour Map

APPENDIX E.

Sample 48-hr Backward Hindcast Trajectories

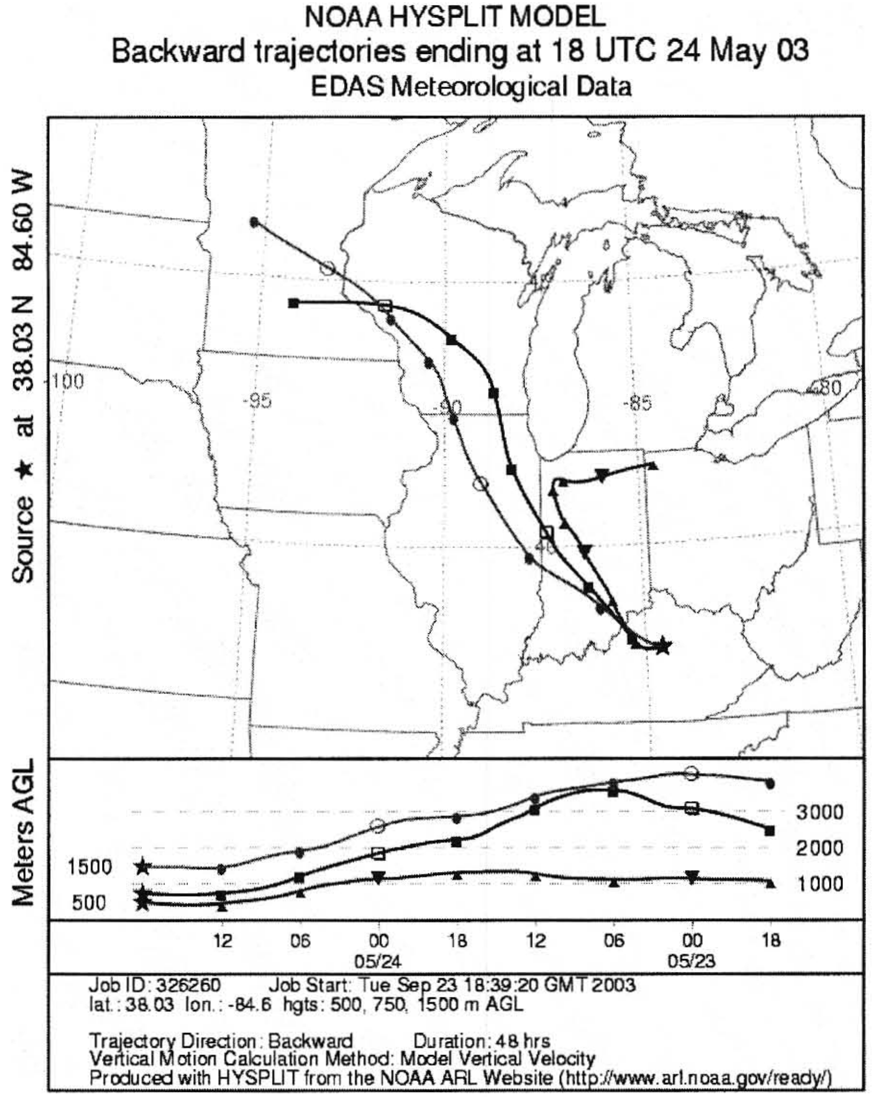

Figure 24. 48-hr Backward Hindcast Trajectories for Lexington

This trajectory data produced using HYSPLIT model, provided by Air Resources Laboratory, NOAA. URL: http://www.arl.noaa.gov/ready/

APPENDIX F.

Sample 36-hr Backward Forecast Trajectory

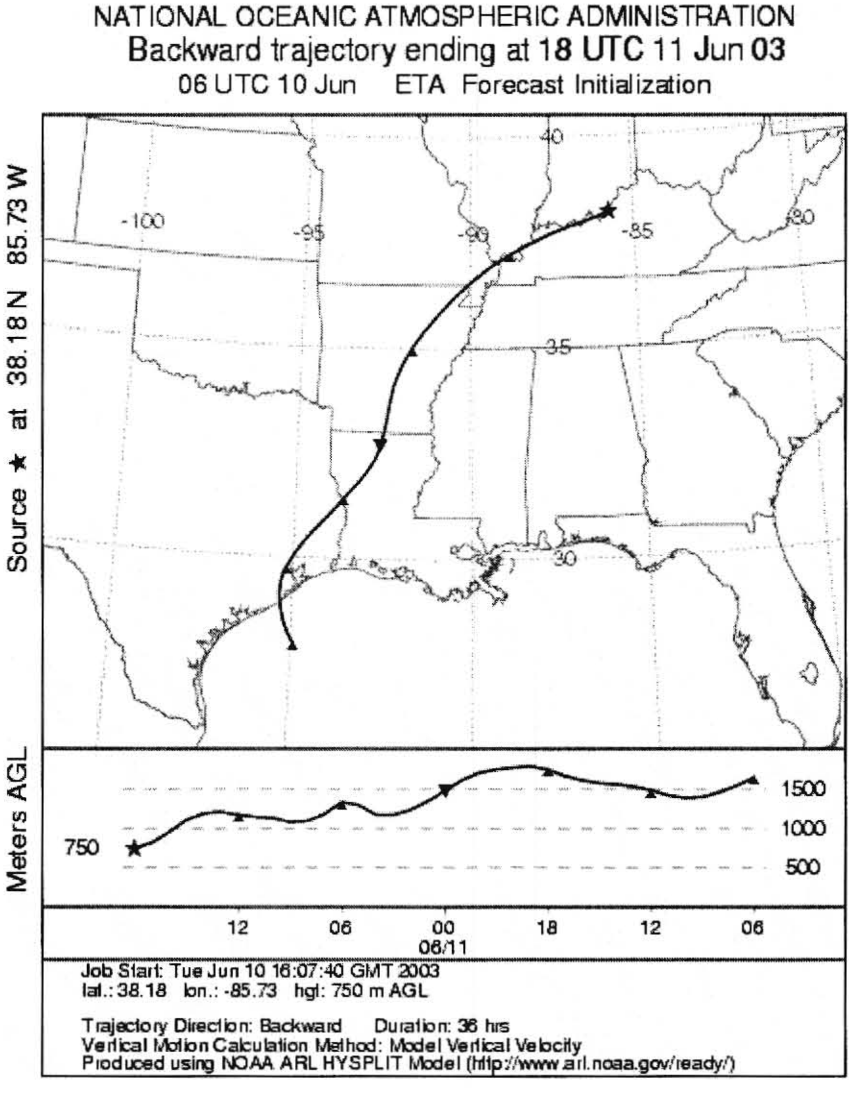

Figure 25. 36-hr Backward Forecast Trajectory for Louisville

This trajectory data produced using HYSPLIT model, provided by NOAA.
URL: http://www.arl.noaa.gov/ready/.

APPENDIX G.

Sample Dataset of NCDC Weather Observations

```
"AUGUST 2002                HUNTINGTON, WV                  "
"TRI-STATE AIRPORT                 (HTS)"
"Lat: 38 22' N     Long: 82 33' W     Elev (Ground):    822 Feet"
"Time Zone:     EASTERN       WBAN: 03860  ISSN #:0198-5655"

(1)
01, 93, 67, 80,  5 , 69, 73,  0, 15,"BR
02, 95, 69, 82,  7 , 70, 73,  0, 17,"TS BR HZ
03, 96, 68, 82,  7 ,   ,   ,  0, 17,"TS TSRA RA FG+ BR HZ
04, 94, 69, 82,  7 , 71, 74,  0, 17,"RA FG+ BR
05, 93, 68, 81,  6 , 71, 74,  0, 16,"RA FG+ BR
06, 80, 66, 73, -2 , 56, 63,  0,  8,"
07, 80, 57, 69, -6 , 52, 59,  0,  4,"
08, 82, 57, 70, -5 , 52, 60,  0,  5,"
09, 87, 54, 71, -4 , 54, 61,  0,  6,"
10, 91, 59, 75,  0 , 61, 66,  0, 10,"
11, 91, 67, 79,  4 , 68, 70,  0, 14,"TS TSRA RA
......

(2)
01,01,"CLR",NC ,    , ,  7.00,"                ", 72, 70, 71, 94,
01,02,"CLR",NC ,    , ,  9.00,"                ", 71, 68, 69, 90,
01,03,"CLR",NC ,    , ,  6.00,"BR              ", 70, 68, 69, 93,
01,04,"CLR",NC ,    , ,  5.00,"BR              ", 69, 67, 68, 93,
01,05,"CLR",NC ,    , ,  5.00,"BR              ", 69, 67, 68, 93,
01,06,"CLR",NC ,    , ,  3.00,"BR              ", 68, 67, 67, 96,
01,07,"CLR",NC ,    , ,  5.00,"BR              ", 70, 69, 69, 97,
01,09,"CLR",NC ,    , , 10.00,"                ", 7 9, 72, 74, 79,
01,10,"CLR",NC ,    , , 10.00,"                ", 83, 71, 75, 67,
01,11,"CLR",NC ,    , , 10.00,"                ", 87, 69, 75, 55,
01,12,"CLR",NC ,    , , 10.00,"                ", 88, 70, 75, 55,
01,13,"SCT",NC ,    , , 10.00,"                ", 90, 70, 76, 52,
01,14,"FEW",NC ,    , , 10.00,"                ", 90, 69, 75, 50,
01,15,"FEW",NC ,    , , 10.00,"                ", 92, 66, 74, 43,
01,16,"SCT",NC ,    , , 10.00,"                ", 92, 68, 75, 46,
01,17,"BKN",065,    , , 10.00,"                ", 91, 70, 76, 50,
01,18,"BKN",060,    , , 10.00,"                ", 89, 70, 76, 53,
......
```

Part (1) consists of the daily peak observations. Part (2) consists of the hourly weather observations.

APPENDIX H.

Sample Ozone Calculator with Friendly Interface

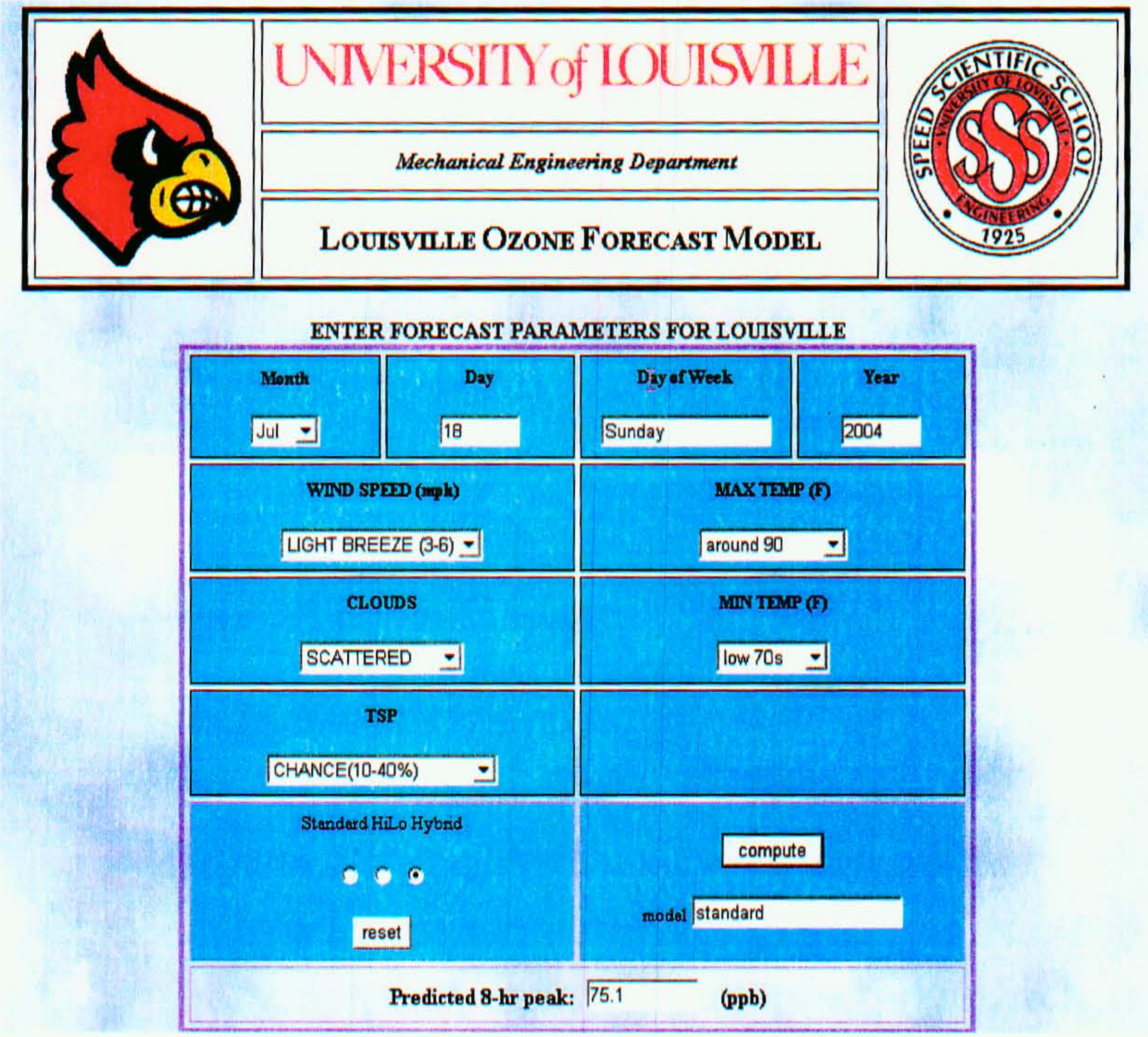

Figure 26. Ozone Calculator for Louisville with Friendly Interface

This calculator based on the non-trajectory ozone forecast model.
Available at: http://www.louisville.edu/~wgcobo01/ozone/ozcalc_sdf.htm

CURRICULUM VITAE

NAME: Yiqiu Lin

ADDRESS: Department of Mechanical Engineering
University of Louisville
Louisville, KY 40292

DOB: Pingdingshan, Henan, China – September 5, 1971

EDUCATION
& TRAINING: B.S., Mechanical Engineering
University of Tianjin, China
1989-93

Worked as Mechanical Engineer
No.6 Institute of Design and Research, China
1993-2001

Pursue the M.S. and PhD degree in
University of Louisville
2001-04

PROFESSIONAL SOCIETIES:
Tau Beta Pi, the Engineering Honor Society