5-2007

Development of fuzzy system and nonlinear regression models for ozone and PM2.5 air quality forecasts.

Yiqiu Lin 1971-
University of Louisville

DEVELOPMENT OF FUZZY SYSTEM AND NONLINEAR REGRESSION MODELS FOR OZONE AND $PM_{2.5}$ AIR QUALITY FORECASTS

By

Yiqiu Lin
B.S., University of Tianjin, China, 1993
M.S., University of Louisville, 2004

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

DOCTOR OF PHILOSOPHY

Department of Mechanical Engineering
University of Louisville
Louisville, KY

May 2007

DEVELOPMENT OF FUZZY SYSTEM AND NONLINEAR REGRESSION MODELS FOR OZONE AND $PM_{2.5}$ AIR QUALITY FORECASTS

Submitted by: ____________________________
Yiqiu Lin

A Dissertation Approved on

March 8, 2007

by the Following Dissertation Committee:

____
Dissertation Director, Dr. W. Geoffrey Cobourn

____
Dr. Julius Wong

____
Dr. John Lilly

____
Dr. Ellen Brehob

____
Dr. Christopher Richards

DEDICATION

This dissertation is dedicated to my parents:

Shipei Lin and Caiyun Huang

And to my wife and son:

Dongmei Zhang and Kevin Y. Lin

ACKNOWLEDGEMENTS

I would like to gratefully and sincerely thank my advisor and mentor Dr. W. Geoffrey Cobourn for his guidance, understanding, patience, and his friendship during my graduate studies at University of Louisville. He encouraged me to face any challenges during the research and the writing of this dissertation. His wisdom, knowledge, and commitment to the highest standards inspired and motivated me. I would also like to thank Dr. Julius Wong, Dr. John Lilly, Dr. Ellen Brehob, and Dr. Christopher Richards for comments and suggestions on this dissertation.

I wish to thank Jerry Sudduth, Tim Smith, Marty Layman, and Steve Sherman for providing the air quality data. The NLR air quality models were developed with support from the federal Congestion Mitigation and Air Quality (CMAQ) Improvement Program, administered through the Kentucky Transportation Cabinet. I also gratefully acknowledge partial support from a University Fellowship from the University of Louisville Research Foundation.

Finally, and most importantly, I would thank my wife Dongmei Zhang, not only for her support, encouragement, and unwavering love, but also for the most amazing miracle she created: our son Kevin Y. Lin. I also wish gratefully thank my parents, Shipei Lin and Caiyun Huang, for giving me invaluable educational opportunities. With all my heart, I dedicate this dissertation to my family.

ABSTRACT

DEVELOPMENT OF FUZZY SYSTEM AND NONLINEAR REGRESSION MODELS FOR OZONE AND $PM_{2.5}$ AIR QUALITY FORECASTS

Yiqiu Lin

March 8, 2007

Ozone forecast models using nonlinear regression (NLR) have been successfully applied to daily ozone forecast for seven metro areas in Kentucky, including Ashland, Bowling Green, Covington, Lexington, Louisville, Owensboro, and Paducah. In this study, the updated 2005 NLR ozone forecast models for these metro areas were evaluated on both the calibration data sets and independent data sets. These NLR ozone forecast models explained at least 72% of the variance of the daily peak ozone. Using the models to predict the ozone concentrations during the 2005 ozone season, the metro area mean absolute errors (MAEs) of the model hindcasts ranged from 5.90 ppb to 7.20 ppb. For the model raw forecasts, the metro area MAEs ranged from 7.90 ppb to 9.80 ppb.

Based on previously developed NLR ozone forecast models for those areas, Takagi-Sugeno fuzzy system models were developed for the seven metro areas. The fuzzy ''c-means'' clustering technique coupled with an optimal output predefuzzification approach (least square method) was used to train the Takagi–Sugeno fuzzy system. Two types of fuzzy models, basic fuzzy and NLR-fuzzy system models, were developed. The basic fuzzy and NLR-fuzzy models exhibited essentially equivalent performance to the existing NLR models on 2004 ozone season hindcasts and forecasts. Both types of fuzzy models had, on average, slightly lower metro area averaged MAEs than the NLR models.

Among the seven Kentucky metro areas Ashland, Covington, and Louisville are currently designated nonattainment areas for both ground level O_3 and $PM_{2.5}$. In this study, summer $PM_{2.5}$ forecast models were developed for providing daily average $PM_{2.5}$ forecasts for the seven metro areas. The performance of the $PM_{2.5}$ forecast models was generally not as good as that of the ozone forecast models. For the summer 2004 model hindcasts, the metro-area average MAE was $5.33\mu g/m^3$.

Exploratory research was conducted to find the relationship between the winter $PM_{2.5}$ concentrations and the meteorological parameters and other derived prediction parameters. Winter $PM_{2.5}$ forecast models were developed for seven selected metro areas in Kentucky. For the model fits, the MAE for the seven forecast models ranged from 3.23 $\mu g/m^3$ to 4.61 $\mu g/m^3$ (~26 – 28% NMAE). The fuzzy technique was also applied on $PM_{2.5}$ forecast models to seek more accurate $PM_{2.5}$ prediction. The NLR-fuzzy $PM_{2.5}$ had slightly better performance than the NLR models.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Ground level ozone (O_3) is one of the six criteria air pollutants that are commonly found throughout the United States. The six criteria pollutants consist of ground level O_3, fine particulate matter ($PM_{2.5}$), carbon monoxide (CO), nitrogen dioxide (NO_x), sulfur dioxide (SO_2), and lead. Ground level O_3, even at low levels, can adversely affect human health. Prolonged exposure to O_3 concentrations over a certain level may cause severe health problems including permanent lung damage, aggravated asthma, or other respiratory illnesses. Ground level O_3 can also have detrimental effects on plants and ecosystems, including damage to plants, reductions of crop yield, and increase of vegetation vulnerability to disease (EPA, 2005a).

The U.S. Environmental Protection Agency (EPA) has set National Ambient Air Quality Standards (NAAQS) for the six criteria pollutants to protect public health (primary standard) and public welfare (secondary standard). Before 1997, the sole NAAQS for ozone was based on 1-hr average concentration, not to exceed 0.12 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 promulgated a more protective standard for ozone. The new primary standard and secondary standard are the same, viz.

0.08 ppm for 8-hr average ozone concentration. To attain this standard, the 3-year average of the fourth-highest daily maximum 8-hour average ozone concentrations measured at each monitor within an area over each year must not exceed 0.08 ppm (EPA, 2005b). Though ground-level O_3 air quality has significantly improved in the past two decades, it remains a critical problem in many communities in the US. Currently there are 255 counties designated as ozone nonattainment areas, based on the NAAQS. Moreover, twenty million people live in nineteen counties designated as "Serious" or "Severe" ozone nonattainment areas (EPA, 2005c).

An accurate ozone forecast model can be used to issue alerts in anticipation of high O_3 levels so that community action can be taken to reduce the emission of O_3-forming compounds in order to avoid a NAAQS exceedence (Hubbard, 1997). It also can provide advanced warning of potentially unhealthful air quality for the people living in these areas. Since 1997, nonlinear regression (NLR) ozone forecast models have been developed and implemented by the University of Louisville for the Louisville metropolitan statistical area (MSA). The Louisville MSA was one of the ozone nonattainment areas in Kentucky. This NLR model uses a group of meteorological parameters as the input predictor variables. It was designed to predict the daily maximum 8-hr average O_3 concentrations among all of the ozone monitors within the metro area. Following the successful implementation of the NLR model for Louisville, more NLR ozone forecast models were developed for selected metro areas in Kentucky. In 2005, there were seven NLR models running on an automated basis, providing ozone forecasts for the Ashland, Owensboro, Bowling Green, Covington, Lexington, Louisville, and Paducah metro areas. These models have been updated each year, using the meteorological and air quality data from the most recent five year period. One of the objectives in this dissertation is evaluating the 2005 NLR ozone forecast models for the seven metro areas in Kentucky.

Previous NLR models have performed well on ozone prediction. For example, for the seven models fitted to 1999-2003 databases, the mean absolute error (MAE) of the fit was typically about 7 ppb, or about 12% of the mean daily peak ozone concentration for the period. Typically the MAEs of the 2004 ozone season forecasts were only about 1- 2 ppb higher than for the original model fits, working out to about 15% of the seasonal mean concentrations. These model performance statistics compare favorably with those of other O_3 air quality models reported in the literature.

Fuzzy modeling is a tool aimed at using the information observed from a complex phenomenon to derive a quantitative model. In recent years, fuzzy methods have been applied to air pollutant forecasting. It has been reported in several papers that fuzzy models performed well on ozone forecasts (Jorquera et al, 1998, Heo et al, 2004, and Ryoke et al, 2000). Developments of the NLR ozone forecast models provided complete databases and a group of ozone predictor variables for development of fuzzy models. Another objective of this dissertation was to develop ozone fuzzy system forecast models for the seven metro areas and compare the performance of fuzzy models and NLR models. A secondary objective, which arose during the course of the research, was to construct combined NLR-Fuzzy system models for comparison with the NLR models.

Fine particulate matter ($PM_{2.5}$) is an important pollutant among the six criteria pollutants. $PM_{2.5}$ consists of microscopic particles that can penetrate deep into the lungs and cause health problems. People with heart or lung diseases are the most likely to be affected by $PM_{2.5}$ pollution. Even healthy people may experience temporary symptoms from exposure to elevated levels of $PM_{2.5}$ pollution. Health problems caused by $PM_{2.5}$ pollution include increase of respiratory disease symptoms, decrease of lung function, developments of chronic bronchitis, nonfatal heart attacks, and premature death in people with heart or lung disease. Fine particles can be carried over long distances by wind and then deposited on ground or water through dry or wet deposition. The wet deposition is often acidic, due to the presence of acidic compounds such as sulfuric acid. Fine particles containing sulfuric acid contribute to rain acidity, or "acid rain". The effects of acid rain include changing the nutrient balance in water and soil, damaging sensitive forests and farm crops, and affecting the diversity of ecosystems. $PM_{2.5}$ pollution is also the major cause of visibility reduction that frequently occurs in many areas in the United States (EPA, 2005a).

The NAAQS for Total Suspended Particles (TSP) were first established in 1971. In 1987, EPA revised the particulate matter (PM) standards and replaced TSP with particles smaller than 10 micrometers (PM_{10}). Ten years later, after a lengthy review, EPA revised the PM standards, setting separate standards for particles smaller than 2.5 micrometers ($PM_{2.5}$). Until recently, applicable NAAQS for $PM_{2.5}$ were 15.0 $\mu g/m^3$ for the annual mean concentration and 65.0 $\mu g/m^3$ for the 24-hr mean. In December 2006 the U.S. EPA lowered the 24-hr NAAQS for PM2.5 to 35 $\mu g/m^3$, based on review of recent health studies. The primary standards and secondary standards are the same. For designation of NAAQS attainment for $PM_{2.5}$, the 3-year average of the 98th percentile of 24-hour concentrations at each population-oriented monitor within an area must not exceed 35 $\mu g/m^3$ and the 3-year average of the annual arithmetic mean $PM_{2.5}$

concentrations from single or multiple community-oriented monitors must not exceed 15.0 μg/m^3 (EPA, 2005b).

Among the six criteria pollutants, $PM_{2.5}$ was responsible for the second most violations of NAAQS in the U.S, next to ground-level ozone. In 2005, 88 million people lived in 208 counties designated as $PM_{2.5}$ nonattainment areas. Most of the $PM_{2.5}$ nonattainment areas were also ozone nonattainment areas (EPA, 2005c). In Kentucky, Covington, Ashland, and Louisville are currently nonattainment areas for both ozone and $PM_{2.5}$. To provide the air pollutant nonattainment areas a better chance to meet the NAAQS and issue advanced alerts on potentially unhealthful air quality days for sensitive people (Groups that are sensitive to ozone or $PM_{2.5}$ include children and adults who are active outdoors, and people with respiratory disease), NLR $PM_{2.5}$ forecast models were developed for selected metro areas in Kentucky. Also NLR-Fuzzy system $PM_{2.5}$ forecast models were developed to seek better model performance.

In summary, the objectives of this study were:

1. Evaluating the updated NLR ozone forecast models for selected metro areas in Kentucky.
2. Developing combined NLR-Fuzzy system ozone forecast models, and comparing the model performance with that of the NLR models.
3. Developing NLR $PM_{2.5}$ forecast models for selected metro areas in Kentucky, and evaluating the model performance based upon the fitted data.
4. Developing NLR-Fuzzy system $PM_{2.5}$ forecast models, and comparing the performance of the NLR-Fuzzy system models and NLR models.

CHAPTER II

BACKGROUND

2.1 Ground level Ozone

Ozone is an odorless, colorless gas. The ozone molecule is composed of three atoms of oxygen (O_3). Ground level ozone refers to the ozone in the earth's lower atmosphere. It is not released directly into air, but formed by complex chemical reactions between volatile organic compounds (VOCs) and nitrogen oxides (NO_x) in the atmosphere. Sunlight plays an important role in ozone formation. The ultraviolet radiation splits a single oxygen atom off the NO_2; via a process called photolysis.

$$NO_2 + h\nu \rightarrow NO + O \tag{2.1}$$

The oxygen atom combines with oxygen (O_2) in the air to form ozone (O_3).

$$O_2 + O \rightarrow O_3 \tag{2.2}$$

In unpolluted air, the nitric oxide formed in reaction (2.1) combines with O_3 to reform NO_2 and O_2, thus completing the NO-NO_2-O_3 photolytic cycle.

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{2.3}$$

In polluted air, complex system of reactions involving volatile organic compounds (VOCs) and nitric oxide (NO) competes with ozone in oxidizing NO (as in Equation (2.4)), thus leading to a buildup of O_3 from reaction (2.2).

$$\mathrm{VOC} + \mathrm{NO} + \mathrm{sunlight} \rightarrow NO_2 + CH_3COOONO_2\,(\mathrm{PAN}) + \mathrm{aerosol} + \mathrm{other\ products} \tag{2.4}$$

Ground level ozone concentration depends on not only the concentrations of NOx and VOC concentrations, but also the VOC/NOx ratio. At high VOC/NOx ratios, ozone formation is controlled by the amount of NOx available, and reaction (2.4) is the main route to regenerate NO_2 from NO. Under this "NOx-limited" situation, decreasing NOx reduces ozone, while decreasing VOC has little or no effect on ozone. But at low VOC/NOx, ozone formation is limited by the amount of VOC available for reaction (2.4), and reaction (2.3) becomes the main route to regenerate NO_2 from NO. In addition, at low VOC/NOx, NO_2 competes with VOC to react with OH radicals, slowing the rate of reaction (2.4). Under this "VOC-limited" condition, reducing VOC reduces ozone, but reducing NOx increases ozone (Schwartz, 2006).

Many urban areas tend to have high levels of ground-level O_3. Some rural areas are also subject to elevated O_3 levels because wind carries ozone and its precursor pollutants hundreds of miles from their original sources. Sunlight and hot weather cause ground-level ozone to form in harmful concentrations in the air. As a result, O_3 is known as a summertime air pollutant. Previous studies have shown that meteorological factors significantly affect ground level O_3 concentrations (Revlett, 1978; Wolff and Lioy, 1978). Daily maximum temperature, relative humidity, and wind speed are among the factors that are strongly related to O_3 concentrations. The relationship between these vital parameters and O_3 concentrations can be represented by nonlinear functions, such as higher order polynomial or exponential function. The other predictor meteorological parameters, including cloud cover, precipitations, atmospheric transmittance, etc, approximately linearly correlated with O_3 concentrations (Lin, 2004).

VOC emissions are produced from numerous combustion sources such as automobiles and power plants, and also industrial processes such as paint coating, printing, and organic chemical producing. NO_x emissions are produced primarily when fossil fuels are burned in power plants, motor vehicles, and industrial boilers. The emission of NO_x from power plants has played a significant role in the phenomena of long-range transport of O_3 and its precursors. By the early 1990s, a new technology for controlling NOx emission, called selective catalytic reduction (SCR), had been demonstrated to be highly effective in reducing NO_x emissions from large sources (Forzatti, 2001). Following the wide availability of SCR technology in the United States, EPA in 1997 issued the NO_x state implementation plan (SIP) call, aimed to mitigate significant transport of NO_x. Under this regulation, many states, particularly in the Midwest, were required to reduce NO_x emissions from point sources dramatically by 2003, through their SIPs. Motor vehicles are the other main source of NO_x and VOC emissions. Nationally, the Clean Air Act Amendments of 1990 mandated increasingly stringent rules to reduce car and truck tailpipe emissions.

NO_x emissions indeed have been reduced due to application of new technology and implementations of pollutant reduction strategies over the past several years. The nationwide NO_x emissions total decreased 15% from 1983 to 2002, and decreased 12% from 1993 to 2002 (EPA, 2005d). In Kentucky, NO_x emissions from point sources totaled 1,624,600 tons in 2002, down 33% from 1998 totals. In the Midwestern states bordering Kentucky (MO, IL, IN, OH, WV, VA, and TN), NOx emissions from regulated utilities in that area totaled 224,490 tons in 2002, down 22% from 1998 totals (Cobourn and Lin, 2004).

Reductions in NOx and VOC emissions have resulted in improvements in ozone air quality across the country. Since 1980, 1-hr average O_3 concentrations have been reduced by 29% and 8-hour average O_3 concentrations have been reduced by 21%. Between 1990 and 2003 there was a 16% improvement for 1-hr average O_3 concentrations and 9% reduction for 8-hr average O_3 concentrations (EPA, 2005e). In Kentucky, Cobourn and Lin (2004) have studied the 8-hr ground level ozone trend in recent years for seven metro areas: Ashland, Bowling Green, Covington, Lexington, Louisville, Owensboro, and Paducah. In the period 2000-2004, there has been a downward trend in upper-end O_3 concentrations (represented by annual 8th maximum) for each of the metro areas. On average, the O_3 concentrations declined by about 10 ppb (13%) during that period. The O_3 concentrations are strongly affected by meteorology. To discern an ozone trend associated with the huge effort of air pollution controls, meteorologically adjusted O_3 concentrations were estimated using the nonlinear regression models developed for these areas. It was demonstrated in the study that the meteorologically adjusted O_3 concentrations also have a downward trend for each of the metro areas, with greater certainties than the unadjusted O_3 concentrations. On average, the meteorologically adjusted O_3 declined by about 10% from 2000 to 2004 (Cobourn and Lin, 2004).

2.2 Fine Particulate Matter ($PM_{2.5}$)

Particulate matter is the term for atmospheric particles, including dust, dirt, soot, smoke, and liquid droplets. $PM_{2.5}$ is the fine particulate matter consisting of particles with diameter 2.5 μm or smaller. $PM_{2.5}$ consists of primary particles, which are directly emitted into the air, and secondary particles, which form in the atmosphere from chemical reactions involving common gaseous pollutants. Primary $PM_{2.5}$ particulate results largely from combustion of fossil or biomass fuels and selected industrial processes. The sources of $PM_{2.5}$ include, but are not limited to, gasoline and diesel combustion, wood stoves and fireplaces, land clearing, wild land prescribed burning, and wild fires. Secondary $PM_{2.5}$ forms through homogeneous and heterogeneous chemical reactions that convert some common gaseous pollutants into very small particles. $PM_{2.5}$ precursors include sulfur oxide compounds (SO_x), nitrogen oxide compounds (NO_x), and VOCs. The observed $PM_{2.5}$ concentrations are dominated by sulfur and nitrogen species in most locations. However, there can also be significant contributions from secondary organic aerosol in some locations (EPA, 1999). Generally, the major components of $PM_{2.5}$ are carbon, sulfate and nitrate compounds, and crustal materials. The chemical makeup of particles varies across the United States. In the air quality region of the industrial midwest, including Kentucky, sulfate compounds are the dominant component of $PM_{2.5}$ (~ 45%) and carbonaceous mass is the second (~ 35%).

Implementation of EPA's air improvement programs has helped reduce $PM_{2.5}$ and its precursors. The Acid Rain Program aimed to reduce releases of SO_2, NO_x, and other pollutants that contributed to the formation of acid rain from coal-fired power plants. For the SO_2 portion of the Acid Rain Program, the first phase began in 1995 and targeted the largest and highest emitting power plants. The second phase (started in 2000) set tighter restrictions on smaller plants. This program will reduce annual SO_2 emissions by 10 million tons (almost half the 1980 level) between 1980 and 2010 (EPA, 2005f). The NOx portion of the Acid Rain Program has helped reduce annual 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, EPA, 2005g). National ozone-reduction programs designed to reduce emissions of NOx and VOCs, such as the NOx SIP call and mandatory rules for reducing car and truck tailpipe emissions required by Clean Air Act Amendments, also have helped reduce carbon and nitrates, both of which are components of $PM_{2.5}$.

As a result of the implementation of EPA's air improvement programs, the national SO_2, NO_x, and VOC emissions decreased 9%, 9%, and 12%, respectively, from 1999 to 2003. The nationwide annual average $PM_{2.5}$ concentrations declined 10% over the same period. Reductions of estimated direct emissions resulted in a 5% decrease of $PM_{2.5}$ concentrations. Decreases of the $PM_{2.5}$ precursor emissions yielded additional reductions. In the industrial Midwest region which includes Kentucky, the annual average $PM_{2.5}$ decreased 9% from 1999 to 2003 (EPA, 2004).

The US EPA has promoted a more stringent 8-hr standard for ground-level O_3 in 1997. In December 2005 EPA proposed revisions to the NAAQS for particle pollution. EPA revised the air quality standards for particle pollution in 2006. The 2006 standards tighten the 24-hour fine particle standard from the current level of 65 micrograms per cubic meter ($\mu g/m^3$) to 35 $\mu g/m^3$, and retain the current annual fine particle standard at 15 $\mu g/m^3$. The new PM2.5 standards became effective since December 17, 2006. It can be expected that the O_3 and $PM_{2.5}$ problems will continue to draw public attention in the near future.

2.3 Characteristics of the Seven Metro Areas

The seven Kentucky metro areas selected for the study were Ashland, Bowling Green, Covington, Lexington, Louisville, Owensboro, and Paducah. All of these metro areas are multi-county areas (Figure 2.1). The Ashland, Covington, and Louisville are currently designated nonattainment areas for both ground level O_3 and $PM_{2.5}$ (EPA, 2005c). Except for the Paducah metro area (MA), the other six are official U.S. Census Bureau metropolitan Statistical areas (MSA). The Covington, Louisville, and Owensboro MSAs include counties from bordering states. These metro areas vary substantially in population. The Covington-Cincinnati MSA (Kentucky-Ohio) and Louisville MSA (Kentucky-Indiana) are the largest and second largest in population, each at more than 1 million. The Ashland-Huntington MSA (Kentucky-West Virginia-Ohio) and Lexington MSA are medium-sized at several hundred thousand. The Owensboro MSA is contiguous with the Evansville IN-KY MSA. The combined area is a medium-sized area of about 450,000 in population. The Bowling Green MSA and Paducah MA are small, at about 100,000 each (Cobourn and Lin, 2004).

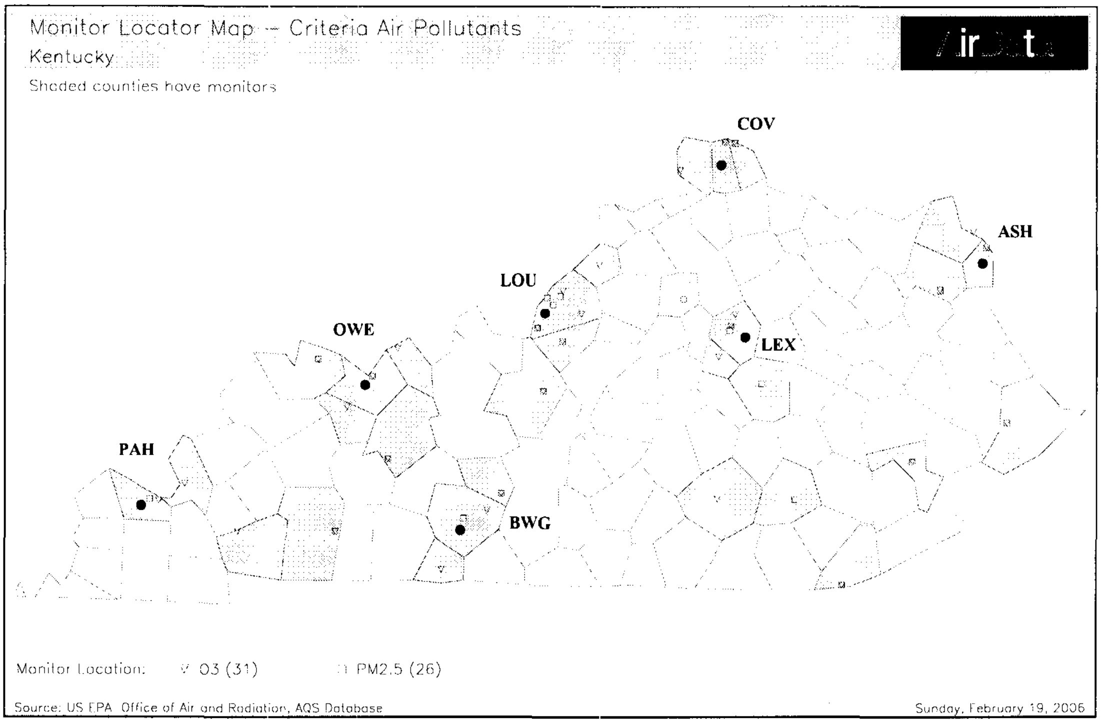

Monitor Locator Map -- Criteria Air Pollutants
Kentucky
Shaded counties have monitors
AirData
COV
ASH
LOU
OWE
LEX
PAH
BWG
Monitor Location: O3 (31) PM2.5 (26)
Source: US EPA Office of Air and Radiation, AQS Database
Sunday, February 19, 2006

Figure 2.1 The seven metro areas with O_3 and $PM_{2.5}$ monitors located in Kentucky

There are a variety of $PM_{2.5}$, VOC, and NO_x emission sources in each of these areas. The emission sources include area sources, such as household and mobile sources, and point sources, including power plants, chemical plants, and various manufacturing facilities. The annual emissions and emission densities vary for each of the multi-county areas (Table 2.1). The estimated annual $PM_{2.5}$ emission total varies from 2987 tons for Bowling Green to 26,677 tons for Owensboro. The average $PM_{2.5}$ emission density varies from 2.8 t/yr/mi^2 for Bowling Green to 17.5 t/yr/mi^2 for Covington. Jefferson County, which contains the city of Louisville, has by far the highest PM2.5 emission density, at 28.5 t/yr/mi^2. Hamilton County (Covington MSA), Floyd County (Louisville MSA), Hancock County and Warrick County (Owensboro MSA) also have high $PM_{2.5}$ emission densities, over 20 t/yr/mi^2.

Table 2.1 Annual Pollutant Emissions in Selected Metro Areas in KY (MA=metro area, ED=emission density)

MA	Area (mi^2)	Popul.	PM2.5 ED ($t/yr/mi^2$)	VOC ED ($t/yr/mi^2$)	NOx ED ($t/yr/mi^2$)	SO_2 ED ($t/yr/mi^2$)	$[PM2.5]_{avg}$ (ug/m3)	$[O3]_{avg}$ (ppb)
ASH	1654	271,882	4.3	14.7	23.0	12.4	18.3	57
BWG	1084	125,980	2.8	10.8	10.1	3.5	15.7	56
COV	967	1,156,111	17.5	70.5	106.3	108.8	18.6	60
LEX	743	346,700	5.6	33.8	27.7	6.1	17.0	52
LOU	1780	1,028,243	10.9	46.0	65.5	63.0	17.7	59
OWE	2363	405,386	11.3	15.6	52.8	99.1	16.5	56
PAH	1123	111,863	3.9	7.7	30.9	30.8	14.3	56

(Source: AirData - Reports and Maps, EPA, 2005h)

VOC and NO_x compounds are the important precursors for both ground level ozone and $PM_{2.5}$. The estimated annual VOC emission total for each area varies from 8658 tons (Paducah) to 81,842 tons (Louisville). The estimated total annual NO_x emissions vary from 10,977 tons (Bowling Green) to 124,812 tons (Owensboro). The average VOC emission density varies from 7.7 $t/yr/mi^2$ (Paducah) to 70.0 $t/yr/mi^2$ (Covington). The NO_x emission density varies from 10.1 $t/yr/mi^2$ (Bowling Green) to 106.3 $t/yr/mi^2$ (Covington). Jefferson County has the highest VOC and NO_x emission densities (152.1 $t/yr/mi^2$ and 232.7 $t/yr/mi^2$). The Hamilton County is the next highest at 118.9 $t/yr/mi^2$ for VOC and 171.2 $t/yr/mi^2$ for NO_x. Several counties, such as McClean County (Owensboro MSA) and Edmonson County (Bowling Green MSA), have the VOC and NO_x emission densities as low as about 2~3 $t/yr/mi^2$.

The VOC and NO_x emission densities vary quite substantially between counties and also between the several metro areas. In contrast, the variation in pollutant O_3 is relatively mild. The five year (2000-2004) average summertime daily domain peak ozone concentration ranged from 52 ppb (Lexington) to 60 ppb (Covington). Nevertheless, the fact is that the greatest problems with ground level O_3 tend to exist in areas where NOx and VOC emission densities are high, for example the Covington and Louisville metro areas (Table 2.1). Clearly these precursor pollutants lead to elevated ozone concentrations. Therefore, reductions in VOC and NO_x emissions will lead to improved O_3 air quality.

Sulfate compounds comprise the biggest component of $PM_{2.5}$ in these Kentucky metro areas. The estimated annual SO_2 emission total for each area varies from 4508 t (Lexington) to 234,212 t (Owensboro). The average SO_2 emission density varies from 6.1 t/yr/mi^2 (Lexington) to 108.8 t/yr/mi^2 (Covington). Floyd County has the highest SO_2 emission densities (322.9 t/yr/mi^2). The Hancock County is the next highest at 301.7 t/yr/mi^2.

The five year (1999-2003) average summertime daily averaged $PM_{2.5}$ concentration ranges from 14.3μg/m^3 (Paducah) to 18.6μg/m^3 (Covington). The high $PM_{2.5}$ concentrations tend to occur in the areas with high emission densities of VOC, NO_x, and SO_2 (for example, Covington and Louisville). The only exception is Ashland: in this metro area, the emission densities of VOC, NO_x, and SO_2 were low but the average $PM_{2.5}$ concentration was relatively high (18.3μg/m^3).

CHAPTER III

LITERATURE REVIEW

Many kinds of ozone and $PM_{2.5}$ forecast models have been described in the literature, including regression models, neural network (NN) models, photochemical transport models, fuzzy system models, and others. Some of the forecast models are being used to provide regional air quality guidance, such as the Community Multiscale Air Quality (CMAQ) model, or making local air quality forecasts, such as the Houston Generalized Additive Model (GAM) and the University of Louisville NLR ozone forecast models. Most of these models have been evaluated on calibration data sets (model fits). Some of them also have been validated on independent data sets using either observed meteorological data (model hindcasts) or forecast meteorological data (model forecasts) as model inputs.

3.1 Ozone Forecast Models

3.1.1 Regression Models

Both linear regression and nonlinear regression models have been employed for ozone forecasting. The general purpose of a linear regression is to learn about the linear relationship between several independent variables and a dependent variable. With the ordinary least squares method, the linear regression procedure will compute model predictions so that the squared deviations between modeled and observed concentrations are minimized.

Robeson and Steyn (1990) tested a bivariate temperature and persistence linear regression model with the ozone data in Fraser Valley of British Columbia, Canada. It was concluded that the bivariate regression model was superior to an autoregressive integrated moving average (ARIMA) model and persistence model. Chaloulakou et al. (2003) proposed a multiple regression model to forecast the next day's hourly maximum O_3 concentration in Athens, Greece. The set of the input variables consisted of eight meteorological parameters and three persistence variables, which were the hourly maximum O_3 concentrations of the previous three days. Testing this linear regression model on four separate test datasets, the MAE ranged from 19.4% to 33.0% of the corresponding average O_3 concentrations. Prybutok et al. (2000) built a simple linear regression model for forecasting the daily peak O_3 concentration in Houston. The final model used four meteorological and O_3 precursor parameters, including O_3 concentration at 9:00 a.m., maximum daily temperature, average nitrogen dioxide concentration between 6:00 a.m. and 9:00 a.m. 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 error statistics of the linear regression model were favorably compared with those of the neural network model built on the same database.

Comrie (1997) developed basic multiple linear regression models and neural network 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. 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. The average observed ozone concentrations ranged from around 40 to 66 ppb for the eight cities. Testing the multiple linear regression models with the quasi-independent subset of 250 observed data, the model hindcasts exhibited MAEs from 8.24 to 13.46 ppb and R^2 from 0.15 to 0.59. The NMAE, which is the ratio of MAE to average ozone concentration, ranged from 16% to 27%. The NN models exhibited MAEs from 7.01 to 12.41 ppb and R^2 from 0.27 to 0.70. The NMAE ranged from 15% to 24%.

Nonlinear regression models are superior to simple linear regression models because they capture the nonlinear relationships between ozone and meteorological parameters. Bloomfield et al. (1996) described a nonlinear regression model to explain the effects of meteorology on O_3 in the Chicago area. The model input variables consisted of a seasonal term, a linear annual trend term, and twelve meteorological variables, including maximum temperature, wind speed, wind direction, relative humidity, specific humidity, dew point temperature, total cloud cover, opaque cloud cover, ceiling height, barometric pressure, and visibility. The observed ozone and meteorological data in 1981-1991 were divided into subsets for model development and validation. The model predictions of the model fits were within ±5 ppb about half the time, and within ±16 ppb about 95% of the time. The RMSE was 8.2 ppb. The model was cross-validated using the indenpendent data subset. The overall RMSE of the cross-validated prediction errors was 8.3 ppb. Bloomfield et al. demonstrated that the meteorological data accounted for at least 50% of the variance of the ozone concentration.

Hubbard and Cobourn (1998) developed a regression model to forecast next-day maximum 1-hr ground-level O_3 concentrations in Louisville, KY. The regression model included a nonlinear temperature term plus additional linear terms. The linear terms included atmospheric transmittance (or length of day), minimum temperature, cloud cover, rainfall, nighttime calms, and day of week. Cobourn and Hubbard (1999) improved this NLR model by using an interactive nonlinear regression term based on maximum temperature, wind speed, and relative humidity. This model was called the hybrid model. It consisted of a standard model fitted to complete database and a Hi-Lo model fitted to the days on which the ozone concentrations were in the upper and lower 10% of the ozone distribution. The model also included a trajectory parameter. For the testing period 1993-1997 (580 days), when keeping the same input variables but exclude the trajectory term, the 1999 model had 5.4% lower MAE (8.99 vs. 9.50 ppb), as compared to 1998 model. Inclusion of the trajectory parameter provided an additional decrease of MAE by 6.0% (8.45 vs. 8.99 ppb).

Cobourn et al. (2000) compared the performance of the NLR model with a three-layer perception neural network (NN) for predicting daily maximum 1-hr O_3 concentration in Louisville, by using data sets for 1998 and 1999 O_3 seasons. The model predictions were compared for the forecast mode and hindcast mode. For the hindcast mode, the NLR model exhibited MAEs of 11.0 and 11.2 ppb for the 1998 and 1999 ozone seasons. The corresponding NMAEs were 15% and 16%. The NN model exhibited MAEs of 12.9 ppb for both of the two years. The NMAEs were 18% and 17% for the 1998 and 1999 O_3 seasons respectively. The model forecasts of the NLR and NN model were comparable. The MAEs for both were around 13.0 for 1998 and 11.8 ppb for 1999.

The corresponding NMAEs were 18% and 15%. During the 1998 and 1999 ozone seasons, the forecast detection rate of 120 ppb threshold exceedances was 42% for each model on 12 exceedences. The hindcast detection rate was 92% for the NLR model and 75% for the NN model.

3.1.2 Neural Network (NN) Models

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 designed to emulate human information processing capabilities such as knowledge processing, speech, prediction, classifications, and control. These models are capable of representing highly non-linear relationships, such as the relationship between ozone concentration and meteorological parameters.

Chaloulakou et al. (2003) developed a NN forecast model to forecast the next day's maximum 1-hr ozone concentration in four locations within the Athens basin, Greece. The NN architecture was the feed forward, multi-layer perceptron topology, consisting of an input layer, a hidden layer, and an output layer. There were 11 nodes in the input layer (eight meteorological and three persistence variables) and one node in the output layer. The input meteorological variables included morning wind speed, nocturnal wind speed, solar radiation, relative humidity, temperature at 850 h Pa, temperature change at 850 h Pa from the previous day, surface temperature range and wind direction index. The three persistence variables referred to the maximum 1-hr O_3 concentrations of the previous three days. The model fit of the NN models for the four locations had

NMAEs ranging from 19.4% to 33.0%. When using the European information threshold for O_3 (180 μg/m^3, or 91.8 ppb), the model detection rate ranged from 0.67 to 0.76 and the false alarm rate ranged from 0.48 to 0.56. In this study, the authors concluded that the NN model provided a considerable improvement in the forecasting of O_3 concentrations over a linear regression model that used the same input parameters.

An innovative neural network model was developed by Wang et al (2003), City University of Hong Kong. This model combines the adaptive radial basis function network with statistical characteristics of ozone to predict the 1-hr daily maximum ozone concentrations in selected specific areas. The input parameters for the model included wind speed, maximum temperature, solar radiation, and the daily maximum O_3 and NOx concentrations from the previous day. In predicting ozone concentrations of the area Tsuen Wan, Kwai Chung, and Kwun Tong for the entire year 2000, the MAEs of model hindcasts were 23.2, 26.3, and 24.8 μg/m^3 (11.8, 13.4, and 12.7 ppb) respectively.

Spellman (1999) described a neural network (NN) model used for predicting the ozone concentrations of five selected cities of the United Kingdom. This two layer NN model had only three predictive parameters, including maximum temperature, hours of sunshine, and previous day's ozone concentration. The model was evaluated on an independent data subset consisting of observed meteorological and air quality data. For the model hindcasts the MAEs ranged from 4.74 ppb to 9.30 ppb, the NMAEs ranged from 12% to 24%, and R^2 ranged from 0.28 to 0.60 for five selected cities.

Balaguer et al. (2002) used a finite impulse response NN model to make 1-day advance predictions of 8-hr average ozone concentrations in eastern Spain. The input variables were observed 24 h lagged observed values of air quality and meteorological inputs, including ambient concentrations of O_3, NO, and NO_2, temperature, solar irradiance, atmospheric pressure, wind speed, and wind direction. The models were evaluated using data from the 1996 to 1999 ozone seasons (July to September). The statistics of the model fits for three sampling sites ranged from 6.39 to 8.8 ppb for MAE and from 0.73 to 0.79 for R^2.

A NN model developed by Elkamel et al. (2001) was applied to predict ozone concentrations around a heavily industrialized area in Kuwait. The meteorological and air quality inputs to the neural network were wind speed, wind direction, relative humidity, daily maximum temperature, solar intensity and the concentration of the pollutants methane, carbon monoxide, carbon dioxide, nitrogen oxide, nitrogen dioxide, sulfur dioxide, non-methane hydrocarbons, and dust. This model was trained using data collected during a period of 60 days. The data fed to the neural network were divided into a training set and a testing set. The NMAEs for the training set and testing set were 11.1% and 12.5% respectively.

3.1.3 Photochemical transport models

Photochemical transport models are numerical models that simulate the transport and chemical transformation of pollutants in the atmosphere. There are two types of photochemical transport models: Eulerian models and Lagrangian models. Photochemical air quality models play an important role in scientific investigation of pollutant processes in the atmosphere and in development of policies to manage air quality. Early in 1973, Reynolds et al. created an Eulerian model, the Urban Airshed Model, for evaluating episodes and air pollution control measures (Russell and Dennis, 2000). After that, many more photochemical transport models were applied 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.

The U.S. EPA developed the Community Multi-scale Air Quality (CMAQ) modeling system, an advanced air quality modeling system designed to approach air quality as a whole by including state-of-the-science capabilities for modeling multiple air quality issues (EPA, 2006). With the model's ability to handle a large range of spatial scales, CMAQ can be used for urban and regional scale model simulations. The CMAQ modeling system simulates various chemical and physical processes that are thought to be important for understanding atmospheric trace gas transformations and distributions. The components of CMAQ system included a meteorology-chemistry interface processor, a photolysis rate processor, an initial conditions processor, a boundary conditions processor, and the CMAQ chemical-transport model. One of the functions of CMAQ system is to provide guidance for O_3 forecasting to the environmental management agencies all over the country. The CMAQ system also has considerably ability to simulate the ambient O_3 concentrations. Eder and Yu (2006) evaluated the performance of CMAQ (Version 4.4, released in 2004) covering the contiguous United States against monitoring data from four nationwide networks. For the simulations of the Peak 1-hr and 8-hr O_3 concentrations during 2001 ozone season, the correlation coefficient (R) was 0.68 and 0.69; the NMAE was 18.3% and 19.6%, respectively.

Flemming et al. (2001) have employed the regional Eulerian model with 3 chemistry mechanisms (REM3), which was a photochemical transport model, operationally to forecast ozone since 1997 at the Freie University, Berlin. The vertical resolution of the model was based on three dynamically changing layers. The chemical mechanism CBM4 was used in the model. The model has been used for making 1, 2, and 3 day advance ozone forecasts with data over Germany from 1997 to 1999, the correlation coefficient (R) spread from 0.9 to 0.77. The disadvantage of this model was that it tended to underestimate the low ozone concentrations.

Another example of an Eulerian model is a photochemical grid model that was used to analyze two ozone episodes in autumn (2000) and winter (2001) seasons in 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. This model has been applied for simulating the variation of O_3 levels for selected episodes. For the autumn episodes, R^2 was 0.865, the coefficient of variation (S) was 0.27, and the index of agreement (d_1) was 0.80. For the winter episodes, values of R^2, S, and d_1 were 0.886, 0.3, and 0.83 respectively.

Wotawa et al. (1998) developed a Lagrangian photochemical box model for providing ozone forecasts for Vienna, Austria. This model consisted of up to 8 vertical and up to 5 horizontal boxes. It simulated emission, chemical reactions, horizontal diffusion, vertical diffusion, dry deposition, wet deposition and synoptic scale vertical exchange. Model input data included a trajectory term, which was calculated using forecast meteorological data. The model predictions for the 1995 O_3 season underestimated O_3 concentrations on most days. The overall median bias was -12.3 ppb. The correlation coefficient (R) was greater than 0.6 for most of the study cases.

3.1.4 Fuzzy system models

Fuzzy modeling is a tool aimed at using the information observed from a complex phenomenon to derive a quantitative model. A fuzzy system is a nonlinear mapping between inputs and outputs. A fuzzification block converts the crisp inputs to fuzzy sets. An inference mechanism uses the fuzzy rules in the rule-base to produce fuzzy conclusions, and a defuzzification block converts these fuzzy conclusions into crisp outputs (Passino and Yurkovich, 1998).

Ryoke et al. (2000) developed a fuzzy O_3 forecast model to describe the relationships between O_3 precursor emissions and daily maximum O_3 concentrations. The estimated emissions of NOx, VOC, CO, and SO_2 were used as model inputs. Meteorological parameters used in this model included mixing height, cloud cover, temperature data, solar radiation, and atmospheric stability. This fuzzy model was used to represent numerous results of the European Monitoring and Evaluation Program (EMEP) model. In this study, the fuzzy model provided better predictions of ozone than a linear regression model with the same input variables. The R^2 between predictions by the fuzzy model and the EMEP ozone model was 0.811, greater than the R^2 between the linear regression model and the EMEP model (0.6708).

Jorquera et al. (1998) compared the performance of fuzzy system, NN, and time serious forecast models in Santiago, Chile. These models were applied to predictions of maximum 1-hr O_3 concentrations. The input variables for these models were daily maximum temperature (for the day of forecast), previous-day daily maximum temperature, and previous-day O_3 concentrations. The fuzzy system model was a Takagi-

Sugeno fuzzy model identified with a fuzzy c-mean algorithm. Testing these models with the observed data from the 1994 ozone season, the errors of the fuzzy model hindcasts were comparable with that of NN model and time series model. For example, the average RSME for four ozone monitors were 23.9 ppb, 23.5 ppb, and 23.5 ppb for the fuzzy model, NN model, and time series model, respectively. Jorquera et al. presented the comparison of various ozone forecast models. However, using only three model input variables may not be adequate for developing an accurate O_3 forecast model. Also these models may not practical for next-day forecasts because "previous-day" maximum parameter O_3 concentrations probably would not be available at time of forecast.

Heo, et al. (2004) applied a fuzzy expert system and neural network combined model to short-term O_3 forecasting in Seoul, Korea. The input variables included meteorological data (temperature, relative humidity, wind direction, wind speed, solar radiation), O_3, and NO_2 concentrations for the previous day. Also the O_3 concentrations, concentrations of O_3 precursors (SO_2, CO, NO_2), and meteorological data were collected from 8:00 to 14:00 and were used as input variables to forecast the maximum ozone concentrations at 15:00. The model was examined by making predictions for O_3 concentrations at seven consecutive hours in a day (8:00 to 14:00), during the 1999 ozone season. The NMAEs of the model hindcasts ranged from 7.4% to 20.4%. This was not particularly impressive considering the short horizon of the predictions. Also, this scheme probably would not work well for next-day forecasts.

3.1.5 Other ozone forecast models

Generalized additive models (GAMs) represent a method of fitting a smooth, nonlinear functional relationship between two variables in a scatter plot of data points. The GAM is resulted by adapting the functional forms in a linear combination fitted by regression techniques. GAMs are effective when the relationship between the variables is expected to be a complex form, not easily fitted by standard linear or non-linear models. GAMs do not involve strong assumptions about the relationship that is implicit in standard parametric regression. Davis et al. (1999) created a GAM ozone forecast model in Houston, TX. In the examined years 1988 and 1991 in Houston, the RMSE of the model hindcasts ranged from 13.2 to 16.3 ppb and the R^2 ranged from 0.66 to 0.73 for the individual stations. For daily domain peak concentrations, the RMSEs were from 18.5 to 22.0 ppb and R^2 were from 0.61 to 0.68.

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).

The Simplified Ozone Modeling System (SOMS) 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. For the three years simulations using SOMS, the model bias was 1.9 ppb, MAE was 12.5 ppb and the R^2 was 0.81. Due to the availability of concentrations NOx and VOC, this model is only adequate for long-term ozone forecasting.

3.2 $PM_{2.5}$ Forecast Models

The problem of fine particulate matter ($PM_{2.5}$) has caused increasing concern after the U.S. EPA established annual and 24-hour NAAQS for $PM_{2.5}$ in 1997. The $PM_{2.5}$ forecast models described in literature are much less abundant than the O_3 forecast models. Nevertheless, they do include regression models, NN models, and photochemical transport models. $PM_{2.5}$ Concentrations, like those of O_3, are related to meteorological conditions. However, $PM_{2.5}$ has a much longer atmospheric life time than O_3, so recent meteorology has less correlation with $PM_{2.5}$. This makes the statistical approach, using meteorological and air quality data, tend to be somewhat less accurate for $PM_{2.5}$ as compared to O_3 modelling.

Ordieres, et al. (2005) proposed a linear regression model comparison with their NN models used for predicting daily average $PM_{2.5}$ concentration on the US-Mexico border in Texas and Chihuahua (Mexico). This simple regression model used 7 input variables, including average temperature, relative humidity, wind speed, wind bearing, wind direction during the first 8 hour of the day, and the average and maximum levels of $PM_{2.5}$ during the first 8 hours of the day. On the 2002 test data set, the R^2 of the model hindcasts was 0.40.

The NN PM2.5 forecast models developed by Ordieres, et al. (2005) included three types of neural network models, which were multilayer perceptron (MLP), square multilayer perceptron (SMLP), and radial basis function (RBF) NN models. These NN models used the same input variables as the simple regression model. On the 2002 test data set, the neural network models had better performance than the linear regression model. The RBF network model, which had the best performance among the three neural network models, had 0.46 for R^2.

Perez, et al. (2000) constructed a NN $PM_{2.5}$ forecast model to make predictions of hourly average $PM_{2.5}$ concentrations in the downtown area of Santiago, Chile. Three forecast models, neural networks, linear regression, and persistence model, were developed to predict $PM_{2.5}$ concentrations at any hour of the day, using the 24 hourly average concentrations measured on the previous day as the input variables. The NMAE of the predictions for the 1994-1995 ozone season (May 1 to September 30) ranged from 30% to 60%. In this study, the authors demonstrated that the $PM_{2.5}$ formation strongly depends on weather conditions. The $PM_{2.5}$ concentrations negatively correlated with wind velocity and relative humidity.

Forsyth county environmental affairs department in Winston-Salem, North Carolina, has been running a "phenomenological" model to forecast year-round $PM_{2.5}$ concentrations for the Triad area of North Carolina (FCEAD, 2005). The input variables for this model include a group of meteorological factors, such as wind speed, wind direction, cloud cover, night length, etc. This model has been used for providing $PM_{2.5}$ forecasts since 1998. The model accuracy was evaluated using the value of air quality index (AQI). For the model forecasts during 2005, the MAE was 10.6 AQI, the bias was -0.8 AQI. The correlation value (R) between forecasts and observations was 0.70. The DR and FAR of the model forecasts in 2005 were 0.56 and 0.55 respectively.

CHAPTER IV

METHODOLOGY

4.1 Air Quality data

4.1.1 Ground Level Ozone Air quality data

The air quality data was quality assured data from the US EPA AQS system, provided by several local agencies, such as the Louisville Air Pollution Control District and the Kentucky Division of Air Quality (KDAQ). The data files consisted of hourly readings for the 8-hr average ozone concentrations from each of the monitors in those metro areas. For each of the metro areas, there are several ozone monitors located within the area. For example, the ozone data used for the Louisville ozone forecast model came from seven ozone monitors, three of which are located in Jefferson County. The other four monitors are situated in the counties that are part of the Louisville MSA, viz. Bullitt Co. and Oldham Co. in Kentucky and Floyd Co. and Clark Co. in Indiana. There were at least three monitors in each metro area (Table 4.1).

Table 4.1 Ozone Monitors in the Seven Metro Aareas

Metro Areas	Symbol	No. of monitors	Location County	Location type	Monitor type	Operation period
Ashland	ASH	3	Greenup, KY	Suburban	SLAMS	1981-present
			Boyd, KY	Suburban	SLAMS	1998-present
			Carter, KY	Rural	Special purpose	1983-present
Bowling Green	BWG	3	Simpson, KY	Rural	Special purpose	1991-present
			Edmonson, KY	Rural	Special purpose	1997-present
			Warren, KY	Rural	Non-EPA	1999-present
Covington	COV	10	Boone, KY	Rural	SLAMS	1975-present
			Kenton, KY	Suburban	SLAMS	1975-present
			Campbell, KY	Suburban	SLAMS	1998-present
			Clermont, OH	Suburban	SLAMS	2001-present
			Butler, OH -1	Suburban	SLAMS	1973-present
			Butler, OH -2	Suburban	SLAMS	1982-present
			Warren, OH	Suburban	SLAMS	2003-present
			Hamilton, OH -1	Rural	SLAMS	1978-present
			Hamilton, OH -2	Suburban	NAMS, SLAMS	1969-present
			Hamilton, OH -3	Urban	SLAMS	1999-present
Lexington	LEX	4	Scott, KY	Rural	Special purpose	1993-2004
			Fayette, KY	Rural	SLAMS	1978-present
			Fayette, KY	Suburban	SLAMS	1979-present
			Jessamine, KY	Suburban	SLAMS	1991-present
Louisville	LOU	7	Jefferson, KY -1	Suburban	SLAMS	1973-present
			Jefferson, KY -2	Suburban	SLAMS	1973-present
			Jefferson, KY -3	Suburban	SLAMS	1992-present
			Bullitt, KY	Urban	SLAMS	1992-present
			Oldham, KY	Rural	SLAMS	1981-present
			Floyd, IN	Suburban	SLAMS	1976-present
			Clark, IN	Suburban	SLAMS	1980-present
Owensboro	OWE	5	Hancock, KY	Rural	SLAMS	1980-present
			McClean, KY	Rural	Special purpose	1991-present
			Henderson, KY -1	Rural	Special purpose	1992-present
			Henderson, KY -2	Suburban	SLAMS	1982-2002
			Daviess, KY	Suburban	SLAMS	1970-present
Paducah	PAH	3	McCracken, KY	Suburban	SLAMS	1980-present
			Livingston, KY	Rural	SLAMS	1981-present
			Graves, KY	Rural	Special purpose	1990-present

Data source: US EPA, AirData - Reports and Maps: Monitor Locator. (Reference: EPA, 2005h)

The forecast models are used to determine whether to announce air quality warnings for local citizens. Therefore, the relevant parameter to forecast is the metro area peak, or domain peak of 8-hr average O_3 concentration for the day. All domain peak concentrations observed during the ozone season for five years were entered into a database. In cases where there were missing monitor data, the daily domain peak was determined from the available monitors, provided that the fraction of available monitors was greater than or equal to 60%.

4.1.2 $PM_{2.5}$ air quality in the seven metro areas

The USEPA Air Quality System (AQS) provides a web link for downloading archived data for the six criteria pollutants, including $PM_{2.5}$. The archived files for $PM_{2.5}$ data consist of the daily average $PM_{2.5}$ concentration readings from each of the $PM_{2.5}$ monitors all over the US. Some monitors sample every day, and some monitors sample every three days. The $PM_{2.5}$ air quality data for the seven metro areas in Kentucky were extracted from the archived files with a data processing program. The $PM_{2.5}$ monitors in Kentucky are distributed over the state, and include urban, suburban, and rural areas. There is only one monitor in the Paducah metro area. The other six metro areas contain several $PM_{2.5}$ monitors (Table 4.2). The appropriate daily $PM_{2.5}$ concentration used in this study is the daily maximum 24-hr recorded concentration of all monitors used in each metro area (the "domain peak"). The AQS monitor data are quality assured data. However, for purposes of assembling a valid set of domain peak values, for these $PM_{2.5}$ databases, we required that at least 50% of the monitors for each area were in operation for each day. Otherwise, the day was excluded from the database.

Table 4.2 $PM_{2.5}$ Monitors in the Seven Metro Areas

Metro Areas	Symbol	No.	County	Monitor ID	Type	Location
Ashland	ASH	4	Boyd Co, KY	21-019-0017-88101	SLAMS	Suburban
			Carter Co, KY	21-043-0500-88101	SLAMS	Rural
			Lawrence Co, OH	39-087-0100-88101	SLAMS	Suburban
			Cabell Co, WV	54-011-0006-88101	SLAMS	Suburban
Bowling Green	BWG	2	Edmonson Co, KY	21-061-0501-88101	Special	Rural
			Warren Co, KY	21-227-0007-88101	SLAMS	Urban
Covington	COV	9	Kenton Co, KY	21-117-0007-88101	SLAMS	Suburban
			Campbell Co, KY	21-037-0003-88101	SLAMS	Suburban
			Hamilton Co, OH	39-061-0006-88101	SLAMS	Suburban
			Hamilton Co, OH	39-061-0040-88101	SLAMS	Urban
			Hamilton Co, OH	39-061-0041-88101	SLAMS	Urban
			Hamilton Co, OH	39-061-0042-88101	SLAMS	Urban
			Hamilton Co, OH	39-061-0043-88101	SLAMS	Suburban
			Hamilton Co, OH	39-061-7001-88101	SLAMS	Suburban
			Hamilton Co, OH	39-061-8001-88101	SLAMS	Suburban
Lexington	LEX	3	Fayette Co, KY	21-067-0012-88101	SLAMS	Suburban
			Fayette Co, KY	21-067-0014-88101	SLAMS	Urban
			Madison Co, KY	21-151-0003-88101	SLAMS	Urban
Louisville	LOU	6	Jefferson Co, KY	21-111-0043-88101	Other	Suburban
			Jefferson Co, KY	21-111-0044-88101	SLAMS	Suburban
			Jefferson Co, KY	21-111-0048-88101	SLAMS	Urban
			Jefferson Co, KY	21-111-0051-88101	Other	Suburban
			Floyd Co, IN	18-043-1004-88101	SLAMS	Suburban
			Clark Co, IN	18-019-0006-88101	SLAMS	Urban
Owensboro	OWE	6	Henderson Co, KY	21-101-0014-88101	SLAMS	Rural
			Daviess Co, KY	21-059-0014-88101	SLAMS	Suburban
			Vanderburgh Co, IN	18-163-0006-88101	SLAMS	Urban
			Vanderburgh Co, IN	18-163-0012-88101	SLAMS	Urban
			Vanderburgh Co, IN	18-163-0016-88101	SLAMS	Urban
			Spencer Co, IN	18-147-0009-88101	SLAMS	Suburban
Paducah	PAH	1	McCracken Co, KY	21-145-1004-88101	SLAMS	Urban

Data source: US EPA, AirData - Reports and Maps: Monitor Locator. (Reference: EPA, 2005h)

4.2 Original observed meteorological data

The surface observed meteorological data were observations from local weather stations (Table 4.3). Most of these were obtained from Edited Local Climatological Data Reports issued by National Climatic Data Center (NCDC, 1999-2004a). Edited meteorological data were not available for Bowling Green, so unedited meteorological data downloaded from the NCDC website (NCDC, 1999-2004b) were used in this case. Data from the Agriculture Weather Center at the University of Kentucky were substituted for missing data (UKAWC, 2005a-b).

Table 4.3 Weather Stations for the Seven Metro Areas

Metro Areas	Station Name	Location	Latitude	Longitude	Elevation
Ashland	Huntington Tri-state Airport	Wayne Co, WV	38° 22'	82° 33'	824 ft
Bowling Green	Bowling Green Warren Country Airport	Warren Co, KY	36° 59'	86° 26'	528 ft
Covington	Cincinnati Northern KY Airport	Boone Co, KY	39° 03'	84° 40'	869 ft
Lexington	Lexington Bluegrass Airport	Fayeette Co, KY	38° 02'	84° 36'	980 ft
Louisville	Louisville Standiford field	Jefferson Co, KY	38° 11'	85° 44'	488 ft
Owensboro	Evansville Regional Airport	Vanderburgh Co, IN	38° 02'	87° 32'	418 ft
Paducah	Paducah Barkley Regional Airport	McCracken Co, KY	37° 03'	88° 46'	413 ft

The meteorological parameters consisted of daily maximum and minimum temperature, hourly surface observations of sky description, precipitation, temperature, dew point, relative humidity, wind direction, and wind speed. The data were examined for errors with a data scanning program. All days were examined for:

- Missing data. Due to the extensive use of data averaging in this analysis, it was essential to identify missing data to ensure that averages were computed correctly with valid data.
- Consistency between temperature, dew point and relative humidity.
- Whether the temperature, dew point, relative humidity, and wind speed were within reasonable ranges.

The anomalous data and missing data were compared to the data from Agriculture Weather Center at the University of Kentucky for verification or substitution. The days that still had incorrect data were excluded from the databases.

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 known source areas of ozone or its precursors. The NOAA Air Resources Laboratory provided a three-dimensional wind trajectory web calculator, using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model, for calculating forward and backward trajectories at various levels for continental US locations (NOAA, 2005). We used the 36-hr air mass trajectories at 750 meters level for Louisville and Lexington for studying the relationship between air mass transportation and the peak ozone concentration (Figure 4.1). The duration of the trajectories was based on the characteristic transport time for ozone and ozone precursors (1-2 days). The trajectory height was chosen to be roughly half of the average summertime mixing height, so that the trajectories would be a mean representation of the transport, which varies in speed and direction throughout the mixed layer. Consideration was given to using trajectories at multiple heights, such as 500 m, 750m, and 1500m (Figure 4.1), but rejected in favor of the simpler approach, in consideration of the time element involved in the ozone forecasting process (Cobourn and Hubbard, 1999).

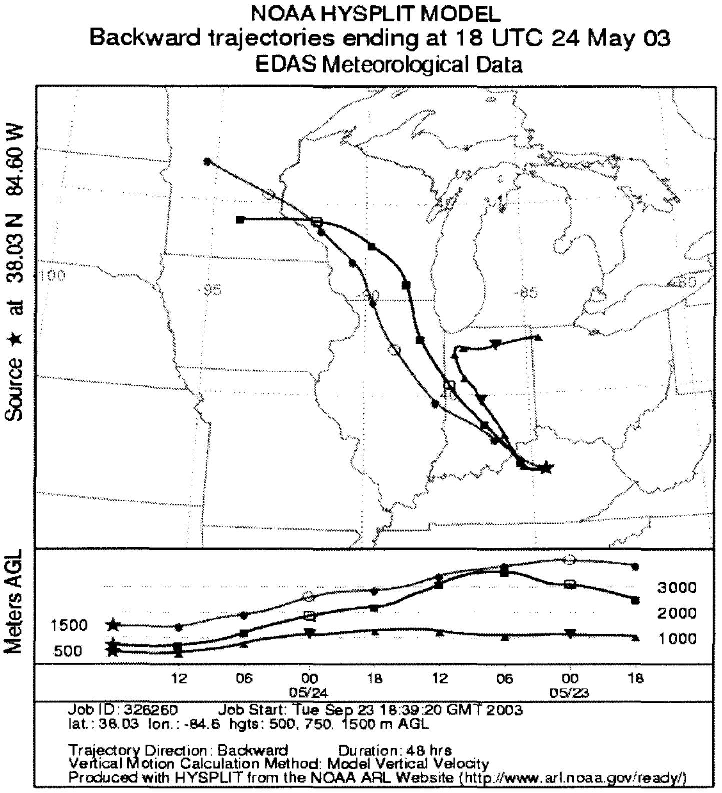

NOAA HYSPLIT MODEL
Backward trajectories ending at 18 UTC 24 May 03
EDAS Meteorological Data
Source ★ at 38.03 N 84.60 W
100
95
85
Meters AGL
1500
500
3000
2000
1000
12
06
00
18
12
06
00
18
05/24
05/23
Job ID: 326260 Job Start: Tue Sep 23 18:39:20 GMT 2003
lat.: 38.03 lon.: -84.6 hgts: 500, 750, 1500 m AGL
Trajectory Direction: Backward Duration: 48 hrs
Vertical Motion Calculation Method: Model Vertical Velocity
Produced with HYSPLIT from the NOAA ARL Website (http://www.arl.noaa.gov/ready/)

Figure 4.1 48-hr backward hindcast trajectories for Lexington

4.3 Model Performance Metrics

Several statistical indices were used to evaluate the performance of the O_3 and $PM_{2.5}$ forecasting models, including square of correlation coefficient (R^2), statistical significance test value (t-value), mean error (Bias), mean absolute error (MAE), normalized mean absolute error (NMAE), root mean square error (RMSE), detection rate (DR), false alarm rate (FAR), success rate (SR), and critical success index (CSI).

4.3.1 Square of Correlation Coefficient (R^2)

Pearson's Correlation Coefficient (R), sometimes called Product Moment Correlation, reflects the degree of linear relationship between two variables (such as observed pollutant concentrations and model predictions). This index varies from 0 to 1, with 0 indicating no relationship and 1.0 indicating perfect relationship. The Square of Pearson's Correlation Coefficient (R^2) represents the percent of the variance in the dependent variable (observed concentrations) explained by the independent variable (model predictions). The R^2 is defined by (Lomax, 2001),

$$R^2 = \frac{\sum_{i=1}^{N}(p_i - \bar{o}_i)^2}{\sum_{i=1}^{N}(o_i - \bar{o}_i)^2} \tag{4.1}$$

where p_i refers to the model predictions for O_3 or PM2.5 concentration and o_i represents the observed values, and $\bar{o}_i$ is the average of the observed pollutant concentrations.

4.3.2 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

$$t = \frac{b_k}{s(b_k)} \tag{4.2}$$

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

$$s(b_k) = \frac{s_{res}}{\sqrt{(n-1)\cdot s_k^2 \cdot (1-R_k^2)}} \tag{4.3}$$

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

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 significant and there is correlation between the corresponding independent variables and the dependent variable. For the multiple linear regressions in this study, at the 0.05 level of significance with ~750 degrees of freedom, the critical t-value is about 2.0 (Lomax, 2001).

4.3.3 Mean error (Bias)

The mean error (Bias) is the arithmetic mean of the errors, given by

$$Bias = \frac{\sum_{i=1}^{n}(p_i - o_i)}{n} \tag{4.4}$$

The Bias for the fitted data in a regression model should be zero. The Bias for forecasted data using a regression model usually is close to zero.

4.3.4 Mean absolute error (MAE)

The mean absolute error (MAE) is the average absolute value of the prediction errors. The MAE is given by

$$MAE = \frac{\sum_{i=1}^{n} | p_i - o_i |}{n} \tag{4.5}$$

4.3.5 Normalized mean absolute error (NMAE)

The average O_3 and PM2.5 concentrations vary from one location to another. Using the same forecast models, the model predictions for the areas with high pollutant concentration levels usually have higher MAEs than the model predictions for the areas with low concentration levels. Therefore, the MAE is not useful for comparing model results from different locations. In this case, the normalized mean absolute error (NMAE) may better evaluate the forecast models for different areas. The NMAE is the ratio of MAE to average pollutant concentrations. Expressed as a percentage, it is given by

$$NMAE = \frac{MAE}{\bar{o}_i} \times 100\% \tag{4.6}$$

4.3.6 Root mean square error (RMSE)

The root mean square error (RMSE) is the square root of the mean of the squares of all the forecast errors, given by

$$RMSE = \sqrt{\frac{\sum (p_i - o_i)^2}{n}} \tag{4.7}$$

Compared to the MAE, the RMSE is more sensitive to outliers.

4.3.7 Detection rate (DR)

The detection rate (DR) is the fraction of the observed exceedences detected by the model. It is calculated by

$$DR = \frac{DE}{EX} \tag{4.8}$$

where DE is the number of detected exceedences, and EX is the number of total observed exceedences during a specified period (e.g. ozone season). The model "detects" an exceedence based on the model prediction exceeding pre-determined alarm threshold. The alarm threshold 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 thresholds for the O_3 and $PM_{2.5}$ forecast models are slightly lower (~5 ppb) than the nominal NAAQS exceedence threshold, so that accurate forecasts (e.g. within a few ppb) just below the exceedence level do not result in "missed exceedences".

4.3.8 False alarm rate (FAR)

A false alarm is an alarm for which the observed concentration did not exceed the alarm threshold. The false alarm rate is defined as the ratio of false alarms (FA) to total alarms (AL) predicted by the model.

$$FAR = \frac{FA}{AL} \tag{4.9}$$

Increasing the alarm threshold tends to reduce both the alarms and false alarms, but the false alarm rate tends to increase. Lowering the alarm threshold would tend to improve the DR and FAR, but increase the number of alarms and false alarms. In prescribing an alarm threshold, public officials must strike a balance between achieving a high DR, without creating too many alarms and false alarms that could erode public confidence in the air quality forecasts.

4.3.9 Critical successes index (CSI)

The critical successes index is the ratio of valid alarms ($AL - FA$) to critical events. Critical events include alarms and undetected exceedences. The CSI can be calculated by

$$CSI = \frac{AL - FA}{AL + EX - DE} \tag{4.10}$$

The CSI is a measure of the model effectivities at critical forecasts, i.e., when the predictions are above alarm levels or concentrations are above exceedence levels.

4.3.10 Success rate (SR)

The success rate is the ratio of the successful predictions (i.e., both the observed exceedences and non-exceedences that were successfully predicted by the model) to overall observed days (OD). The successful predictions can be obtained by subtracting the false alarms and undetected exceedences from the overall observed days. The SR is given by

$$SR = \frac{OD - FA - (EX - DE)}{OD} \tag{4.11}$$

Since most days during the ozone season are uneventful, the SR is usually a high percentage.

CHAPTER V

OZONE FORECAST MODELS

In 1997, the first hybrid nonlinear regression (NLR) ozone forecast model for Louisville was developed at University of Louisville. Due to the successful implementation of the NLR model for Louisville, more NLR ozone forecast models were developed for other selected metro areas in Kentucky. In 2005 there were seven NLR models providing ozone predictions for the metro areas Ashland, Owensboro, Bowling Green, Covington, Lexington, Louisville, and Paducah. Since the local ozone pollution is affected in part by the local and regional emissions, climate, and land use, a separate fitting process was employed for each metro-area model. Each of the metro area databases consisted of ozone air quality data and meteorological data from consecutive ozone seasons. The databases were updated each year by adding the air quality and meteorological data from the most recent ozone season, and removing the data of the earliest ozone season.

Development of the NLR models has led to the compiling of several sets of complete databases of ozone concentrations and related meteorological parameters for the seven metro areas. Using 1999-2003 database set, fuzzy system ozone forecast models were developed for the seven metro areas for application to the 2004 forecast season. Moreover, combined NLR-fuzzy models were synthesized with the objective of attaining a set of more accurate ozone forecast models.

5.1 Ozone Prediction Parameter Development

The training data used for developing the ozone forecast models for the seven metropolitan areas consisted of ozone air quality data and a group of candidate ozone predictors. The ozone predictors were derived from observed meteorological data and other factors that play important roles in ozone concentrations, such as pollutant transport and ozone air quality trend. The database for each of the metro areas was built to manage the data and generate the parameters used in the ozone forecast models. Each database contained data from five ozone seasons (May to September). The maximum number of days in each database was thus 765 days. Due to some missing ozone data or meteorological data, the total number of days in the databases ranged from 750 to 760. Ozone prediction parameters were derived from meteorological data, air mass trajectories, and other deterministic factors. There were four classes of parameters used in this study, including observed meteorological parameters, derived meteorological parameters, deterministic parameters, and other parameters.

5.1.1 Observed meteorological parameters

The observed meteorological parameters consisted of daily maximum and minimum temperature (Tmax, Tmin), average temperature (Tavg), dew point temperature (Dewpt), cloud cover (CC), relative humidity (RH), mid-day wind speed (WS), rain (Rain), and thunder storm occurrences (TS). The parameters Tmax and Tmin were instantaneous values of extrema from the datasets, not extremes of the hourly data. To reduce the random fluctuations of the hourly observed data, the parameter Tavg, Dewpt, CC, RH, and WS were averaged over several hours. The span of averaging interval was chosen to be 10 A.M. to 4 P.M. During that time of day the ozone levels were highest.

Daily maximum temperature is the most powerful meteorological variable for forecasting ground-level ozone. This is because the rates of photochemical reactions are highly sensitive to temperature, and high air temperatures are usually associated with strong solar radiation, sunny skies, stagnant circulation, and subsiding upper air. The scatter plot of O_3 against the parameter Tmax was used for studying the response of O_3 on Tmax (For example, Figure 5.1). It was found that a second-order polynomial provides a good fit to the data. The values of coefficient of determination (R^2) for the regressions were about 0.4 (Lin, 2004). In a two-way linear regression, the parameter Dewpt correlated positively to ground-level O_3. The dew point provides a lower limit value on the minimum temperature due to the latent heat of condensation of water. So the parameters Dewpt and Tmin strong correlated to each other in the multiple linear regression. In this study, the parameters Tmin and Tavg were not used as direct ozone predictors. The Tavg and Dewpt were used to estimate the relative humidity; the Tavg and Tmin were used to calculate "special relative humidity" (RHx) parameter.

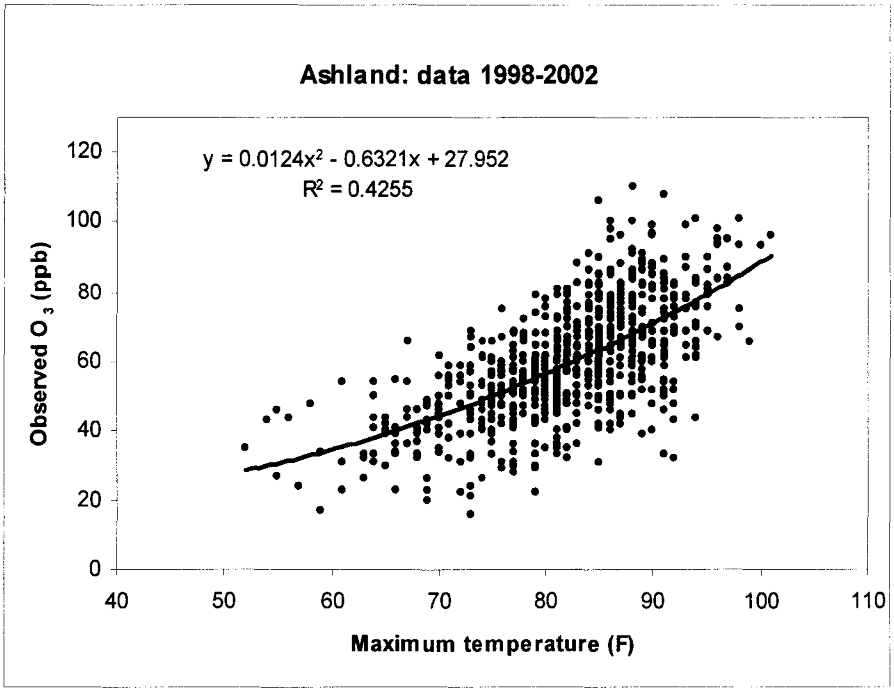


Figure 5.1 Second-order polynomial regression of O_3 concentrations vs. Tmax (Data from Ashland 1998-2002 ozone season)

Cloud cover is negatively correlated with ground-level ozone concentrations since clouds reduce solar radiation intensity available to drive the ozone forming photochemical reactions. Sky condition was reported using encoded descriptions such as “Clear”, “Overcast”, etc. In order to develop the parameter CC for numerical analysis, it was necessary to convert the encoded descriptions to the equivalent tenths of cloud cover (Table 5.1).

Table 5.1 Tenth of Cloud Cover Converted by Sky Condition Descriptions

Sky Condition Description	Symbol	CC value (tenth)
Clear	CLR	0.5
Few cloud	FEW	1.5
Scatter	SCT	3
Broken	BKN	7
Overcast	OVC	9.5

Surface wind speed affects the dilution and mixing of air pollutants. High wind speed reduces pollutant concentrations. Some air pollutant concentration models theoretically explained this phenomenon. For example, in both the Gaussian plume diffusion model

$$C = \frac{q}{2\pi \cdot U \cdot \sigma_y \sigma_z} \exp(-\frac{y^2}{2\sigma_y^2}) \exp(\frac{-(z-h)^2}{2\sigma_z^2}) \tag{5.1}$$

and the fixed-box model (De Nevers, 1995)

$$C = b + \frac{q \cdot l}{U \cdot h} \tag{5.2}$$

local pollutant concentration, C, is inversely proportional to wind speed, U. Ozone concentration is negatively correlated with wind speed (Figure 5.2). With databases of Ashland, Bowling Green, Owensboro, and Paducah, a variety of functional forms were fitted to the wind speed data. The best model was found to be a nonlinear exponential function

$$[\mathrm{O_3}] = \beta \exp(\theta \cdot \mathrm{Mdwind}) \tag{5.3}$$

where Mdwind refers to mid-day wind speed, β and θ are coefficients. On average, the determination coefficients R^2 for the above function in a two-way regression was about 0.03 (Lin, 2004). This form was used as part of the nonlinear term in the model.

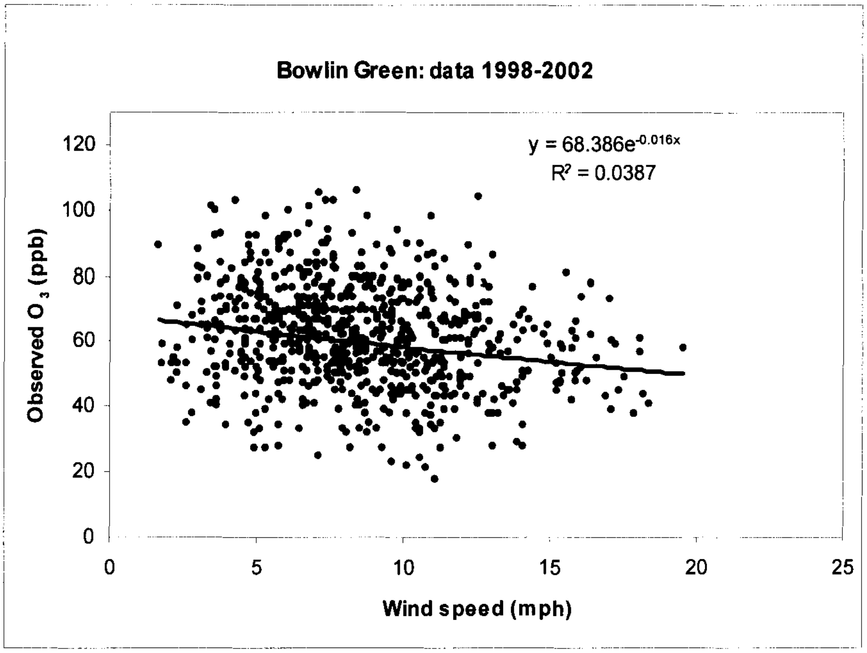


Figure 5.2 Nonlinear regression of O_3 concentrations vs. surface wind speed (Data from Bowling Green)

The precipitation parameter "Rain" used in this study referred to the daily precipitation recorded in the NCDC data files. The rainfall reduces ozone levels by directly scavenging O_3 and O_3 precursors. On the other hand, rainfall is associated with increased cloud cover and increased convective activity. All these factors would reduce ozone levels. Thunderstorm occurrences were selected as a parameter based on two reasons: first, thunderstorms are usually accompanied with heavy rain and unstable atmospheric conditions; second, a forecasted thunderstorm probability can be obtained 24 hours in advance. The parameter TS was defined as following: If the thunderstorm occurred in time period 6 A.M. to 5 P.M, the parameter was assigned value "1", otherwise it was assigned value "0".

5.1.2 Derived meteorological parameters

Derived meteorological parameters consisted of maximum and minimum temperature departure (Tmx_dep and Tmn_dep), and special relative humidity (Rhx). The National Climatic Data Center provided the "normal" climatological conditions for metro areas all over the nation. The normal daily maximum and minimum temperatures are the 30-year average values computed from the data recorded during the period 1971-2000 (NCDC, 2003). The parameter Tmx_dep and Tmn_dep were obtained by calculating the differences between the daily maximum or minimum temperatures and the corresponding normal values. In our previous nonlinear regression ozone forecast models, either Tmx_dep or Tmn_dep were significantly correlated with ozone concentrations (Lin, 2004).

The relative humidity is calculated from the partial pressure of water (function of dew point temperature) and the saturated vapor pressure of water at the temperature (function of temperature). The calculated relative humidity correlated better than the National Weather Service measured relative humidity.

$$RH = \frac{Psat(Dewpt)}{Psat(Tavg)} \tag{5.4}$$

In this study, we define a "special relative humidity" parameter, as follow,

$$RHx = \frac{Psat(T\,min)}{Psat(T\,max)} \tag{5.5}$$

where Psat( ) is a polynomial function used for calculating the saturation vapor pressure of water. In the previous ozone forecast models, the parameter RHx was used in the nonlinear term because of higher statistical significance in the nonlinear regression. The parameter RH sometimes was used as a linear term if it was statistically significant in the multiple linear regression.

5.1.3 Deterministic parameters

The deterministic parameters consisted of normal maximum temperature (Tmx_nrm), normal minimum temperature (Tmn_nrm), length of day (LOD), clear sky atmospheric transmittance (Xmitt), holiday (Hol), Saturday (Sat), and Friday (Fri).

The clear sky atmospheric transmittance was derived from the average intensity of solar radiation at noon received at ground level, which drives the photochemical ozone formation process. The parameters Xmitt and LOD both are calculated with day of year, zenith angle, and altitude angle of the metro areas location. Since the LOD and Xmitt strongly correlate with each other, the one that performed better in the regressions was selected as the independent parameter in the forecast model.

Saturday, Friday, and holiday were considered as parameters because on the weekend or holiday, the reduction of traffic and manufacturing could reduce the emission of ozone's precursors, VOC and NOx. Each of the three parameters, Sat, Fri, and Hol, has been statistically significant in some previous forecast modes.

5.1.4 Other ozone prediction parameters

This category includes the statistical parameter local ozone trend (Trend) and two transport parameters, air mass trajectory (Traj) and 48-hr ozone transport (OZ48).

The parameter Trend was developed based on the fact that the regional average ozone concentrations have declined over the past five years. Cobourn and Lin (2004) studied the ozone trend for six Kentucky metro areas. On average, the meteorologically adjusted ozone concentrations declined about 6 ppb during the five recent O_3 seasons 1998-2002. Emissions of NOx and VOC compounds also declined during this period (EPA 2005).

The trajectory parameter "Traj" reflects the influence of the transport of ozone and its precursors. To determine the value of this parameter, the forecaster would compare the 750m air trajectory to a map that displays an envelope developed by Coboum and Hubbard (1999) encompassing most of the large NOx emission sources (Figure 5.3). The parameter would be assigned as a value of 1.0 if the originating backward trajectory was fully inside the envelope, a value of zero if originating outside the envelope, and a value of 0.5 if originating inside the envelope, but lying in proximity to the envelope boundary.

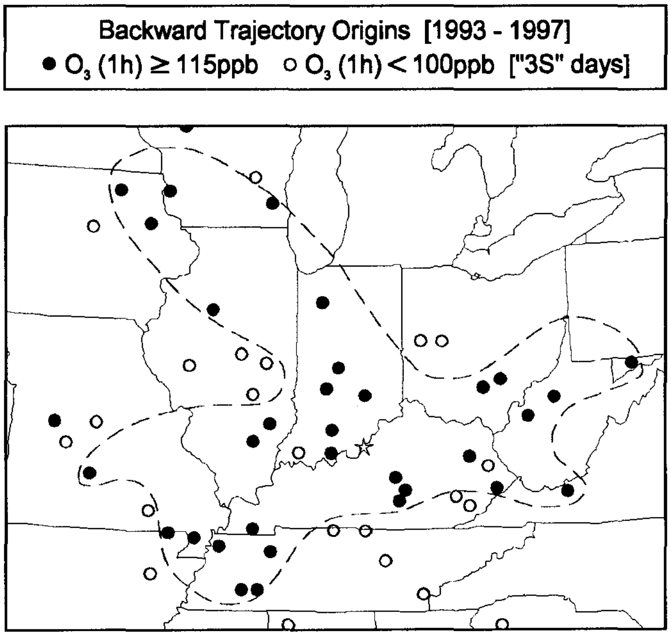

Backward Trajectory Origins [1993 - 1997]
● O3 (1h) ≥ 115ppb ○ O3 (1h) < 100ppb ["3S" days]

Figure 5.3 Origins of the 36-hr backward trajectories at 750m elevation on high-ozone days during the Period 1993-1997 (Louisville pattern. Cobourn, 1999)

The parameter OZ48 accounts for the influence of the long distance transported ozone on the local ozone concentrations. The value of parameter OZ48 was determined by comparing the figures of the 48-hour backward air trajectory to the national ozone concentration contour map available from the EPA AIRNOW web site (Figure 5.4). If the 48-hour backward trajectory came from the areas with high ozone concentrations, the parameter OZ48 was assigned as a value of 1.0. Otherwise it was assigned as a value of zero.

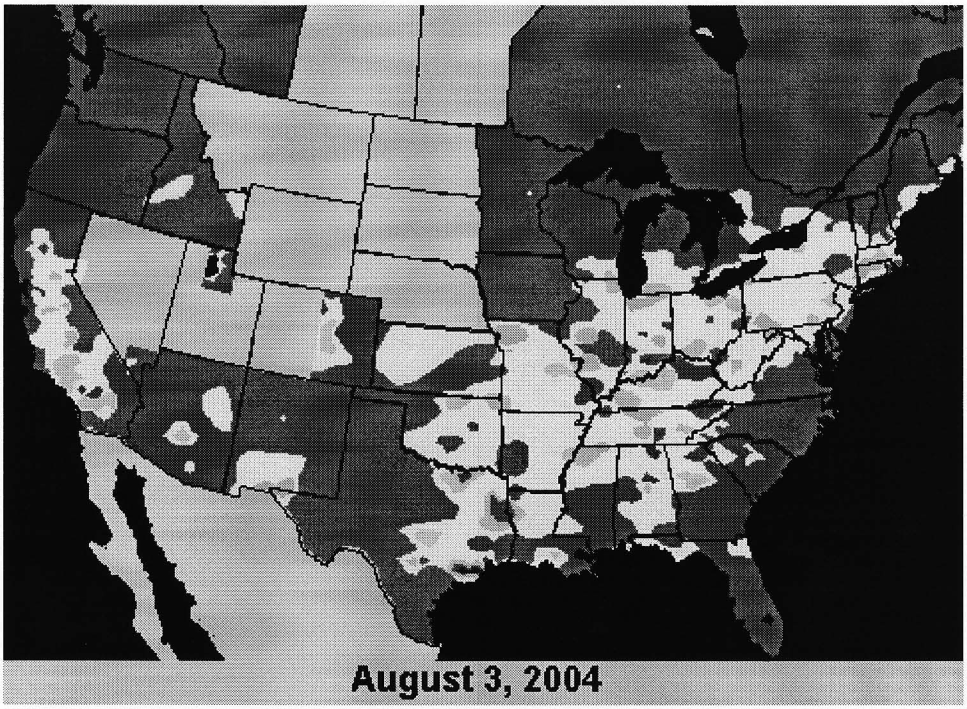


Figure 5.4 Sample AIRNOW Map (1-hr ozone concentrations) (EPA, 2005 j)

The four classes of ozone prediction parameters are summarized in Table 5.2. These candidate predictor variables were included in databases for possible use in the ozone forecast models.

Table 5.2 Candidate Variables for Ozone Forecast Models

Class	Parameter	Description	Units	Timing
Observed Meteorological Parameter	Tmax	maximum temperature	ºF	Daily instantaneous
	Tmin	minimum temperature	ºF	Daily instantaneous
	Tavg	average temperature	ºF	10 am to 4 pm avg.
	Dewpt	dew point temperature	ºF	10 am to 4 pm avg.
	CC	cloud cover		10 am to 4 pm avg.
	WS	wind speed	mph	10 am to 4 pm avg.
	Rain	rain	inch	daily
	TS	thunder storm		5 am to 5 pm
Derived Meteorological Parameter	Tmx_dep	max. temperature departure	ºF	daily
	Tmn_dep	min. temperature departure	ºF	daily
	RH	special relative humidity 2		10 am to 4 pm avg.
	RHx	special relative humidity 4		10 am to 4 pm avg.
Deterministic Parameter	Tmx_nrm	normal max. temperature	ºF	daily
	Tmn_nrm	normal min. temperature	ºF	daily
	LOD	length of day	hours	daily
	Xmitt	atmospheric transmittance		noon
	Trend	ozone trend	ppb/year	annual
	Traj	air mass trajectory		36-hr backward
	OZ48	ozone transportation		48-hr backward
	Hol	holiday		
	Sat	Saturday		
	Fri	Friday		

5.2 Evaluation of 2005 NLR ozone forecast models

The NLR model has been shown to be more accurate than linear models and at least as accurate as neural network models for O_3 forecasting (Cobourn et al, 2000). It also has been shown that the previous NLR forecast models for six metro areas in Kentucky (Ashland, Owensboro, Bowling Green, Lexington, Louisville, and Paducah) performed well on providing O_3 forecasts for those metro areas (Cobourn and Hubbard, 1999; Lin, 2005). In the year of 2003, one more NLR ozone forecast model for the Cincinnatti-Covington metro statistical area (MSA) was developed and implemented. The 2005 ozone forecast models for the seven metro areas refer to the models fitted to the databases of 2000-2004 ozone seasons. These models were designed to predict the daily maximum ozone concentrations in the summer ozone season of 2005. In this study, the 2005 ozone forecast models for all the seven metro areas will be evaluated on both calibration data sets and independent data sets.

5.2.1 The NLR ozone forecast models for seven metro areas

The operational O_3 forecast models were hybrid models. A standard model and a "Hi-Lo" model were developed separately for each metro area model. The standard model was fitted to the complete database, 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 and lower 10% of the ozone distribution. This technique reduced the influence of the middle days on the outcome of the regression coefficients. Both the high and low ozone concentration days were included in the Hi-Lo model to preserve the numerical range of the predictors and predictand. The Hi-Lo model was invoked when high ozone-prone meteorological conditions, called 3S criteria, were forecast. The 3S criteria account for three important weather characteristics associated with high ozone level: sunny, sultry, and stagnant, as follows:

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

This switching strategy increased the detection rate and increased the explained variance, without significantly changing the bias or MAE error for the model (Cobourn and Hubbard, 1999).

The nonlinear regression model was characterized by a nonlinear term (Nonlin). The nonlinear term was actually a separate prediction model, obtained in a nonlinear regression fitting process. The nonlinear term had the same form in each of the seven NLR forecast models, as follows:

$$Nonlin = (a_1 + (a_2 T\,max + a_3 T\,max^2)\,exp(a_4 WS))\,exp(a_5 Rhx) \tag{5.6}$$

This function accounted for the nonlinear behavior of ozone with regard to maximum temperature, wind speed, and relative humidity. As explained the special relative humidity term Rhx was used in the nonlinear term because the statistical significance in the nonlinear regression was slightly better than that of the relative humidity, RH. A separate regression process was employed for each metro-area model, yielding a unique set of constants a_i for each metro area (Table 5.3).

Table 5.3 Coefficients for the Nonlinear Term for the 2005 NLR Ozone Models

Coef.	ASH	BWG	CVG	LEX	LOU	OWE	PAH
Standard model							
a_1	86.87	88.47	85.91	80.18	90.97	90.03	80.33
a_2	-3.04	-2.02	-3.47	-1.53	-3.63	-2.97	-1.95
a_3	0.041	0.028	0.049	0.024	0.049	0.039	0.027
a_4	-0.112	-0.052	-0.076	-0.041	-0.076	-0.079	-0.049
a_5	-0.010	-0.012	-0.010	-0.012	-0.011	-0.011	-0.010
Hi-Lo model							
a_1	70.28	88.47	-2.41	-238.87	60.34	47.92	80.33
a_2	-2.95	-2.02	-0.14	6.77	-2.92	-1.39	-1.95
a_3	0.046	0.028	0.027	-0.023	0.051	0.032	0.027
a_4	-0.057	-0.052	-0.021	-0.003	-0.050	-0.015	-0.049
a_5	-0.013	-0.012	-0.013	-0.019	-0.014	-0.017	-0.010

The nonlinear term was used as a predictor variable in the multiple linear regression. The final equation for the forecast model, with predicted O_3 as the dependent variable, consisted of an intercept and a group of explanatory terms. The general form of the final equation used for the seven NLR models is,

$$\begin{aligned} O_3 &= b_0 + b_1 Nonlin(T\max, WS, RHx) + b_2 Xmitt + b_3 Trend + b_4 RH + b_5 Tmn\_dep \\ &+ b_6 WS + b_7 CC + b_8 Dewpt + b_9 Traj + b_{10} OZ48 + b_{11} TS \end{aligned} \tag{5.7}$$

Whether a parameter was used as an independent variable depended on statistical significance in the multiple linear regression. As described in Chapter II, Section D, if the t-value of a parameter was greater than 2.0 in the regression process, the parameter was used as an independent variable. The final models for each of the metro areas were each slightly different from the general form equation. All the models included the term Nonlin, Xmitt, and Trend. None of the models included all predictor variables and no two models had a common predictor variable set. The model coefficients were unique to each area, but the values of the parameter were of the same order of magnitude for both the standard and Hi-Lo models (Table 5.4 and Table 5.5).

Table 5.4 Model Coefficients for 2005 NLR Ozone Forecast Models
(Standard Regression)

Variable	Coef.	ASH	BWG	CVG	LEX	LOU	OWE	PAH
Intercep		-182.90	-173.64	-237.61	-152.32	-200.50	-233.62	-220.32
Nonlin	b1	0.76	0.71	0.88	0.58	0.75	0.77	0.94
Xmitt	b2	333.68	333.55	396.96	300.35	362.56	409.88	385.16
Trend	b3	-1.78	-0.72	-0.38	-0.90	-1.50	-1.04	-1.34
RHx	b4	-0.15	-0.33	-0.16	-0.25	-0.29	-0.20	
Tmn_dep	b5		0.17	0.11	0.12	0.09	0.12	0.18
WS	b6	-0.52	-0.56		-0.30		-0.29	-0.27
CC	b7	-0.58			-0.45		-0.59	-0.84
Dewpt	b8							-0.28
Traj	b9				2.05	3.68		
OZ48	b10				8.03	8.43		
TS	b11					-2.60		

Table 5.5 Model Coefficients for the Seven NLR Ozone Forecast Models.
(Hi-Lo Regression)

Variable	Coef.	ASH	BWG	CVG	LEX	LOU	OWE	PAH
Intercep		-198.16	-249.07	-319.28	-231.60	-221.12	-255.81	-246.99
Nonlin	b1	0.62	0.84	0.97	0.66	0.69	0.92	1.05
Xmitt	b2	387.10	458.19	528.79	425.87	405.13	447.51	425.05
Trend	b3	-2.94	-0.92		-1.40	-2.13	-1.98	-2.48
RHx	b4	-0.21	-0.50	-0.26	-0.37	-0.31	-0.41	
Tmn_dep	b5							
WS	b6	-0.60	-0.55					
CC	b7	-1.36						-1.33
Dewpt	b8							-0.34
Traj	b9					5.55		
OZ48	b10				8.58	10.17		
TS	b11							

The parameters Traj and OZ48 were used for the Louisville and Lexington forecast models. Application of Traj and OZ48 resulted in an improvement in the model accuracy (Cobourn and Hubbard, 1999). However, when the model is used for operational forecasts, the values of the Traj and OZ48 need to be determined manually by an air quality professional. Louisville and Lexington have had professional ozone forecasters during recent ozone seasons. The other communities did not. Therefore, the two transport parameters were not used in the models applied to the automated internet ozone forecasts.

5.2.2 Evaluating models with calibration data set

Model performance on the calibration data set was evaluated by comparing the model estimates with the observed ozone concentrations. For the seven 2005 ozone forecast models, the R^2 for the model fits ranged from 0.72 (for Ashland) to 0.80 (for Covington). That indicates that the ozone forecast models can explain at least 72% of the local ozone variance. The bias for each hybrid model was near zero. This was expected, since the standard and Hi-Lo basic models each had zero bias for the model fits. The MAE and RMSE were used to evaluate the deviation of the predicted values from the observed values. The MAE for the seven forecast models ranged from 5.57 ppb (for Lexington) to 7.32 ppb (for Ashland). The NMAE varied little by location, and was typically 11~12% (Table 5.6).

Table 5.6 Statistics of the Model Performance on Calibration Data Sets

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH
Bias (ppb)	0.33	0.61	0.39	0.05	0.12	0.12	0.13
MAE (ppb)	7.32	6.55	6.76	5.57	6.32	6.60	6.78
RMSE (ppb)	9.11	8.39	8.61	7.08	8.19	8.25	8.45
NMAE	12.8%	11.7%	11.3%	10.7%	10.7%	11.8%	12.1%
$[O_3]_{avg}$	57.1	55.9	60.0	51.8	59.1	56.0	56.0
R^2	0.72	0.79	0.80	0.75	0.79	0.72	0.72
Count	750	742	757	762	763	748	759

The forecast skill of an ozone forecast model was evaluated with the indexes DR, FAR, CSI, and SR. These detection indexes indicate the effectiveness of a model in predicting high ozone concentrations. The values of the DR, FAR, and CSI were affected by the alarm threshold. The unhealthy limit of NAAQS is 85 ppb for 8-hr ground-level ozone. However, an alarm threshold slightly lower than the NAAQS unhealthy limit could significantly increase the detection rate without issuing too many false alarms. In this study, the alarm threshold was chosen as 80 ppb. For the seven areas, the Louisville and Lexington models had relatively high DR and low FAR, due to the use of transport parameters Traj and OZ48; the Bowling Green model had the lowest DR of 0.38 and the highest FAR of 0.54, probably because the unedited meteorological data were used to build the database (Table 5.7). These statistics compare favorably with those of the previous NLR ozone forecast models.

Table 5.7 Detection Statistics for the 2005 Ozone Forecast Models (1999-2004 calibration data, Threshold 80 ppb)

Statistic	Sym	ASH	BWG	CVG	LEX	LOU	OWE	PAH
Detection Rate	DR	0.55	0.38	0.60	0.56	0.71	0.40	0.57
False Alarm Rate	FAR	0.28	0.54	0.23	0.17	0.15	0.31	0.38
Critical Success Index	CSI	0.50	0.33	0.54	0.63	0.66	0.35	0.43
Success Rate	SR	0.96	0.97	0.94	0.99	0.97	0.97	0.98
Events	EV	58	36	93	16	70	31	30
Detected Exceedences	EX	22	6	42	5	39	10	12
Exceedences	DE	40	16	70	9	55	25	21
Alarms	AL	40	26	65	12	54	16	21
False Alarms	FA	11	14	15	2	8	5	8

The scatter plot of the model fits versus the observed ozone concentrations illustrates the correspondence between model estimates and observations. Figure 5.5 is a sample scatter plot for the Ashland forecast model. The relatively dense scatter of points near to the diagonal line indicates the good correlation and agreement between predictions and observations. Scatter plots for the other metro areas had a similar pattern.

The residual is defined as the difference between the observed and predicted values. 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 concentration. The residuals plot for the seven ozone forecast models each had a pattern similar to the example plot for the Ashland forecast model (Figure 5.6).

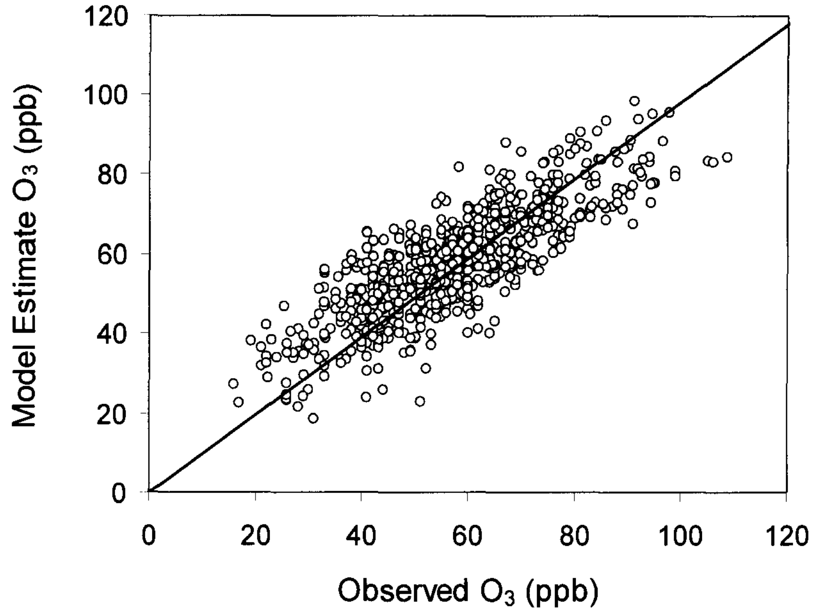


Figure 5.5 Scatter plot of model estimates against observed O_3 (Ashland, 2000-2004). The diagonal indicates the perfect correspondence line.

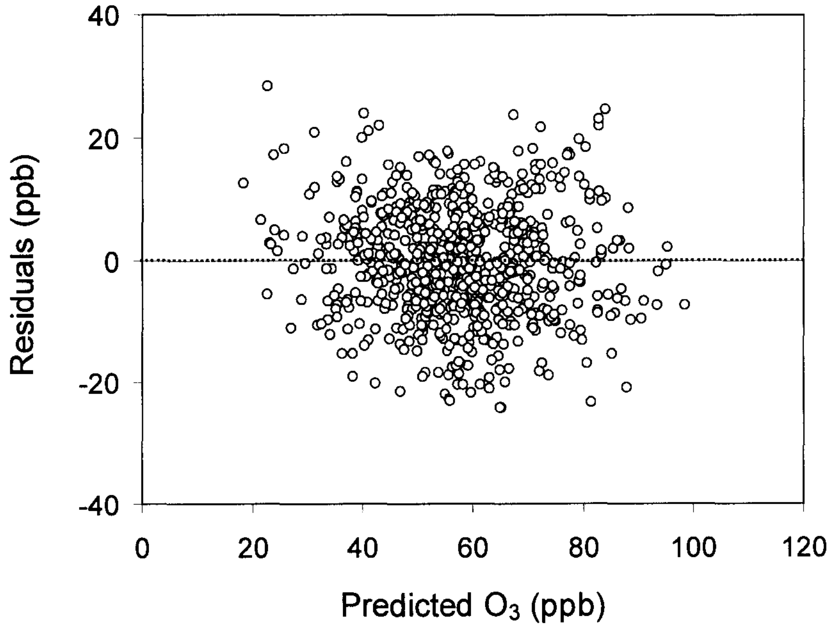


Figure 5.6 Residuals of the hybrid model versus the predicted ozone concentrations. (Ashland, 2000-2004)

5.2.3 Model performance in predicting O_3 concentrations of 2005 ozone season

To test the final NLR-fuzzy and basic-fuzzy models on an independent data set, the models were used to hindcast the peak ozone concentrations during the 2005 ozone season. The Bias of the model hindcasts ranged from -3.1 ppb (Louisville) to 2.8 ppb (Bowling Green). The MAE of the model hindcast ranged from 5.40 ppb (Lexington) to 7.20 ppb (for Bowling Green), on average for the seven areas, 6.27 ppb. The MAE was 9.8% – 12.7% of the corresponding average observed O_3 for each of the seven metro areas (Table 5.8).

Table 5.8 Statistics of the NLR Ozone Model Hindcasts for 2005

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH
Bias (ppb)	1.60	2.80	1.30	0.10	-3.10	-1.30	-0.60
MAE (ppb)	5.90	7.20	6.50	5.40	6.70	6.00	6.20
NMAE	10.3%	12.7%	10.3%	9.8%	10.8%	10.3%	11.0%
$[O_3]_{avg}$	57.3	56.9	63.2	55.0	61.8	58.3	56.6
detected exceedence	2	0	12	0	5	0	0
exceedence	3	1	18	0	8	2	1
false alarms	0	5	5	4	2	0	0
alarms	5	6	25	11	11	0	0

For most of the metro areas, there were typically just a few NAAQS exceedences during 2005 (Table 5.8). Thus, the annual critical forecast statistics for most metro areas were not statistically meaningful. Taken together, though, the combined statistics for all metro areas during 2005 ozone seasons (33 total exceedence days) provide an indication of the critical forecast performance. For all Kentucky metro areas during the study period, the DR was 0.58 and the FAR was 0.26. This critical forecast performance was reasonably good and compares favorably with other reported operational O$_3$ forecast models.

The model hindcasts tracked the day-to-day ozone variation reasonably well, as the time series plot for Louisville shows (Figure 5.7).

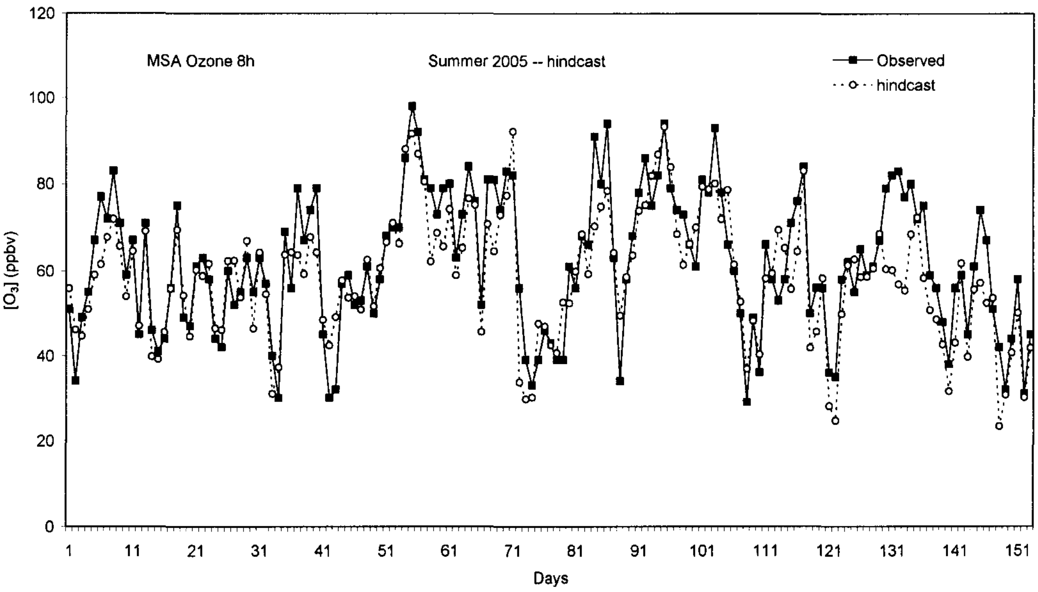


Figure 5.7 Time series of observed 8-hr ozone in Louisville and hindcasts for the NLR model during the 2005 ozone season.

During the 2005 ozone forecast season, the meteorological forecast data used for the NLR model forecasts were saved. This data consisted of text files of the model output statistic (MOS) forecasts from the daily 1200 UTC NGM numerical weather model runs for each metro area. The availability of this data made it possible to evaluate the NLR models in the forecast mode. The Bias of the model forecasts ranged from -6.0 ppb (Owensboro) to 0.5 ppb (Covington). The MAE of the model forecasts ranged from 7.90 ppb (Bowling Green) to 9.80 ppb (Owensboro), on average, 8.61 ppb. The MAE was 12.8% – 16.8% of the corresponding average observed O$_3$, which was greater than those of the model hindcasts for each of the areas (Table 5.9). For the combined statistics for all metro areas during 2005 ozone seasons, the DR was 0.64 and the FAR was 0.57.

Table 5.9 Statistics of the NLR Ozone Model Forecasts for 2005

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH
Bias (ppb)	-1.90	-0.60	0.50	-3.30	-5.10	-6.00	-5.60
MAE (ppb)	8.40	7.90	8.10	8.30	9.10	9.80	8.70
NMAE	15.7%	13.9%	12.8%	15.1%	15.7%	16.8%	15.4%
$[O_3]_{avg}$	57.3	56.9	63.2	55.0	61.8	58.3	56.6
detected exceedence	3	0	13	0	5	0	0
exceedence	3	1	18	0	8	2	1
false alarms	15	6	8	2	2	0	1
alarms	5	7	30	9	12	1	1

Figure 5.8 is an example of time series plot for the Louisville forecast model. The predictions are seen to agree quite closely with the observed concentrations on most days.

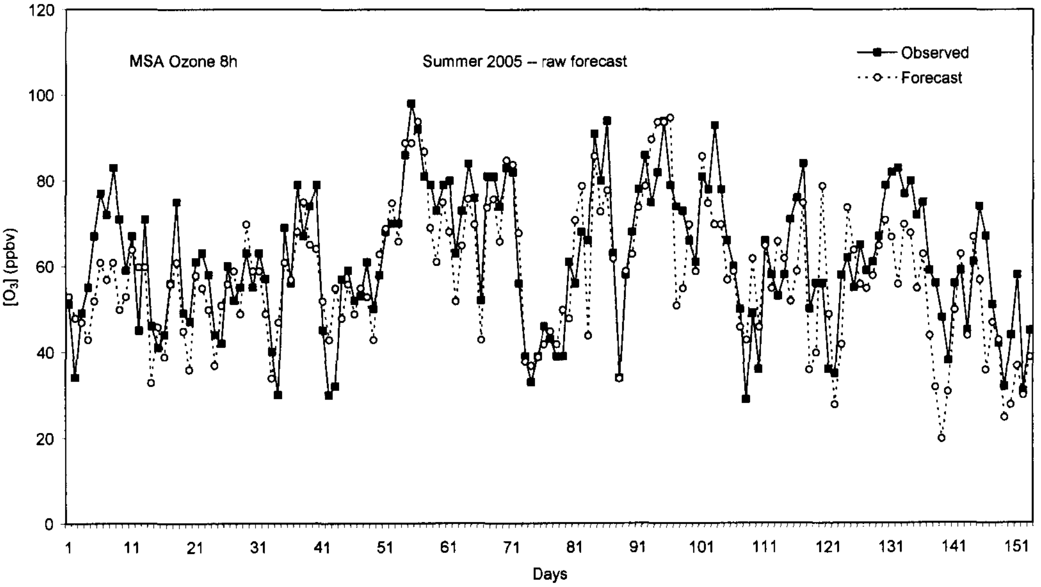


Figure 5.8 Time series of observed 8-hr ozone in Louisville and forecasts for the NLR model during the 2005 ozone season.

5.3 Development of Fuzzy System Ozone Forecast Models

5.3.1 Theory of fuzzy systems

5.3.1.1 Standard fuzzy system

Fuzzy modeling is a tool aimed at using the information observed from a complex phenomenon to derive a quantitative model. A general fuzzy system consists of four parts: a rule base, an inference mechanism, a fuzzification interface, and a defuzzification interface. It has the following structure:

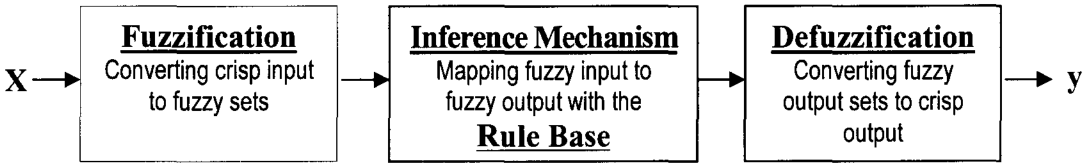

X
Fuzzification
Converting crisp input to fuzzy sets
Inference Mechanism
Mapping fuzzy input to fuzzy output with the
Rule Base
Defuzzification
Converting fuzzy output sets to crisp output
y

Figure 5.9 Structure of a general fuzzy system

The inputs and outputs consist of real numbers. The fuzzification block converts the crisp inputs to fuzzy sets. The inference mechanism uses the fuzzy rules in the rule-base to produce fuzzy conclusions, and the defuzzification block converts these fuzzy conclusions to crisp outputs (Passino and Yurkovich, 1998).

A fuzzy system is a static nonlinear mapping between inputs and outputs. The mapping of the inputs to the outputs for a fuzzy system is in part characterized by a set of rules in if-then form,

$$\textbf{If } \text{premise } \textbf{Then } \text{consequent} \tag{5.8}$$

The inputs of the fuzzy system are associated with the premise, and the outputs are associated with the consequent. The standard form of a multi-input single-output (MISO) of a linguistic rule is

$$Rule\ i: \qquad If\ X\ is\ A^i, then\ y\ is\ B^i;\quad i = 1,\cdots,R \tag{5.9}$$

where $X=(x_1, x_2, \ldots, x_n)$ is the set of input variables, the number of input variables is n; y is the output variable. The fuzzy sets $A^i = (A_1^i, A_2^i, \cdots, A_n^i)$ and B^i are input and output fuzzy sets, respectively. A fuzzy set is used to heuristically quantify the meaning of linguistic variables, values, and rules. It is a crisp set of pairings of elements coupled with their associated membership values defined by membership functions. Rule i above states that if a given input set X can associated with a known pattern, then a rule specific to A^i will give an estimate of the associated output y (Jorquera et al, 1998).

The membership function associates with fuzzy sets A^i and B^i. It maps the elements of the input or output variables to [0,1]. The membership function describes the "certainty" that an element of the variables may be classified linguistically as a specific linguistic value. There are many choices for the shape of the membership function, including singleton, triangular, trapezoidal, and Gaussian membership functions, etc. (Figure 5.10). These membership functions each provide a different meaning for the linguistic values that they quantify. The shape of membership functions is chosen by the fuzzy system designer.

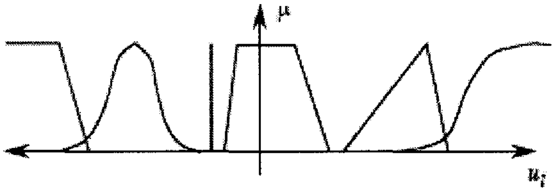


Figure 5.10 Typical membership functions for fuzzy system models (Passino and Yurkovich, 1998)

The fuzzification process specifies how the fuzzy system will convert its numeric inputs into fuzzy sets so that they can be used by the fuzzy system. Generally, the "singleton fuzzification" is used in implementations to produce a fuzzy set for the inputs, which describes the certainty of the input taking on its measured value. The singleton fuzzification was used as the membership function. This method simplifies computational complexity to the inference process and achieves comparable functional capabilities with other fuzzification methods (Gaussian and triangle fuzzification, etc.).

The inference mechanism has two steps. The first step is determining the extent to which each rule is relevant to the current situation as characterized by the inputs. In this step, a membership value μ_i is formed for the i^{th} rule's premise that represents the certainty that each rule's premise applies to the given inputs. Second, the inference step determines implied fuzzy sets by combining the membership values for the premise and consequent.

A number of defuzzification strategies exist. The typical defuzzification techniques for the implied fuzzy sets include center of gravity (COG) and center-average. With center-average defuzzification, a crisp output y is chosen using the centers of each of the output membership functions and the maximum certainty of each of the conclusions represented with the implied fuzzy sets. When using singleton fuzzification and center-average defuzzification, a mathematical representation of a MISO fuzzy systems is

$$y = \frac{\sum_{i=1}^{R} b_i \mu_i}{\sum_{i=1}^{R} \mu_i} \tag{5.10}$$

where R is the number of rules, b_i is the center of the output membership function, μ_i is the premise membership value for the i^{th} rule.

5.3.1.2 Takagi-Sugeno fuzzy system

When the consequence part of the rule uses a function g_i instead of a linguistic term with an associated membership, the fuzzy system is referred to as a Takagi-Sugeno (T-S) fuzzy system, or "functional fuzzy system". The ith rule of a MISO functional fuzzy system has the form

$$If\ X\ is\ A^i, then\ y = g_i \quad i = 1, \cdots, R \tag{5.11}$$

The premise of this rule is the same as that for the standard fuzzy system. However, the consequent uses a function g_i that does not have an associated membership function. Virtually any function can be used for g_i depending on the application. The independent variables of function g_i may include the input variables of the fuzzy system $(x_1, x_2, \cdots, x_n)$ on any other variables. In this study, g_i is defined as an affine function using the input variables of the fuzzy system as its independent variables,

$$g_i = a_{i,0} + a_{i,1}x_1 + \cdots + a_{i,n}x_n \tag{5.12}$$

The crisp output of a T-S fuzzy system is a weighted average of the outputs g_i for $i = 1, \cdots, R$. It is given by

$$y = \frac{\sum_{i=1}^{R} \mu_i g_i}{\sum_{i=1}^{R} \mu_i} \tag{5.13}$$

The function $\mathrm{g_i}$ defines an affine relationship between the inputs and output. The T-S fuzzy system performs a nonlinear interpolation between linear mappings.

5.3.2 Methodology

The design and construction of a fuzzy system is somewhat of an art, in that many possibilities exist, and the designer must make choices based upon skill and experience. The basic choices confronting the designer concern the set of predictor parameters used in the fuzzy model, the type of fuzzy model, type of membership functions, and number of rules. In this study, a special T-S fuzzy system with predefined membership functions was employed.

Fuzzy identification refers to the process of determining the parameters of a fuzzy system (usually includes membership function centers, widths, coefficients, etc), by calibrating the fuzzy system with the training data set. There are several methods that can be used for fuzzy identification, including the least square (LS) method, recursive least square method (RLS), gradient methods, etc. Fuzzy clustering of input data with least square approach to finding consequents was used to identify the T-S fuzzy system in this study. Fuzzy clustering is the partitioning of the input portion of the training data into fuzzy subsets based on similarities between the data. The well-known fuzzy c-means algorithm was used to cluster the input data. Fuzzy c-means is an iterative algorithm used to find input membership functions.

The T-S fuzzy system in this study has the general form described by Eq. 5.11-5.13. However, the membership function shape was predefined. It is calculated using an alternating optimization algorithm (Passino and Yurkovich, 1998),

$$\mu_i(\mathbf{x}) = \left[\sum_{k=1}^{R} \left(\frac{|\underline{\mathbf{x}} - \underline{\mathbf{v}}^i|^2}{|\underline{\mathbf{x}} - \underline{\mathbf{v}}^k|^2} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{5.14}$$

Here, the input variable $\underline{\mathbf{x}}$ and cluster center $\underline{\mathbf{v}}^i$ are vectors with the dimension equal to the number of input variables. The parameter R represents the number of rules and clusters of the fuzzy model. The parameter "m>1" is referred to as the fuzziness factor, which determines the amount of overlap of the clusters. A smaller value of m represents a smoother membership function. Before applying the c-means algorithm to find cluster centers $\underline{\mathbf{v}}^i$, the parameters R and m need to be selected by the designer. Selection of R and m in the first design is somewhat arbitrary. The final value of R and m will be determined by comparing the performance of the fuzzy system on both the training data sets and independent testing data set.

A fuzzy c-means algorithm is used to find the cluster centers $\underline{\mathbf{v}}^i$, so as to determine the value of membership μ_{ij}, which is the membership of ith data pair in the jth cluster. The fuzzy c-means algorithm is realized by minimizing the objective function,

$$J = \sum_{j=1}^{M} \sum_{i=1}^{R} (\mu_{ij})^m \, |\underline{x}^j - \underline{v}^i|^2 \tag{5.15}$$

where M is the number of input-output data pairs in the training data set. Here a data pair refers to an input vector $\underline{x}^j$ with n elements $[x_1, x_{2,} \cdots, x_n]^j$ and its resulting output vector y^j with only one element for a MISO fuzzy system. R is the number of clusters, also the number of rules. Minimization of the objective function results in cluster centers being determined to represent clusters of data.

The fuzzy c-means method is an iterative algorithm. In the first iteration, the initial cluster centers $\underline{\mathbf{v}}_0^{\mathbf{i}}$ for each of the clusters (rules) were randomly chosen so that the initial cluster centers were evenly distributed within the data range. Then a new cluster $\underline{v}_{new}^i$ is calculated with the following equation,

$$\underline{v}_{new}^i = \frac{\sum_{j=1}^{M} x^j (\mu_{ij}^{new})^m}{\sum_{j=1}^{M} (\mu_{ij}^{new})^m} \tag{5.16}$$

where $\mu_{\mathbf{ij}}^{\mathbf{new}}$ is the membership value for the i^th^ rule with j^th^ data pair. It is given as,

$$\mu_{ij}^{new} = \left[\sum_{k=1}^{R} \left(\frac{| \underline{x}^j - \underline{v}_{old}^i |^2}{| \underline{x}^j - \underline{v}_{old}^k |^2} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{5.17}$$

In this equation, $\underline{v}_{old}^i$ is the cluster center obtained from the previous iteration. In the first iteration, it is the initial cluster center $\underline{v}_0^i$. The $\underline{v}_0^i$ needs to be carefully chosen to avoid $| \underline{x}^j - \underline{v}_{old}^i | = 0$. In that case the $\mu_{\mathrm{ij}}^{\mathrm{new}}$ is undefined. The distance between the new cluster $\underline{v}_{new}^i$ and previous cluster center $\underline{\mathbf{v}}_{\mathbf{old}}^{\mathbf{i}}$ is defined by,

$$\varepsilon_d^i = | \underline{v}_{new}^i - \underline{v}_{old}^i | \tag{5.18}$$

The value of ε_d^i is compared to an error tolerance ε_c. The tolerance ε_c is the amount of error allowed in calculating the cluster centers. Usually it is a small number, designated by the designer. If $\varepsilon_d^i < \varepsilon_c$ for all the cluster centers, the cluster centers $\underline{v}_{new}^i$ accurately represent the input data. Let the current $\underline{v}_{new}^i$ be the final cluster centers $\underline{v}^i$. Otherwise iteratively repeat the process untill the final cluster centers are found.

With the cluster centers $\underline{v}^i$ determined, the premise part of the fuzzy system is defined. Then we can apply the weighted least square method to find the coefficients of the linear function g_i expressing the consequent of rule R_i in fuzzy system. The coefficients for the ith rule are expressed by a vector $\underline{a}_i = [a_{i,0}, a_{i,1}, \cdots, a_{i,n}]$.

The equation used to compute a_i is given by

$$\underline{a}_i = (\overline{X}^T D_i^2 \overline{X})^{-1} \overline{X}^T D_i^2 Y \tag{5.19}$$

where $\overline{X}$ and Y are matrices composed of input and output training data. The matrix D_i is a diagonal matrix containing the values of corresponding membership functions. They are defined as,

$$\overline{X} = \begin{bmatrix} 1 & \cdots & 1 \\ x^1 & \cdots & x^M \end{bmatrix}^T \tag{5.20}$$

$$Y = [y^1, \quad \cdots, \quad y^M]^T \tag{5.21}$$

$$D_i = \begin{bmatrix} \mu_{1,i} & & & \\ & \mu_{2,i} & 0 & \\ & 0 & \cdots & \\ & & & \mu_{M,i} \end{bmatrix} \tag{5.22}$$

With the cluster centers v^i and coefficients a_i determined using the training data, the Takagi-Sugeno fuzzy system is constructed.

5.3.3 Construction of basic-fuzzy system and NLR-Fuzzy system ozone forecast models

The Takagi-Sugeno fuzzy system ozone forecast models were developed with the clustering method for the seven metro areas: Ashland, Owensboro, Bowling Green, Covington, Lexington, Louisville, and Paducah. The training data were created by the databases which consisted of ozone air quality data and meteorological data over a five year period, 1999-2003. The number of training data pairs (M) for the seven metro areas depend on the effective data in each database. In this study, data pairs used for the fuzzy system models ranged from 741 (for Bowling Green) to 764 (for Louisville). Development of the previous 2004 NLR ozone forecast models provided the input variables for the fuzzy system models. The candidate input variables have been described in section 5.1 (Table 5.2). All model terms shown in the table were statistically correlated with local O_3 at the 95% confidence level.

There were two types of T-S fuzzy models developed in this study: basic-fuzzy models and NLR-fuzzy models. The input variables used in the basic-fuzzy models were the same observed meteorological data and deterministic parameters used in the NLR model (Table 5.10). The NLR-fuzzy models used similar sets of variables, except that the nonlinear term from the NLR model replaced the three meteorological variables Tmax, WS, and RH which had been incorporated into this term (Table 5.11).

Table 5.10 Input Variables for Seven Metro Area Basic-fuzzy System O_3 Models

Variables	ASH	BWG	CVG	LEX	LOU	OWE	PAH
x_1	Nonlin	Nonlin	Nonlin	Nonlin	Nonlin	Nonlin	Nonlin
x_2	Xmitt	Xmitt	Xmitt	Xmitt	Xmitt	Xmitt	Xmitt
x_3	Trend	Trend	Trend	Trend	Trend	Trend	Trend
x_4	RH	RH	RH	RH	RH	Tmn_dep	Tmn_dep
x_5	CC	Tmn_dep	Tmn_dep	OZ48	OZ48	CC	CC
x_6	WS	WS		Traj	Traj	Dewpt	Dewpt
x_7				CC	Tmn_dep	WS	WS
x_8					TS		
Count	6	6	5	7	8	7	7

Table 5.11 Input Variables for Seven Metro Area NLR-fuzzy System O_3 Models

Variables	ASH	BWG	CVG	LEX	LOU	OWE	PAH
x_1	Tmax	Tmax	Tmax	Tmax	Tmax	Tmax	Tmax
x_2	WS	WS	WS	WS	WS	WS	WS
x_3	RHx	RHx	RHx	RHx	RHx	RHx	RHx
x_4	Xmitt	Xmitt	Xmitt	Xmitt	Xmitt	Xmitt	Xmitt
x_5	Trend	Trend	Trend	Trend	Trend	Trend	Trend
x_6	RH	RH	RH	RH	RH	Tmn_dep	Tmn_dep
x_7	CC	Tmn_dep	Tmn_dep	OZ48	OZ48	CC	CC
x_8				Traj	Traj	Dewpt	Dewpt
x_9				CC	Tmn_dep		
x_{10}					TS		
Count	6	6	5	7	8	7	7

A computer program (Appendix A.1) was used for training the fuzzy system models. Equations 5.14 - 5.18 were applied to realize the iterative process for finding cluster centers $\underline{\mathbf{v}}^{\mathrm{i}}$. The initial cluster centers with 5 rules and $m = 5$ for Louisville NLR-fuzzy system model are shown in Table 5.12. The error tolerance ε_{c} used for finding each cluster center $\underline{\mathbf{v}}^{i}$ was chosen as 0.01. Equations 5.19 – 5.22 were used for finding the coefficients $\underline{\mathbf{a}}_i$ of the fuzzy consequence.

Table 5.12 Initial Cluster Centers v_0^i for Louisville NLR-fuzzy System O_3 Model

Variables	v_0^i	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	v_0^1	36.03	53.62	71.22	88.82	106.41
Xmitt	v_0^2	0.62	0.62	0.63	0.64	0.65
Trend	v_0^3	3.60	2.80	2.00	1.20	0.40
RH	v_0^4	92.47	77.41	62.35	47.29	32.23
OZ48	v_0^5	0.10	0.30	0.50	0.70	0.90
Traj	v_0^6	0.10	0.30	0.50	0.70	0.90
Tmn_dep	v_0^7	-15.10	-6.30	1.50	9.30	17.10
TS	v_0^8	0.90	0.70	0.50	0.30	0.10

To determine the optimum combination of R and m values for application in the model, a series of NLR-fuzzy models for the Louisville metro area were developed with different combinations of R (1, 3, 5, 10, 15, 20, 25, 30) and m (1.5, 2, 3, 4, 5, 6). The resulting models were evaluated by comparing the model estimates from the 1999-2003 calibration period (training data) and model hindcasts from the 2004 ozone season (test data) with the observed O_3 concentrations, using mean absolute error (MAE) as the criterion of performance. With the training data, the models achieved the best performance with the combination m=3 and R=25 (Figure 5.11). However, with the 2004 test data, the combination m=5 and R=5 produced the best results (lowest MAE). At higher values of R the performance was equivalent up to about 15 rules; then deteriorated thereafter, indicating that over-training had occurred at higher R values (Figure 5.12). So for simplicity and to avoid over-training, the combination R=5 and m=5 was chosen for all of the NLR-fuzzy models and basic-fuzzy models.

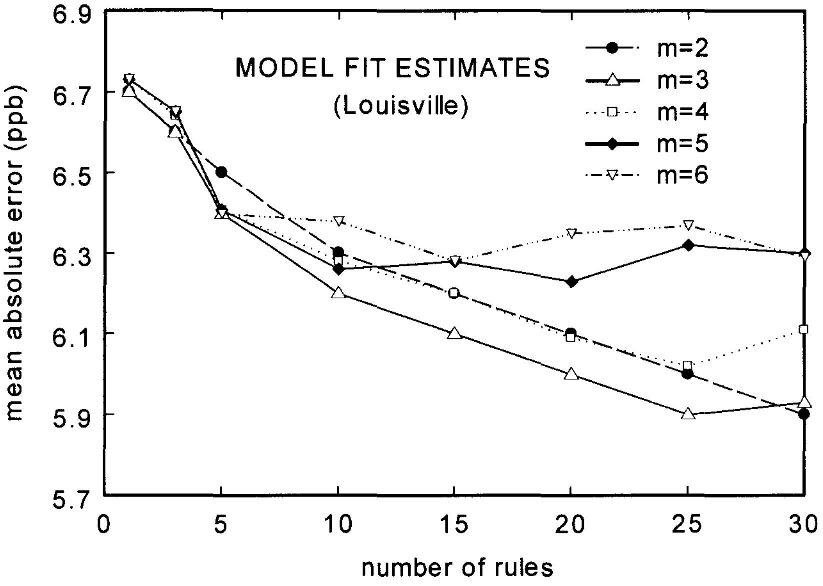


Figure 5.11 Variation of the mean absolute error of the NLR-fuzzy model fit for Louisville (1999-2003) for selected values of fuzziness factor (m) and number of rules (R).

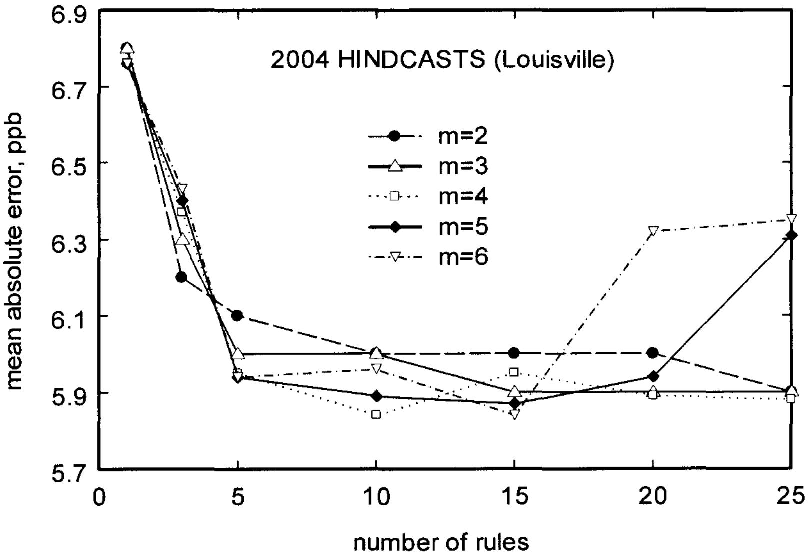


Figure 5.12 Variation of the mean absolute error of 2004 NLR-fuzzy model hindcasts for Louisville, for selected values of fuzziness factor (m) and number of rules (R).

The final fuzzy system models were characterized by the number of rules R, the fuzziness factor (m), the fuzzy cluster centers ($\underline{\mathbf{v}}^{i}$), and the coefficients of the fuzzy consequence ($\underline{\mathbf{a}}_{i}$). In addition, a training parameters, ε_c, representing the error tolerance in the iterative training process was designated for each model. Each of the seven fuzzy system models were configured with the same number of rules (R=5) and fuzziness factor (m=5). All models were trained using the same error tolerance ($\varepsilon_c = 0.01$). The cluster centers and coefficients were unique for each model. As an example, the parameters for Louisville NLR-fuzzy system model are listed in the Table 5.13 and Table 5.14. Each of the rules is actually a linear regression model which evaluates the O_3 concentration based upon the input data. The value of the membership function for the ith rule gives the weight of this rule for the given data pair consisting of the input data vector and the observed O_3 concentration. The predicted O_3 concentration determined by the fuzzy model is the weighted average of the outputs for the five rules.

Table 5.13 Final Cluster Centers v_f^i for Louisville NLR-fuzzy system O_3 Model

Variables	**v^i**	**Rule 1**	**Rule 2**	**Rule 3**	**Rule 4**	**Rule 5**
Nonlin	V_1^i	42.93	58.37	75.17	55.44	100.06
Xmitt	V_2^i	0.64	0.65	0.65	0.64	0.64
Trend	V_3^i	2.04	2.30	1.98	1.91	1.21
RH	V_4^i	87.76	69.80	55.73	45.18	36.33
OZ48	V_5^i	0.12	0.12	0.19	0.00	0.41
Traj	V_6^i	0.12	0.20	0.38	0.12	0.49
Tmn_dep	V_7^i	1.96	3.88	2.86	-6.76	1.19
TS	V_8^i	0.41	0.35	0.05	0.00	0.00

Table 5.14 Coefficients a_i for Louisville NLR-fuzzy System O_3 Model

Variables	Coef.	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-185.53	-117.95	-285.18	-303.26	-180.65
Nonlin	a_1	0.42	0.54	0.81	0.79	0.44
Xmitt	a_2	379.65	281.76	482.46	508.14	392.20
Trend	a_3	-2.57	-3.07	-1.36	-0.94	-0.30
RH	a_4	-0.38	-0.49	-0.23	-0.17	-0.54
OZ48	a_5	13.21	7.35	10.48	9.35	7.43
Traj	a_6	0.39	6.18	2.53	3.48	-2.87
Tmn_dep	a_7	0.17	-0.10	-0.04	0.35	-0.10
TS	a_8	-3.33	-1.49	-6.05	-3.10	-5.55

The parameters for Louisville Basic-fuzzy system model are listed in the Table 5.15 and Table 5.16. Model parameters for other NLR-fuzzy and basic-fuzzy system models refer to Appendix B. In Table 5.13 - 5.15, Rule 5 appears to be associated with meteorological conditions conducive to high ozone. Rule 1 appears to be associated with conditions conducive to low ozone, and rules 2-4 appear to be associated with medium ozone concentrations.

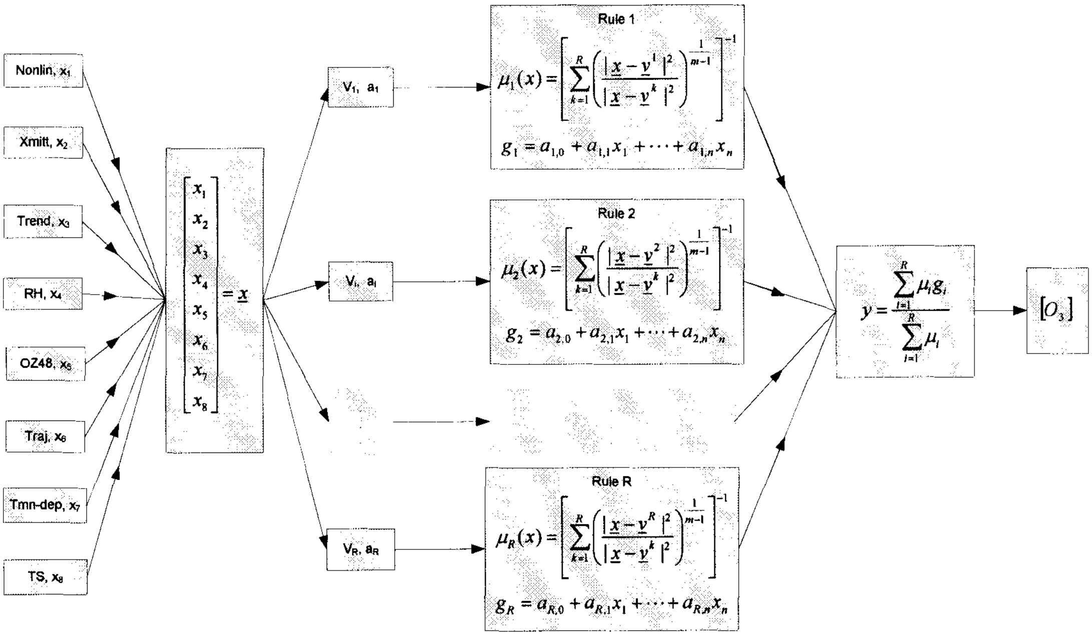
Table 5.15 Final Cluster Centers $v_f^{\,i}$ for Louisville Basic-fuzzy System O_3 Model

Variables	v^i	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Tmax	V_1^i	77.53	85.18	89.25	75.79	93.76
WS	V_2^i	8.81	9.04	8.29	8.67	6.07
RHx	V_3^i	85.84	71.83	60.08	51.87	45.05
Xmitt	V_4^i	0.64	0.65	0.65	0.64	0.64
Trend	V_5^i	1.88	2.19	1.99	2.07	1.05
RH	V_6^i	87.74	71.15	56.30	45.22	35.55
OZ48	V_7^i	0.14	0.12	0.18	0.00	0.35
Traj	V_8^i	0.13	0.22	0.35	0.15	0.40
Tmn_dep	V_9^i	3.05	3.40	3.25	-7.73	-0.80
TS	V_{10}^i	0.40	0.37	0.08	0.00	0.00

Table 5.16 Coefficients a_i for Louisville Basic-fuzzy System O3 Model

Variables	Coef.	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
intercept	a_0	-211.54	-133.55	-235.24	-307.48	-225.46
Tmax	a_1	0.53	0.65	0.85	0.90	1.29
WS	a_2	0.15	-0.41	-0.97	-0.51	-1.07
RHx	a_3	-0.17	-0.14	-0.35	-0.38	-0.36
Xmitt	a_4	399.23	298.06	432.92	518.30	355.47
Trend	a_5	-2.41	-3.37	-1.79	-0.76	-0.61
RH	a_6	-0.32	-0.54	-0.35	-0.19	-0.32
OZ48	a_7	12.79	6.67	10.66	11.94	9.60
Traj	a_8	-1.78	5.85	3.03	3.16	-1.00
Tmn_dep	a_9	-0.08	-0.24	0.13	0.35	-0.12
TS	a_{10}	-3.64	-1.91	-5.19	-8.64	-5.09

The fuzzy system model output can be computed using Equation 5.12 – 5.14 with the parameters determined above. The computer program in Appendix A.2 was used to test the fuzzy system models. The flow chart in Figure 5.13 illustrates the algorithm of a Takagi-Sugeno fuzzy system.



Nonlin, x_1
Xmitt, x_2
Trend, x_3
RH, x_4
OZ48, x_5
Traj, x_6
Tmn-dep, x_7
TS, x_8
$[x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8]^T = \underline{x}$
V_1, a_1
V_i, a_i
V_R, a_R
Rule 1
$\mu_1(x) = \left[\sum_{k=1}^{R}\left(\frac{|\underline{x}-\underline{v}^1|^2}{|\underline{x}-\underline{v}^k|^2}\right)^{\frac{1}{m-1}}\right]^{-1}$
$g_1 = a_{1,0} + a_{1,1}x_1 + \cdots + a_{1,n}x_n$
Rule 2
$\mu_2(x) = \left[\sum_{k=1}^{R}\left(\frac{|\underline{x}-\underline{v}^2|^2}{|\underline{x}-\underline{v}^k|^2}\right)^{\frac{1}{m-1}}\right]^{-1}$
$g_2 = a_{2,0} + a_{2,1}x_1 + \cdots + a_{2,n}x_n$
Rule R
$\mu_R(x) = \left[\sum_{k=1}^{R}\left(\frac{|\underline{x}-\underline{v}^R|^2}{|\underline{x}-\underline{v}^k|^2}\right)^{\frac{1}{m-1}}\right]^{-1}$
$g_R = a_{R,0} + a_{R,1}x_1 + \cdots + a_{R,n}x_n$
$y = \frac{\sum_{i=1}^{R}\mu_i g_i}{\sum_{i=1}^{R}\mu_i}$
$[O_3]$

Figure 5.13 Flow chart of the algorithm of the Tasagi-Sugeno fuzzy system

5.3.4 Model Validation

5.3.4.1 Validation of the fuzzy system models on the calibration data set.

Performance of the seven basic-fuzzy and NLR-fuzzy system ozone forecast models on the calibration data set was evaluated by comparing the model estimates with the observed ozone concentrations from the calibration period, 1999-2003. For each of the seven models, the statistics of the model fit were good (Table 5.17 and 5.18). For both the basic-fuzzy and NLR-fuzzy system models, the biases were close to zero. For the basic-fuzzy system models, the MAE of the model fits ranged from 5.65 ppb (Lexington) to 6.96 ppb (Ashland). On average, the MAE of the seven models was 6.47 ppb, or 10.8% of the mean daily peak O_3 concentration for the period (NMAE). For the NLR-fuzzy system models, the MAE of the model fits ranged from 5.83 ppb (Lexington) to 6.70 (Ashland). The MAE of the fits was on average 6.34 ppb. The average NMAE was 10.5%. Due to the application of the parameter Traj and OZ48 that accounted for the pollutant transport, the NMAE for Louisville and Lexington were slightly lower than those of the other cities. The MAE of Louisville basic-fuzzy and NLR-fuzzy were 6.53 and 6.41 ppb, which were 10.3% and 10.1% of the mean daily peak O_3 concentration respectively. The correlation coefficient R^2 was 0.75 for the basic-fuzzy model and 0.76 for the NLR-fuzzy model. The Lexington model fit had the lowest MAE of 5.65 for basic-fuzzy model and 5.83 ppb for NLR-fuzzy model, which were much lower than the average MAE of the seven models. The low MAE of Lexington was probably because the mean O_3 concentration was lower than that of the other cities.

Table 5.17 Error Statistics of Model Fits for the Basic-fuzzy O_3 Models

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
bias (ppb)	-0.12	-0.02	-0.04	0.10	0.12	0.10	0.20	0.05
MAE (ppb)	6.96	6.12	6.89	5.65	6.53	6.33	6.78	6.47
RMSE (ppb)	8.70	7.90	8.96	7.30	8.35	8.10	8.60	8.27
NMAE (%)	11.4%	10.2%	11.0%	10.2%	10.3%	10.6%	11.4%	10.8%
$[O_3]_{avg}$	60.8	59.8	62.4	55.5	63.5	59.5	59.3	60.12
R^2	0.74	0.72	0.71	0.73	0.75	0.70	0.67	0.72
Count	747	741	758	763	764	751	751	

Table 5.18 Error Statistics of Model Fits for the NLR-fuzzy O_3 Models

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
bias (ppb)	0.10	0.14	0.30	0.02	0.33	0.26	-0.49	0.09
MAE (ppb)	6.70	6.20	6.51	5.83	6.41	6.30	6.41	6.34
RMSE (ppb)	8.47	7.99	8.37	7.46	8.23	8.06	8.20	8.11
NMAE (%)	11.0%	10.4%	10.4%	10.5%	10.1%	10.6%	10.8%	10.5%
$[O_3]_{avg}$	60.8	59.8	62.4	55.5	63.5	59.5	59.5	60.14
R^2	0.72	0.71	0.74	0.72	0.76	0.71	0.71	0.72
Count	747	741	758	763	764	751	751	

The NAAQS for ozone is 0.08 ppm. The value of 85 ppb is used for determination of exceedence, since 84 ppb (0.084 ppm) rounds to 0.08 ppm. For our ozone forecast models, the alarm threshold for unhealthy O_3 concentration was chosen as 80 ppb. Louisville and Covington (Cincinnatti MSA) are the two largest metro areas and both have a long standing ozone problem. The number of ozone exceedence days in Louisville and Covington were 98 and 95 respectively during the period 1999-2003. The NLR-fuzzy forecast model for Louisville and Covington successfully detected 85% and 71% of the respective local ozone exceedence days. The false alarm rates were 0.16 and 0.21 respectively. For the other areas, the NLR-fuzzy forecast model predicted at least

50% of the local ozone exceedence days, with the FAR less than 0.37 (Table 5.19). The forecast skills of the basic-fuzzy system models were close to the corresponding NLR-fuzzy system models (Table 5.20).

Table 5.19 Detection Statistics of Model Fit for the NLR-fuzzy O_3 Models (Calibration data 1999-2003, threshold = 80 ppb)

Statistic	Sym.	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
Detection Rate	DR	0.69	0.54	0.71	0.50	0.85	0.54	0.54	0.62
False Alarm Rate	FAR	0.22	0.24	0.21	0.37	0.16	0.26	0.20	0.24
Critical Success Index	CSI	0.62	0.53	0.63	0.46	0.74	0.47	0.49	0.56
Detected Exceedences	DE	45	21	67	11	83	26	26	39.9
Exceedences	EX	65	39	95	22	98	48	45	58.9
Alarms	AL	81	41	106	30	118	38	35	65.1
False Alarms	FA	18	10	22	11	19	10	7	13.9

Table 5.20 Detection Statistics of Model Fit for the Basic-fuzzy O_3 Models (Calibration data 1999-2003, threshold = 80 ppb)

Statistic	Sym.	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
Detection Rate	DR	0.69	0.54	0.65	0.55	0.79	0.48	0.56	0.61
False Alarm Rate	FAR	0.22	0.21	0.19	0.29	0.14	0.14	0.24	0.20
Critical Success Index	CSI	0.63	0.54	0.60	0.53	0.72	0.46	0.48	0.57
Detected Exceedences	DE	45	21	62	12	77	23	25	37.9
Exceedences	EX	65	39	95	22	98	48	45	58.9
Alarms	AL	79	39	98	28	107	29	34	59.1
False Alarms	FA	17	8	19	8	15	4	8	11.3

Graphical techniques are also useful for evaluating model performance. A scatter plot of model estimates versus observed O_3 concentration visually depicts how well the model fits the data over the entire range of observations. For example, the scatter plot of model estimates vs. observations for the Louisville NLR-fuzzy model indicated a good fit between model estimates and ozone observations for the 1999-2003 calibration data set (Figure 5.14).

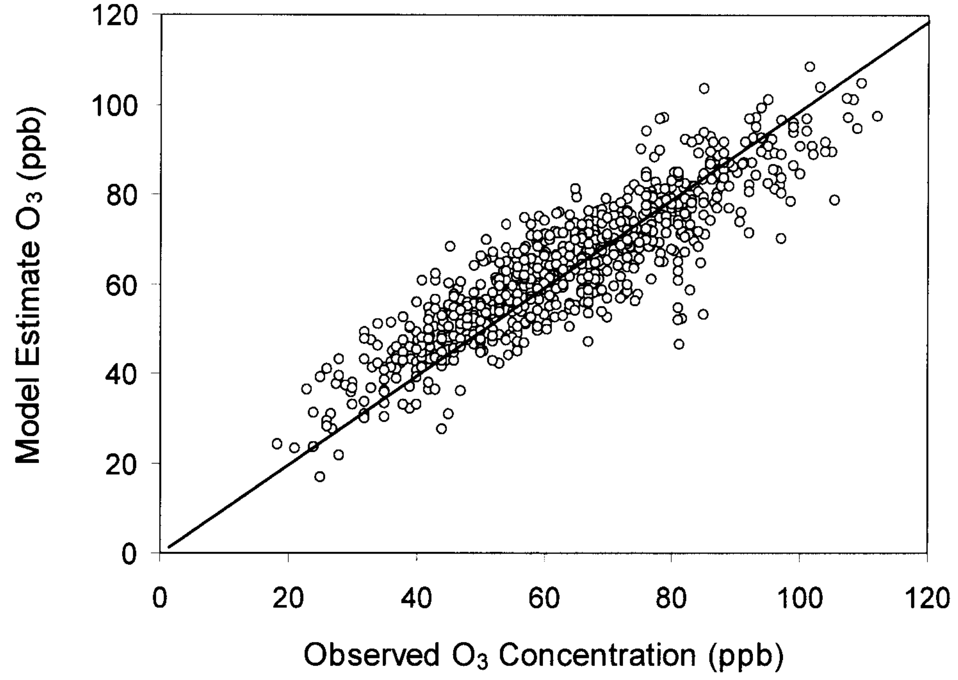


Figure 5.14 Scatter plot of NLR-fuzzy model estimates against observed O_3 concentrations for Louisville

A scatter plot of residuals ($[O_3]_{pred}$ $-[O_3]_{obs}$) versus predicted O_3 concentration indicates the distribution of the prediction errors through the range of observations. For example, the scatter plot of residuals vs. predicted O_3 concentration for Louisville (1999-2003) indicates that the model is essentially unbiased throughout the range of predictions (Figure 5.15). Scatter plots for the other metro areas fuzzy system models were similar to those of the Louisville model.

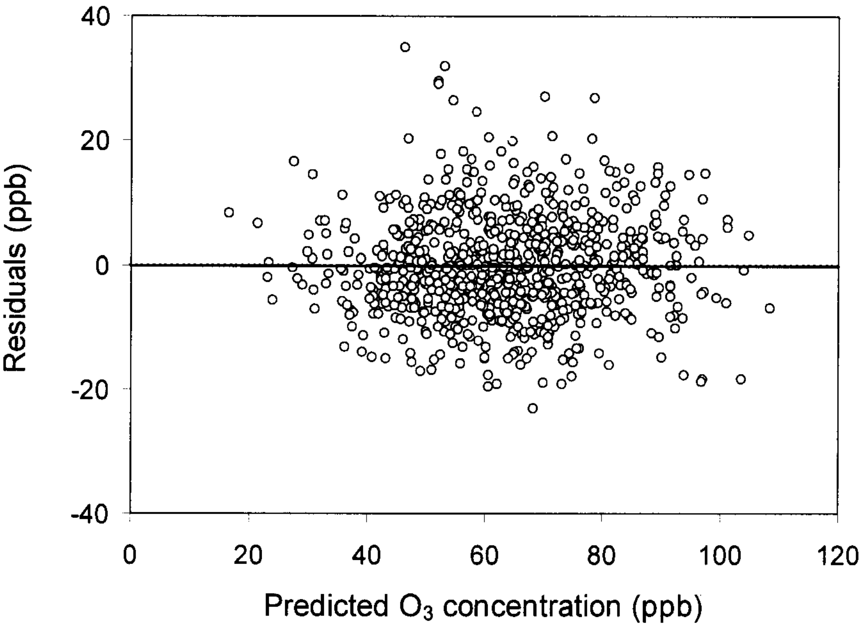


Figure 5.15 Residual plot of the NLR-fuzzy model prediction error versus predicted ozone concentrations (Louisville)

5.3.4.2 Validation of the fuzzy system models with and independent data set

The ozone forecast models were designed to provide ozone forecasts using forecast meteorological data. The basic-fuzzy and NLR-fuzzy system ozone forecast models for the seven metro areas were tested with independent data, by using the fuzzy system models to make both forecasts and hindcasts of the O_3 concentrations during 2004 ozone season. Model predictions made with observed meteorological data are called hindcasts, and model predictions made with forecasted meteorological data are forecasts. Errors in the forecast meteorological data tend to reduce the accuracy of the ozone forecast models.

The model hindcast errors for the year of 2004 were comparable to the model fit errors, but slightly higher, as expected (Table 5.21 and Table 5.22). For the NLR-fuzzy system models, the seven-city average MAE was 7.43 ppb, as compared to 6.34 ppb for the model fits. The average MAE for the basic-fuzzy system models was 7.61, compared to 6.47 ppb for the model fit. The biases of the model hindcasts were all positive, ranging from 1.01 ppb (Louisville) to 9.31 ppb (Ashland) for NLR-fuzzy models, and ranging from 0.96 ppb (Louisville) to 10.10 ppb (Ashland) for basic-fuzzy models. The unusually high bias for the Ashland hindcasts could not be explained. This high systematic error produced a high MAE for Ashland. The Louisville and Lexington models both performed measurably better than the other models. The hindcast MAEs for these two models were both lower than the seven-city average, by more than 1.0 ppb.

Table 5.21 Error Statistics of NLR-fuzzy Model Hindcasts for 2004 Ozone Season

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
Bias (ppb)	9.31	5.79	5.48	2.01	1.01	5.69	6.23	5.65
MAE (ppb)	10.57	7.17	7.37	5.50	5.94	7.31	8.16	7.43
RMSE (ppb)	12.81	8.86	8.90	6.70	7.62	8.92	10.13	9.13
NMAE (%)	21.2%	15.4%	13.0%	12.0%	11.1%	15.6%	16.5%	14.7%
$[O_3]_{avg}$	49.8	49.9	56.7	46.0	53.6	50.0	49.3	50.8
R^2	0.71	0.79	0.90	0.66	0.71	0.77	0.89	0.78
Count	152	150	151	151	152	151	152	

Table 5.22 Error Statistics of Basic-fuzzy Model Hindcasts for 2004 Ozone Season

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
Bias (ppb)	10.10	5.84	5.57	2.19	0.96	5.41	6.22	5.90
MAE (ppb)	11.21	7.17	7.65	5.47	5.93	7.71	8.15	7.61
RMSE (ppb)	13.47	8.85	9.30	6.69	7.66	9.26	10.05	9.33
NMAE (%)	22.5%	15.4%	13.5%	11.9%	11.1%	15.4%	16.5%	15.0%
$[O_3]_{avg}$	49.8	49.9	56.7	46.0	53.6	50.0	49.3	50.8
R^2	0.71	0.81	0.87	0.73	0.71	0.81	0.92	0.79
Count	152	150	151	151	152	151	152	

The NLR-Fuzzy system ozone forecast models also performed well when the models were tested in the forecast mode with forecasted meteorological data as input (Table 5.23 and Table 5.24). The average MAE of the model forecasts for the NLR-fuzzy and basic-fuzzy models were 7.78 ppb and 8.03 ppb respectively, which was about 5% and 6% higher than the values of model hindcasts. The characteristic degradation of model accuracy in going from model fit estimates to model hindcasts, and then to model forecasts is illustrated in Figure 5.16. This degradation is normally observed for an ensample of forecasts, for example for an ozone season. For particular forecasts, it is sometimes the case that the errors of the meteorological forecasts compensate for the built-in model errors, due to random causes unexplained by the model. The forecasts of the Ashland and Paducah models had lower MAEs than for the hindcasts. The probable reason is because the systematic component of the MAE was lowered, since the model biases in both cases dropped significantly in going from hindcast to forecast.

Table 5.23 Error Statistics of NLR-fuzzy Model Forecasts for 2004 Ozone Season

Statistic	**ASH**	**BWG**	**CVG**	**LEX**	**LOU**	**OWE**	**PAH**	**Average**
Bias (ppb)	6.17	1.50	3.71	1.09	0.38	2.21	2.11	2.45
MAE (ppb)	8.88	7.56	8.26	6.42	7.66	7.96	7.71	7.78
RMSE (ppb)	11.20	9.60	10.48	7.81	9.76	10.18	9.81	9.83
NMAE (%)	17.8%	15.2%	15.6%	15.0%	15.3%	15.9%	15.6%	15.3%
$[O_3]_{avg}$	49.8	49.9	56.7	45.9	53.6	50.0	49.3	50.8
R^2	0.70	0.78	0.85	0.64	0.66	0.86	0.77	0.75
Count	152	150	151	151	152	151	152	

Table 5.24 Error Statistics of Basic-fuzzy Model Forecasts for 2004 Ozone Season

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
Bias (ppb)	6.95	1.25	5.07	1.45	0.35	3.14	3.08	2.90
MAE (ppb)	9.57	7.71	8.58	6.50	7.61	8.13	8.13	8.03
RMSE (ppb)	11.88	9.76	10.85	7.86	9.78	10.39	10.27	10.11
NMAE (%)	19.2%	15.5%	15.1%	15.2%	15.2%	16.3%	16.5%	15.8%
$[O_3]_{avg}$	49.8	49.9	56.7	45.9	53.6	50.0	49.3	50.8
R^2	0.70	0.80	0.81	0.70	0.65	0.88	0.87	0.77
Count	152	150	151	151	152	151	152	

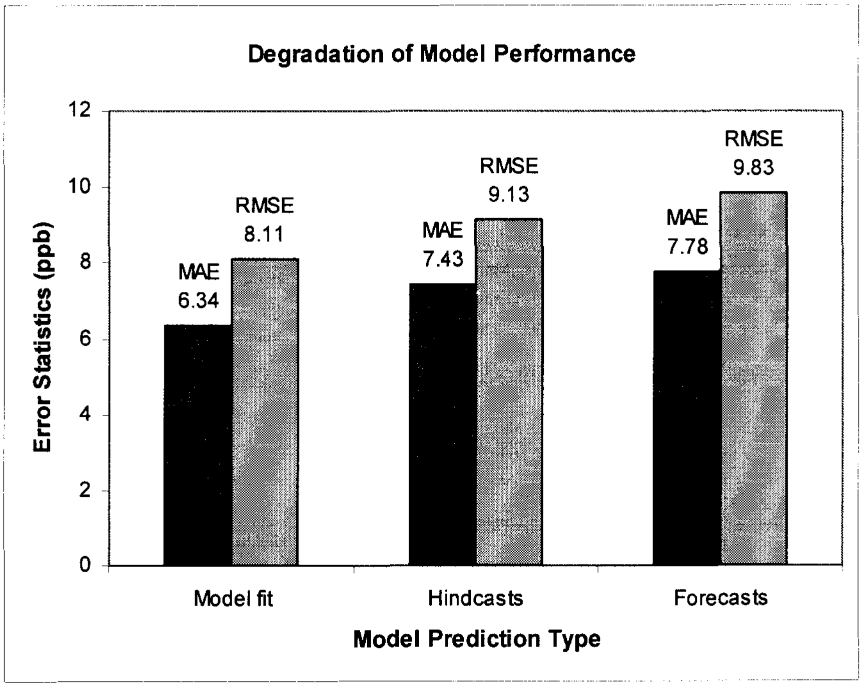


Figure 5.16 Degradation of model performance (data for NLR-fuzzy models) in going from model fit estimates to hindcasts to forecasts. Statistics are model averages for the seven cities. The forecast lead time is approximately 24 hours.

The R^2 values for all hindcasts and forecasts were good, demonstrating that a large portion of the O_3 variation (70% or better in most cases) was explained by the models. This fact is further demonstrated by time series plots of observed O_3 with model predictions (hindcasts and forecasts of NLR-fuzzy models) for the seven models during the 2004 ozone season (Appendix C, Figure A1 to Figure A7), in which all model hindcasts and forecasts tracked the day-to-day ozone variation reasonably well.

5.3.5 Comparison of NLR-fuzzy, basic-fuzzy, and NLR models

The NLR ozone forecast models are well-established and have been used for operational next-day ground-level ozone forecasts for several metropolitan areas in Kentucky since 2005. The performance of the NLR models were compared with NLR-fuzzy and basic-fuzzy models described in the previous sections. Comparison of the model fit estimates for these ozone forecast models showed the NLR-fuzzy and basic-fuzzy models had slightly better performance statistics than that of the NLR model. The performance of the NLR-fuzzy model and basic-fuzzy model were close. For example, the seven city average MAE for the NLR-fuzzy models and basic-fuzzy models were 6.34 ppb and 6.47 ppb respectively. The corresponding value for the NLR models was 6.76 ppb.

Model hindcasts and forecasts for 2004 ozone season for the NLR-fuzzy, basic-fuzzy, and NLR models were compared (Table 5.25 and Table 5.26). During the 2004 ozone forecast season, the meteorological forecast data used for the NLR model forecasts were saved. This data consisted of text files of the model output statistic (MOS) forecasts from the daily 1200 UTC NGM numerical weather model runs for each metro area. The availability of this data made it possible to compare the NLR-fuzzy, basic-fuzzy and NLR regression models operating in the forecast mode.

Table 5.25 Statistics of 2004 Model Hindcasts for the Ozone Forecast Models

	Statistic	ASH	BWG	COV	LEX	LOU	OWE	PAH	Average
NLR-fuzzy	Bias (ppb)	9.3	5.8	5.5	2.0	1.0	5.7	6.2	5.6
	MAE (ppb)	10.6	7.2	7.4	5.5	5.9	7.3	8.2	7.4
	NMAE (%)	21.2%	15.4%	13.0%	12.0%	11.1%	15.6%	16.5%	15.7%
Basic-fuzzy	Bias (ppb)	10.5	5.8	5.6	2.5	1.0	5.4	6.2	5.0
	MAE (ppb)	11.4	7.2	7.7	5.5	5.9	7.7	8.1	7.6
	NMAE (%)	22.9%	15.4%	13.5%	12.0%	11.0%	15.4%	16.4%	15.1%
NLR	Bias (ppb)	9.6	5.7	5.6	2.5	1.0	5.7	5.4	5.6
	MAE (ppb)	10.8	7.4	7.5	5.8	6.8	7.8	8.0	7.7
	NMAE (%)	21.7%	15.9%	13.3%	12.6%	12.6%	15.7%	16.2%	15.3%
	[O3]avg	49.8	49.9	56.7	45.9	53.6	50.0	49.3	50.8
	sample size	152	150	151	151	152	151	152	

Table 5.26 Statistics of the 2004 Model Forecasts for the Ozone Forecast Models

	Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
NLR-fuzzy	Bias (ppb)	6.2	1.5	3.7	1.1	0.4	2.2	2.1	2.5
	MAE (ppb)	8.9	7.6	8.3	6.4	7.7	8.0	7.7	7.8
	NMAE (%)	17.9%	15.1%	15.6%	15.0%	15.3%	15.9%	15.6%	15.3%
Basic-fuzzy	Bias (ppb)	7.0	1.2	5.1	1.4	0.4	3.1	3.1	2.9
	MAE (ppb)	9.6	7.7	8.6	6.5	7.6	8.1	8.2	8.0
	NMAE (%)	19.3%	15.4%	15.1%	15.1%	15.2%	16.2%	16.6%	15.9%
NLR	Bias (ppb)	7.0	1.3	3.7	1.5	0.3	3.0	2.0	2.7
	MAE (ppb)	9.3	7.8	8.6	6.5	8.2	8.5	8.1	8.1
	NMAE (%)	18.7%	15.7%	15.1%	15.2%	15.3%	17.0%	16.4%	16.0%
	[O3]avg	49.8	49.9	56.7	45.9	53.6	50.0	49.3	50.8
	sample size	152	150	151	151	152	151	152	

For both the model hindcasts and forecasts, the NLR-fuzzy models had equivalent or slightly better performance statistics than those of the NLR models and basic-fuzzy models. For the model forecasts, the average MAE for the seven metro areas was 7.8 ppb.

This statistic was comparable to the average MAEs of the basic-fuzzy model (8.0 ppb) and NLR model (8.1 ppb). The 2004 ozone season in Kentucky was significantly cooler and wetter than usual (as was the case for most of the eastern U.S.), with the result that there were very few NAAQS exceedences. Therefore, there was insufficient data for reliable comparison of the DR and FAR statistics. The sample time series plot for Louisville, June 2004 (Figure 5.17) showed the model hindcasts for the three types of ozone forecast models tracked the day-to-day ozone variation reasonably well.

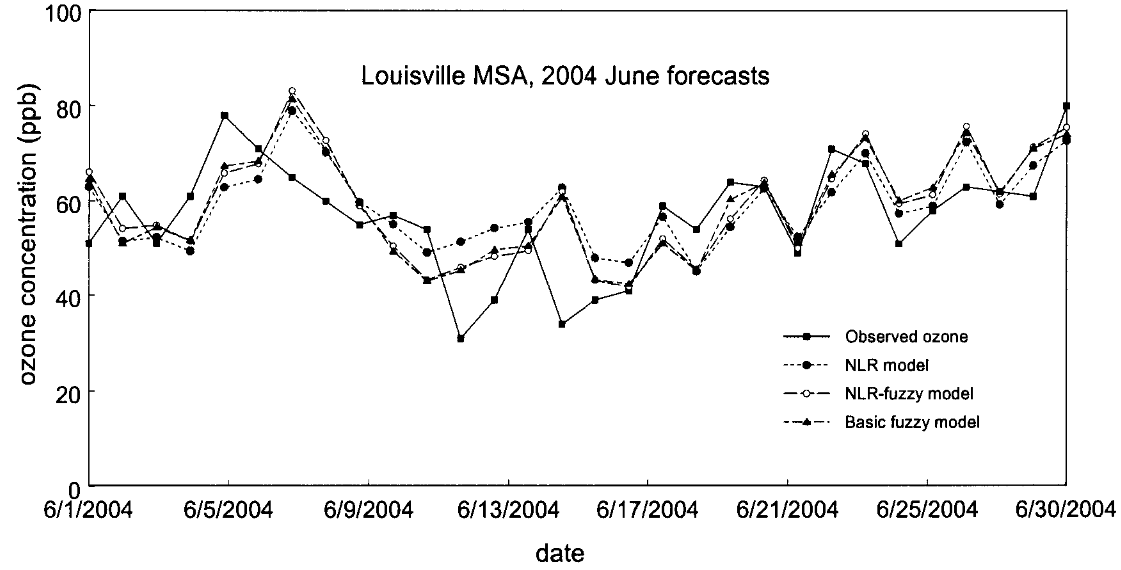


Figure 5.17 Time series of observed 8-hr ozone concentrations in Louisville and forecasts from the NLR, basic-fuzzy and NLR-fuzzy models during June of 2005.

CHAPTER VI

$PM_{2.5}$ FORECAST MODELS

$PM_{2.5}$ concentrations are correlated with ground-level O_3 concentrations during the summer ozone season. One reason for this is that like O_3, much of the summertime $PM_{2.5}$ is photochemically generated, and the NO_x and VOCs are common precursors for $PM_{2.5}$ and ground-level O_3. The formation of both the secondary $PM_{2.5}$ and ground-level ozone are significantly affected by weather conditions. The time series plot (Figure 6.1) using the data for Louisville 2001 summer season demonstrated the similarity of the variation between 24-hr average $PM_{2.5}$ concentrations and the daily maximum 8-hr O_3 concentrations.

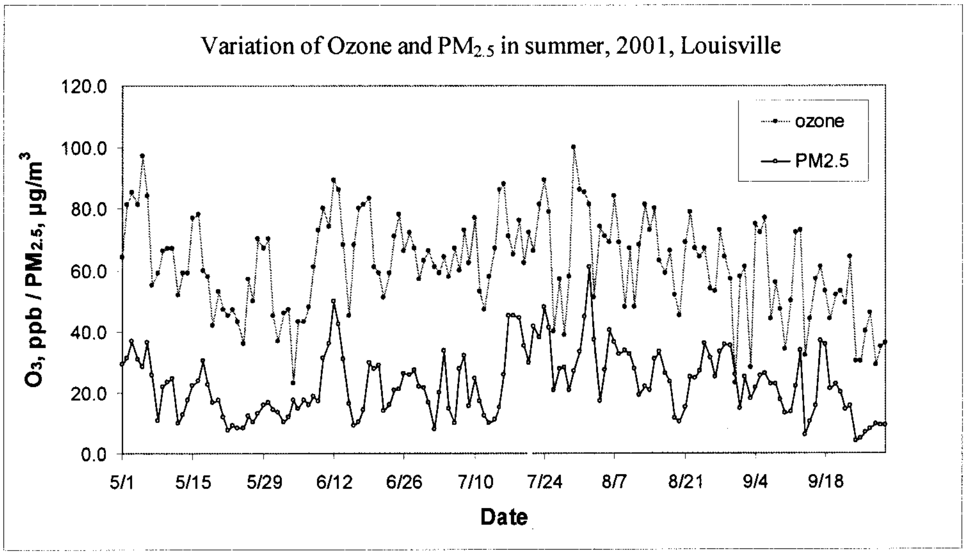


Figure 6.1 Variation of ozone and $PM_{2.5}$ in summer 2001 (Louisville)

In this study, the summer $PM_{2.5}$ forecast models for seven selected metro areas in Kentucky (Ashland, Bowling Green, Covington, Lexington, Louisville, Owensboro, and Paducah) were developed based on the databases for the ozone forecast models. The same candidate prediction parameters as used for ozone (Table 5.2) were correlated with the summertime $PM_{2.5}$ concentrations. The parameters that were statistically significant in the regression processes were used in the $PM_{2.5}$ forecast models. Also, exploratory research was done to find new prediction parameters for $PM_{2.5}$ forecasting.

High $PM_{2.5}$ concentrations were mostly observed in the summertime. In the other seasons, lower temperature and less solar radiation reduced the photo-chemical reactions that form the secondary $PM_{2.5}$. However, $PM_{2.5}$ concentrations were higher in the wintertime than in springtime. This was probably due to increases of the primary $PM_{2.5}$. The use of fossil fuel, including gas, oil, and coal, in home heating ovens and industrial boilers in winter increased the primary $PM_{2.5}$ emissions. Also, the mixing heights tend to be lower in the winter, thus reducing the dilution of emitted particles. In this study, exploratory research was conducted to find the relationship between the winter $PM_{2.5}$ concentrations and the available meteorological parameters and other derived prediction parameters. The winter $PM_{2.5}$ forecast models were developed for seven selected metro areas in Kentucky.

6.1 Preliminary Data Analysis

The $PM_{2.5}$ concentrations have a similar seasonal pattern for each metro area: In the summer season, especially in the warmest period June through August, the $PM_{2.5}$ concentrations were significantly higher than those in the other months, indicating the important influence of high temperature and photochemistry on secondary $PM_{2.5}$ formation. In winter, the $PM_{2.5}$ concentrations reached a secondary, much smaller peak, in January or February, primarily due to the greater fuel use for heating. For example, for the Louisville $PM_{2.5}$ data in the period 1999-2003, the monthly average $PM_{2.5}$ concentrations were high in June and August (21.9 and 24.1 $\mu g/m^3$ respectively) and peaked at 27.4 $\mu g/m^3$ in July. In Louisville, the 5-year monthly average $PM_{2.5}$ concentration was lowest at 14.3 $\mu g/m^3$ in April (Figure 6.2).

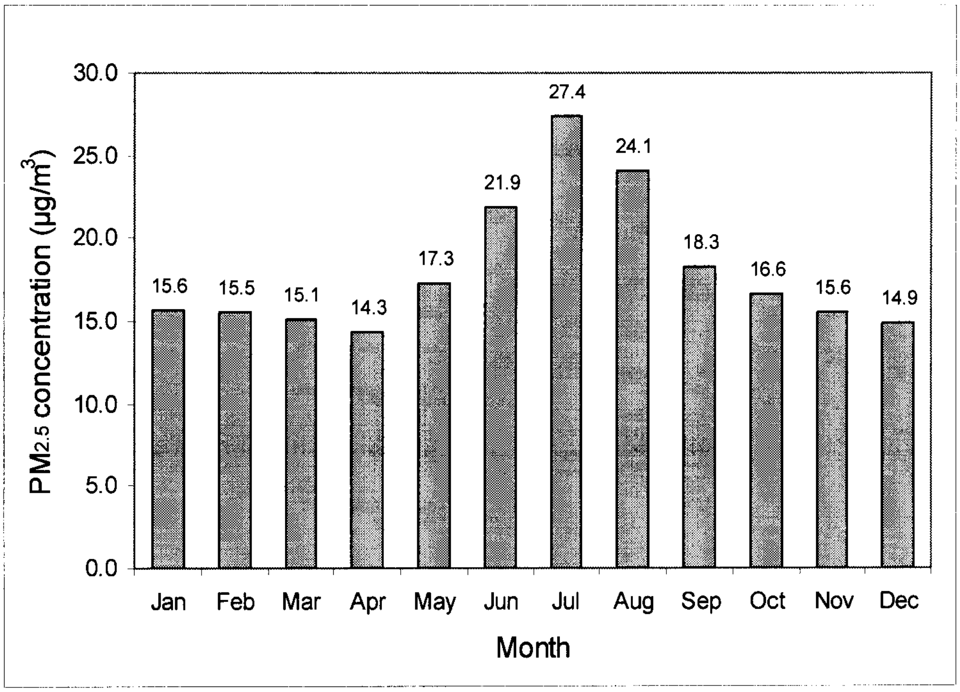


Figure 6.2 Monthly average $PM_{2.5}$ concentrations for Louisville. (Data: 1999-2003)

The $PM_{2.5}$ concentrations in the unhealthy for sensitive groups category (>40 $\mu g/m^3$) mostly occurred in summer. For example, for the data of Louisville over the 1999-2003 period, there were 56 $PM_{2.5}$ exceedence days, 44 of which occurred in summer (May to September), and 39 occurred in June, July, or August. Extremely high $PM_{2.5}$ concentrations occurred on July 4th for each of the five years (Figure 6.3). This was undoubtedly due to the use of fireworks on that day.

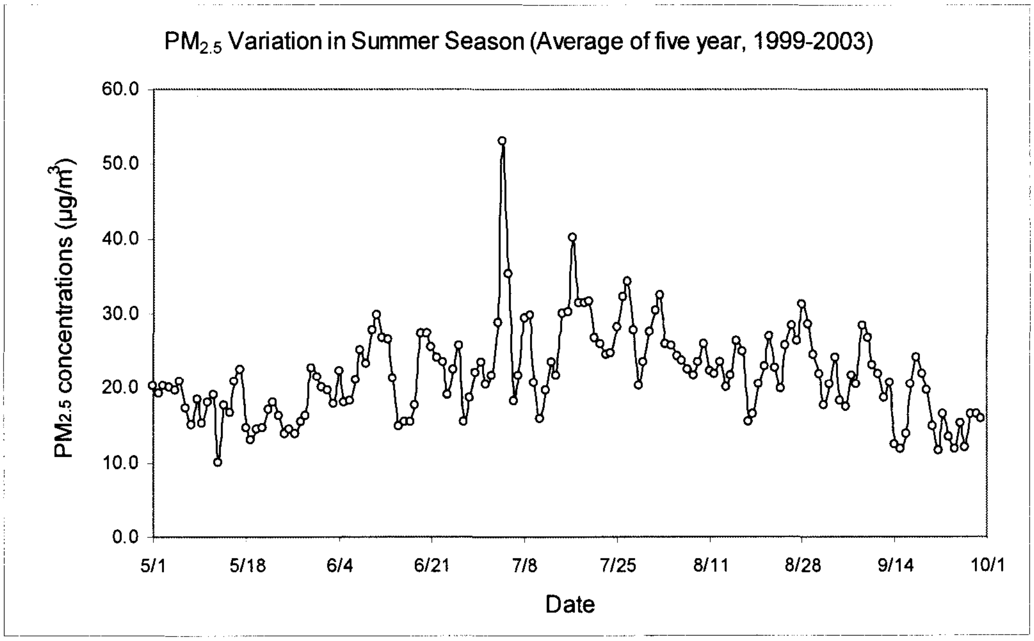


Figure 6.3 Variation of daily $PM_{2.5}$ concentration in summer season, Louisville

As discussed in Chapter II, the national and regional $PM_{2.5}$ concentrations have declined significantly in recent years. The $PM_{2.5}$ trend of the seven metro areas in Kentucky was consistent with nationwide and regional trends. The annual average $PM_{2.5}$ concentrations decreased from 1999 to 2003 for all seven metro areas (Figure 6.4). From 1999 to 2003, the Ashland area achieved the greatest decline of 16.3% and the

Owensboro area had the smallest decrease of 2.1%. On average, the annual average $PM_{2.5}$ concentrations decreased 9.9% for the seven metro areas. The 8th maximum $PM_{2.5}$ concentration for each year is a good statistic to represent the upper end of the distributions, because the year-to-year variation is less random in nature than for the highest few values, which are more sensitive to aberrant events (Cobourn and Lin, 2004). For the 8th maximum $PM_{2.5}$ concentrations, the Ashland area decreased 17.3%. All of the other areas exhibited declines during the period, with the exception of the Covington area, in which this statistic increased 3.8% over the period. On average, the 8th maximum $PM_{2.5}$ concentration of the seven metro areas declined by 9.4% (Figure 6.5).

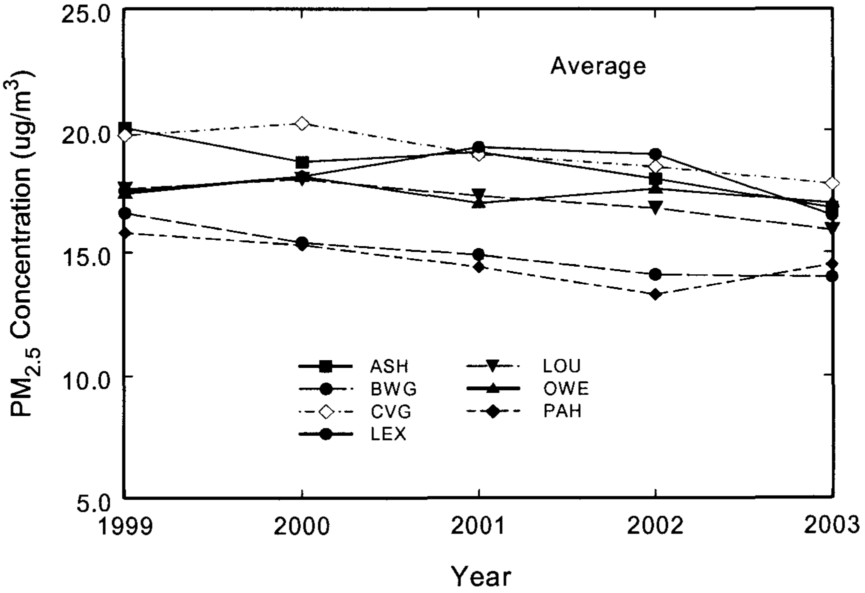


Figure 6.4 Inter-annual patterns of annual average daily $PM_{2.5}$ concentrations for the seven Kentucky metro areas for the period 1999-2003.

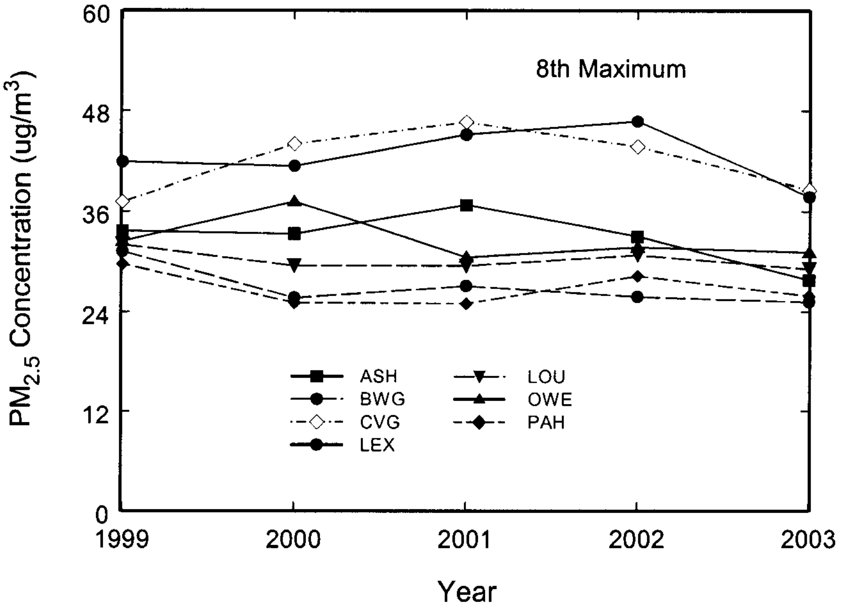


Figure 6.5 Composite inter-annual patterns of annual 8th maximum daily $PM_{2.5}$ concentrations for the seven Kentucky metro areas for the period 1999-2003.

The drop of $PM_{2.5}$ concentrations from 1999 to 2003 was partly due to the fact that 1999 had a hot, dry summertime and 2003 had a cool, wet summertime. Therefore, the scale of the downward trends may better reflect the trends of the $PM_{2.5}$ concentrations in those metro areas. The scale of the downward trends can be ascertained by determining best-fit linear trend lines through the data. For both the annual average and 8^{th} maximum $PM_{2.5}$ data, the trend line slopes were negative for each metro area, except the 8^{th} maximum value for Covington area (Table 6.1). For the annual average $PM_{2.5}$ concentration, the average linear rate of decline for the seven areas was -0.4 $\mu g \cdot m^{-3} \cdot yr^{-1}$. The magnitude of the annual average $PM_{2.5}$ concentration decline rate was less than 8^{th} maximum $PM_{2.5}$ concentration decline rate (-0.6 $\mu g \cdot m^{-3} \cdot yr^{-1}$). This is because these trends were generated from a larger data set (~765) than that of the 8^{th}

maximum trend statistics (~40). Note also that the statistical uncertainty of the slope of the annual average trendline, expressed in terms of standard error, was much less than the corresponding uncertainty for the 8th maximum (Table 6.1).

Table 6.1 Estimated Trend Line Slopes in $PM_{2.5}$ Concentration ($\mu g.m^3.yr^{-1}$, 1999-2003)

	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
8th maximum	-1.2	-1.2	0.2	-0.5	-0.3	-0.8	-0.4	-0.6
Std error	1.0	0.6	1.4	0.4	1.3	0.9	0.7	
Annual mean	-0.7	-0.6	-0.6	-0.5	-0.1	-0.1	-0.5	-0.4
Std error	0.2	0.1	0.1	0.1	0.4	0.1	0.2	

The trends of PM2.5 concentrations in summertime were also investigated by comparing the seasonal average and 8th maximum $PM_{2.5}$ concentrations in the summer of 1999 through 2003 (Figure 6.6 and Figure 6.7).

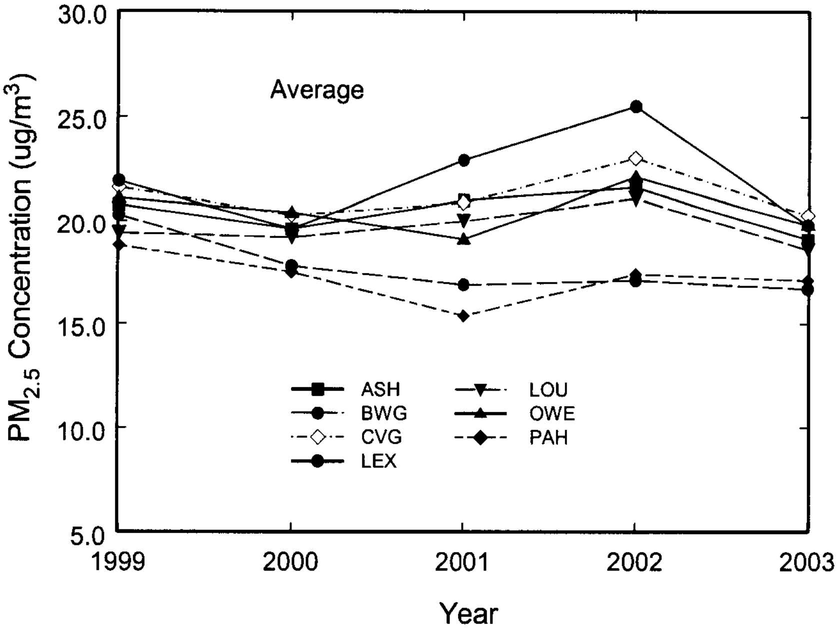


Figure 6.6 Inter-annual patterns of summer average daily $PM_{2.5}$ concentrations for the seven Kentucky metro areas for the period 1999-2003.

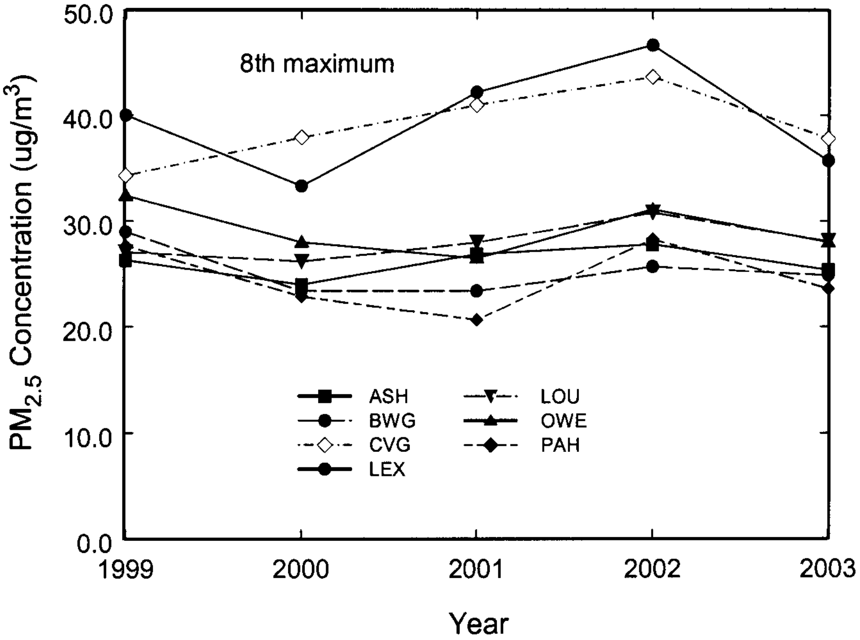


Figure 6.7 Composite inter-annual patterns of summer 8th maximum daily $PM_{2.5}$ concentrations for the seven Kentucky metro areas for the period 1999-2003.

High $PM_{2.5}$ concentrations in 2002 summer, due to the high temperature in that season, affected the linear trend lines for some of the areas. Though most of the areas both seasonal average and 8th maximum $PM_{2.5}$ concentrations for 2002 summer season were lower than those of 1999 summer season, the corresponding $PM_{2.5}$ trend lines may be upward. Counting the seasonal average trend lines, four of the seven lines are downward and three lines are upward. For the 8th maximum trend lines, three lines are downward and four lines are downward. The average linear rate of decline for the seven was -0.3 for seasonal average and 0.1 for 8th maximum $PM_{2.5}$ concentrations (Table 6.2). The summer $PM_{2.5}$ is affected by meteorology. Therefore, based on the linear fits, there is really no clear trend for the summer $PM_{2.5}$ concentrations for those areas.

Table 6.2 Estimated Trend Line Slopes for Summer $PM_{2.5}$ Concentration (μg.m^{3}.yr^{-1} 1999-2003)

	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
8th maximum	0.2	-0.3	0.3	0.6	0	-0.2	-0.1	0.1
Std error	0.6	0.4	0.2	0.4	0.2	0.3	0.3	
Seasonal mean	-0.3	-0.9	0	0.1	0.1	-0.2	-0.6	-0.3
Std error	0.8	0.3	0.8	1	0.4	0.8	0.7	

6.2 $PM_{2.5}$ predictors

The parameters for ozone forecast models listed in Table 4.4 were also used as candidate input variables for $PM_{2.5}$ forecast models. In addition, a new parameter was developed for $PM_{2.6.}$ The $PM_{2.5}$ concentrations are affected by the meteorological conditions, because atmospheric photochemical reactions also form $PM_{2.5}$ in the summertime. Daily maximum temperature is related to secondary $PM_{2.5}$ formation through the temperature-dependent homogeneous and heterogeneous reaction rates, which convert some common gaseous pollutants into very small particles. A second-order polynomial best represents the relationship between maximum temperature and $PM_{2.5}$ concentrations (Figure 6.8 a). The relative humidity is negatively correlated with PM2.5 concentrations with a second-order polynomial (Figure 6.8 b). Daily rain reduces $PM_{2.5}$ levels by directly scavenging particulate matter and its precursors. A straight line can represent the relationship between daily rain and $PM_{2.5}$ concentrations (Figure 6.8 c). The mid-day wind speed dilutes the concentrations of $PM_{2.5}$ and its precursors. This phenomenon could be theoretically explained by the Gaussian plume diffusion model and fixed-box model (Equation 5.1 and 5.2). The exponential function provides a good fit to the wind speed and $PM_{2.5}$ concentration data (Figure 6.8 d).

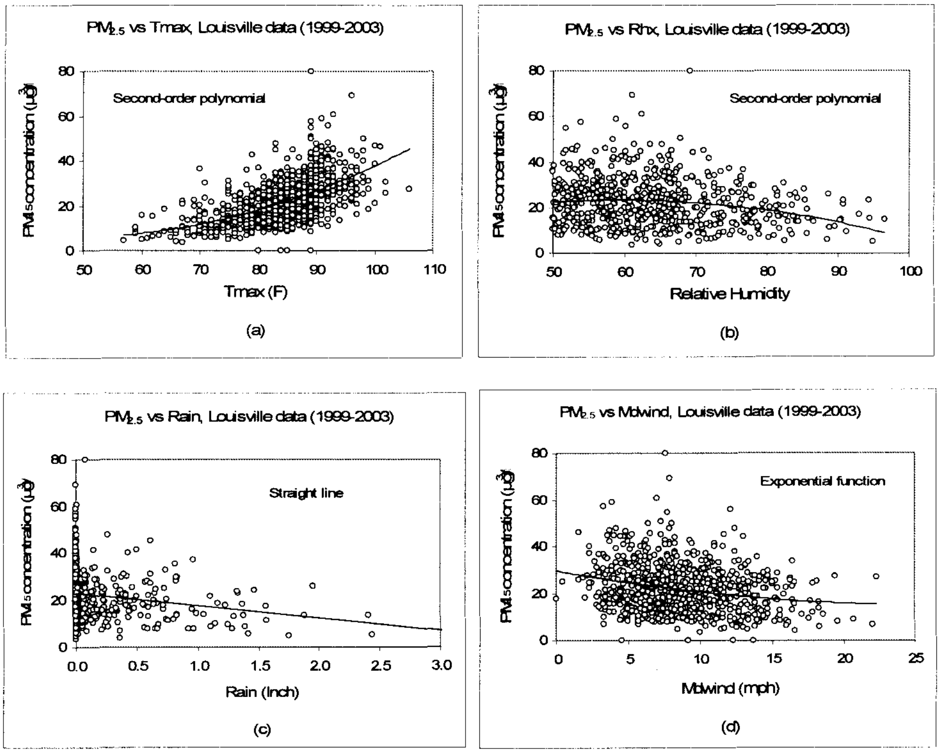


Figure 6.8 Scatter plots of $PM_{2.5}$ vs. Tmax, RH, Rain, and WS (summer).

The trend parameter was included in the $PM_{2.5}$ forecast model, based on the fact that the observed $PM_{2.5}$ concentrations have declined gradually in the past decade in each of the seven metropolitan areas. $PM_{2.5}$ and its precursors, particularly NO_x, can be transported over distances of several hundred kilometers or more. For the Louisville and Lexington models, the two trajectory-based parameters “OZ48” and “traj” used in the ozone models were added to account for the transport of $PM_{2.5}$ and its precursors. These parameters were not included in the other models due to logistical and staffing limitations.

The clear sky atmospheric transmittance at noontime accounts for the solar radiation which drives the photochemical ozone formation. The minimum temperature departure may serve as a day-to-day modifier of the seasonal effect. Cloud cover directly reduces solar radiation. The parameter Saturday, Friday, and Holiday account for the aberrant events of $PM_{2.5}$ emissions in special days. The special events occurred in holiday, such as using fireworks, could significantly increase $PM_{2.5}$ concentrations. In Saturday and Friday, reduction of traffic and manufacturing could reduce the emission of VOC and NO_x, which are the precursors of secondary $PM_{2.6.}$

A wind rose diagram displays the frequency (percentage) of wind directions for a specific location over a specified period of time (Figure 6.9). In this study, a binary parameter “wind rose” was developed based on the daily resultant wind direction (also referred to wind sector), which relate to the short distance transportation of the particulate matter from local sources.

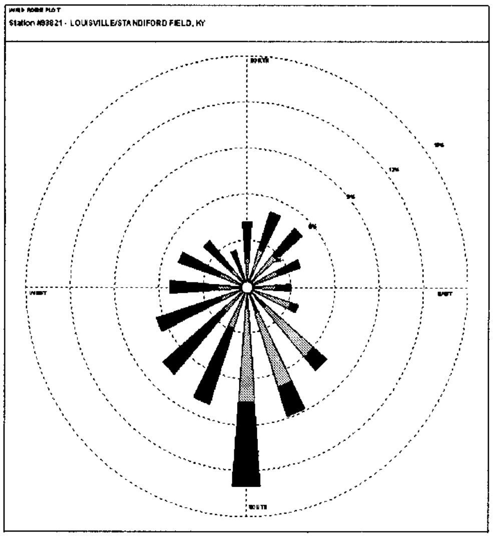


Figure 6.9 Wind rose diagram for June 1 – June 31, 2002, Louisville (NRCS, 2006).

The Louisville $PM_{2.5}$ concentrations and meteorological data during 1999-2003 winter season (November through March) were studied to develop the parameter wind rose. First, a NLR $PM_{2.5}$ forecast model using aforementioned parameters was developed and was used to estimate the $PM_{2.5}$ concentrations for a calibration data set. The model estimates were compared with the observed $PM_{2.5}$ concentrations. The days with $PM_{2.5}$ concentrations that were highly under-estimated (over 10 $\mu g/m^3$) and highly over-estimated (over 8 $\mu g/m^3$) by the model were selected for investigation. To determine the value of this parameter wind rose, we found the wind direction sectors associated with under-estimated $PM_{2.5}$ concentrations and the sectors associated with over-estimated $PM_{2.5}$ concentrations, based on the air quality data and wind rose data. During the investigating period, there were 39 under-estimated days and 30 over-estimated days. It was found that daily wind directions in the 190°-240° sector were associated with 41% of the under-estimated days (16 of 39); daily wind directions in the 250°-320° sector were associated with 40% of the over-estimated days (12 of 30). So the sectors 190°-240° and

250°-320° were designated as the under-estimated sector and the over-estimated sector respectively. The daily wind directions were obtained from Local Climatological Data Reports issued by National Climatic Data Center (NCDC, 1999-2004a). The parameter "wind rose" was assigned a value 1.0 if the daily resultant wind direction was in the under-estimated sector, a value -1.0 if the daily resultant wind direction was in the over-estimated sector, and a value 0.0 for other sectors. A case study using $PM_{2.5}$ air quality data and wind rose data in 1999-2003 for Louisville area showed a significant influence of daily wind direction on the $PM_{2.5}$ concentrations. When a linear function was fitted to the data, the straight line had a slope of 3.7 and a determination coefficient R^2 of 0.06. The parameter wind rose was significant in the multiple linear regression for the Louisville $PM_{2.5}$ forecast model, with a coefficient of 1.91 and t-value of 4.73.

6.3 Summer $PM_{2.5}$ Forecast Models

Nonlinear regression (NLR) models were developed to forecast the $PM_{2.5}$ concentrations in the summer season for the metro areas Ashland, Bowling Green, Covington, Lexington, Louisville, Owensboro, and Paducah. The databases used for developing the summer $PM_{2.5}$ forecast models consist of data during the summer ozone season (May to September), over the five year period 1999-2003. For the Louisville and Covington area, the observed $PM_{2.5}$ data were available for each day of each season. The Louisville and Covington databases had 760 days and 669 days of complete data after the data screening process. For the other metro areas, the $PM_{2.5}$ concentrations were reported every three days. The effective number of days for these areas ranged from 215 (Paducah) to 254 (Owensboro), after removing the days with missing air quality data or meteorological data.

The operational $PM_{2.5}$ forecast model for Louisville was a hybrid model, assembled with a similar fitting procedure as the ozone hybrid model, viz. a standard model fitted to the complete database and a Hi-Lo model fitted to the data set with 10% upper and 10% lower $PM_{2.5}$ concentrations. Based on the fact that during the warm season high $PM_{2.5}$ concentrations usually occur on days with high temperature and low surface wind speed, the criteria used for invoking the Hi-Lo model was defined as follows:

- Maximum temperature greater than 90 °F;
- Wind speed less than 7.6 mph;

The thresholds of maximum temperature and wind speed were determined by analyzing the Louisville data during a five year period. These switching criteria were slightly different from the 3S criteria for the ozone forecast models. The parameter cloud cover was not used as a switching criterion for $PM_{2.5}$ models, because a direct relationship between the cloud cover and high $PM_{2.5}$ concentrations was not observed through the Louisville data.

Except for the Louisville model, the operational $PM_{2.5}$ forecast models for the other six metro areas were standard NLR models that were only fitted to the complete databases. The hybrid technique was not applied on those models, because the databases for these areas were smaller and so the number of days for developing the Hi-Lo model (typically 80-90) was insufficient for developing a valid statistical model.

6.3.1 Model development

As with the ozone models, development of each nonlinear regression $PM_{2.5}$ forecast model entailed two separate fitting procedures: one for the nonlinear term, and another for the linear terms. In the first step, a nonlinear term was developed considering the nonlinear behavior of $PM_{2.5}$ with part of the candidate input parameters. It has been shown that maximum temperature, wind speed, and relative humidity are significantly correlated with $PM_{2.5}$ concentrations with nonlinear functions (Figure 6.7). Many forms of the interactive nonlinear function were studied on the each of the seven databases. The combination of a second order polynomial function for maximum temperature, an exponential function for wind speed, and a second order polynomial function for relative humidity worked best. The final form of the nonlinear term was the same for each of the seven models:

$$Nonlin = (a_1 + a_2 T\,max + a_3 T\,max^2)\,exp(a_4 WS)(Rhx3 + a_5 Rhx3^2) \qquad (6.1)$$

The area specific coefficients a_1 to a_5 were determined through a separate fitting process for each metro area (Table 6.3). For the Louisville $PM_{2.5}$ forecast model, the standard model and the Hi-Lo model used the same nonlinear term. The nonlinear terms for the model of Bowling Green, Owensboro, and Paducah excluded the second order term of Tmax because the coefficient a_3 was not statistically significant in the regression process with the interactive nonlinear function.

Table 6.3 Nonlinear Coefficients for Seven $PM_{2.5}$ Models

Coefficient	ASH	BWG	CVG	LEX	LOU	OWE	PAH
a_1	1.35	-1.14	2.67	2.92	1.70	-2.13	-1.48
a_2	-0.04	0.02	-0.08	-0.08	-0.05	0.04	0.03
a_3	0.00046	0.00000	0.00068	0.00065	0.00050	0.00000	0.00000
a_4	-0.0449	-0.0096	-0.0313	-0.0295	-0.0277	-0.0386	-0.0492
a_5	-0.0084	-0.0098	-0.0082	-0.0089	-0.0072	-0.0093	-0.0100

In the second step, the full nonlinear model was assembled by adding candidate linear terms to the nonlinear "sub-model", and in a stepwise regression procedure various regression models for predicting $PM_{2.5}$ concentrations were examined. The stepwise regression method does not examine all combinations, so there is no guarantee of an absolute "best" model. Therefore after the stepwise procedure, physical reasoning and previous model building experience was applied to examine other combinations and determine the final model. In this process, the t-statistic was used to judge the significance of each parameter in the regression, and the correlation coefficient (R^2) was used to evaluate the overall model performance. The final equation for the forecast model, consisted of an intercept and a group of up to ten explanatory variables,

$$\begin{aligned}\text{PM2.5} = {} & b_0 + b_1\text{Nonlin} + b_2\text{Trend} + b_3\text{Dewpt} + b_4\text{Rain} + b_5\text{CC} \\ & + b_6\text{Windrose} + b_7\text{Hol} + b_8\text{Sat} + b_9\text{Tmn\_dep} + b_{10}\text{Traj} + b_{11}\text{OZ48}\end{aligned} \tag{6.2}$$

Some of the parameters were dropped from the models if they were not significant in the linear regression. Therefore, the input variables and their coefficients were unique for each $PM_{2.5}$ forecast model (Table 6.4).

Table 6.4 Model Coefficients for the Seven Metro Area NLR $PM_{2.5}$ Models

Variable	Coef.	ASH	BWG	CVG	LEX	LOU	OWE	PAH
Intercept	b0	-6.13	-0.38	-6.72	-6.50	-3.85	-3.34	1.76
Nonlin	b1	0.63	1.00	0.78	0.71	0.60	0.83	0.92
Trend	b2	0.24	-0.33		0.36	0.42	-0.22	-0.26
Dewpt	b3	0.25		0.16	0.17	0.16	0.11	
Rain	b4	-1.65		-3.56	-2.85	-3.42		
CC	b5	-0.59	0.43	0.35				
Windrose	b6			2.23		1.91	1.18	0.78
Hol	b7			6.03	27.02	23.41	31.45	16.57
Sat	b8			-1.75		-1.53		
Tmn_dep	b9					0.14		
Traj	b10				1.50	4.51		
OZ48	b11				6.99	6.79		

The nonlinear term was significant in each forecast model. It was a strong contributor for each forecast model with a t-value of 3.5 or more. Except the Covington model, the other six forecast models included the parameter Trend. The parameter Trend reflects the $PM_{2.5}$ trend in the summer season after removing the meteorological influence on the $PM_{2.6}$. The trend term had negative coefficients for Bowling Green, Owensboro, and Paducah models and had positive coefficients for the other models. These results were consistent with the summer $PM_{2.5}$ trend study for selected metro areas in Kentucky (Section 6.1). The new parameter Windrose was used in four of the seven models. The holiday parameter coefficient was more than 20.0 for the Lexington, Louisville, and Owensboro models. The parameter Traj and OZ48 was available only for the Lexington and Louisville models. These two parameters were statistically significant in both models.

6.3.2 Model validation on calibration data set and independent data set

The NLR $PM_{2.5}$ forecast models for the seven metro areas were first validated on the calibration data set. The standard regression models had a Bias of zero. For Louisville model, applying hybrid technique slightly increased the Bias to 0.33 (Table 6.5). The MAE for the seven forecast models ranged from 4.84μg/m^3 (for Lexington) to 6.55μg/m^3 (for Owensboro). The MAE was 24.4%-28.9% of the corresponding average $PM_{2.5}$ concentrations for each metro area. On average, the forecast models for the seven metro areas explained about 45% of the variation in $PM_{2.5}$ concentrations, based on the correlation coefficient R^2 of the multiple linear regression. The Louisville and Lexington models had R^2 values of 0.54 and 0.48 respectively. This was higher than those of the other models, probably because these models included the Traj and OZ48 parameters. The performance of the $PM_{2.5}$ forecast models, characterized by MAE and R^2 of the model fit, was generally lower than that of the ozone forecast model. This was true also for each metro area. A probable explanation for this fact is that the $PM_{2.5}$ has a longer atmospheric residence time than ozone. Current meteorological conditions are related to local and regional transport and dispersion of pollutants and pollutant precursors, and also the formation of secondary pollutant. Ambient $PM_{2.5}$, in contrast to ambient ozone, consists of particles that have been airborne for extended times, up to two weeks. The amounts of these aged particles have little to do with recent atmospheric conditions.

Table 6.5 Statistics of the Model Fit for the Seven NLR $PM_{2.5}$ Models (1999-2003)

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
Bias (ug/m^3)	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.05
MAE (ug/m^3)	6.43	4.78	6.41	4.84	6.45	6.55	4.99	6.18
RMSE (ug/m^3)	6.87	6.12	7.19	6.11	7.24	7.39	6.41	6.76
NMAE (%)	26.1%	27.0%	26.5%	24.6%	24.4%	27.1%	28.9%	26.2%
$[PM2.5]_{avg}$	20.4	17.7	21.2	19.7	22.0	20.5	17.3	19.80
R^2	0.42	0.37	0.46	0.48	0.54	0.46	0.40	0.45
Count	237	238	669	245	760	254	215	

The unhealthy limit of NAAQS for daily average $PM_{2.5}$ concentrations is 40 $\mu g/m^3$ (unhealthy for sensitive groups). In the calibration period 1999-2003, there were 23 “unhealthy” days recorded in Louisville. Due to the small databases for the other metro areas, there were few unhealthy days in the calibration period for each area (typically less than 10 days out of 215-245 possible days). To make the critical performance indexes (DR, FAR, and CSI) statistically significant, the forecast skill of the $PM_{2.5}$ forecast model was evaluated with only the Louisville data. When the alarm threshold was set at the NAAQS unhealthy limit $40\mu g/m^3$, the Louisville $PM_{2.5}$ forecast model detected 12 of the 23 unhealthy days. The detection rate was 0.52. This value of DR was comparable to that of the ozone forecast model. However, this $PM_{2.5}$ forecast model issued 32 false alarms, resulting in a high false alarm rate of 0.73 and a low critical success index of 0.22. Using an alarm threshold slightly lower than the NAAQS unhealthy limit would increase the forecast skill of the model. With the alarm threshold of $38\mu g/m^3$, the detection rate was 0.57, the false alarm rate was 0.65, and the critical success index was 0.29.

The sample scatter plot of the model fits versus the observed $PM_{2.5}$ concentrations for Louisville illustrates the correspondence between model fits and observations (Figure 6.10). Scatter plots for the other models had a similar pattern.

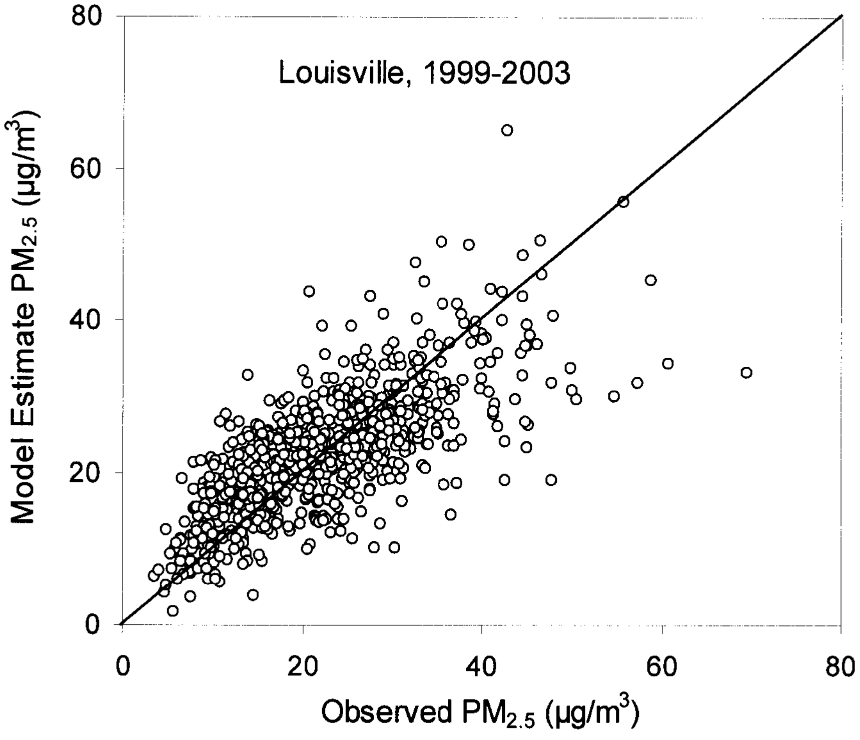


Figure 6.10 Scatter plot of model estimates against observed $PM_{2.5}$ concentrations for the Louisville model. The diagonal indicates the line of perfect agreement.

To test the seven $PM_{2.5}$ forecast models with an independent data set, we used the forecast models for the seven metro areas to predict the daily $PM_{2.5}$ concentrations of the 2004 summer season, using the observed meteorological data as model inputs. The error statistics of the model hindcasts for each forecast model were slightly higher than those of the model fit (Table 6.6). The average MAE of the seven model hindcasts was 6.43$\mu g/m^3$, compared to 6.18$\mu g/m^3$ for the model fits. The average RMSE of the model hindcasts was 6.46$\mu g/m^3$. This value was actually lower than the average RMSE of the model fits (6.76μg/m^3). This is because RMSE gives a relatively high weight to large errors. The average MAE and RMSE of the model hindcasts was 31.7% and 38.4% of the 2004 average $PM_{2.5}$ concentrations of the seven metro areas. It so happened that for 2004 the model hindcast biases for each metro area were positive, ranging from 0.57μg/m^3 (Bowling Green) to 4.48μg/m^3 (Louisville). It is usual for the bias to be non-zero for a test data set, due to year-to-year variability in climate and pollutant transport. The $PM_{2.5}$ concentrations during the 2004 ozone season (16.9μg/m^3) were significantly lower than the average of the previous five year average (19.8μg/m^3). Part of the variations of the $PM_{2.5}$ concentrations is due to factors not explained by the forecast models. Year-to-year variability in these unknown factors can produce either positive or negative bias.

Table 6.6 Statistics of the Model Hindcasts for Seven NLR $PM_{2.5}$ Models (2004)

Statistic	**ASH**	**BWG**	**CVG**	**LEX**	**LOU**	**OWE**	**PAH**	**Average**
Bias (ug/m^3)	2.68	0.57	1.52	2.23	4.48	2.61	1.13	2.17
MAE (ug/m^3)	6.59	6.03	6.40	4.63	6.87	6.48	6.40	6.43
RMSE (ug/m^3)	6.74	6.14	6.86	6.65	7.19	6.37	6.24	6.46
NMAE (%)	30.5%	32.0%	29.1%	27.2%	33.2%	33.2%	37.1%	31.7%
$[PM2.5]_{avg}$	18.3	16.7	18.6	17.0	17.7	16.5	14.3	16.9
R^2	0.51	0.26	0.34	0.51	0.97	0.68	0.43	0.53
Count	51	50	51	51	153	51	44	

The time series plot of model hindcasts and observed $PM_{2.5}$ for Louisville during 2004 was a typical pattern that reflects the performance of the $PM_{2.5}$ forecast model (Figure 6.11). The model hindcasts tracked the $PM_{2.5}$ variation reasonably well.

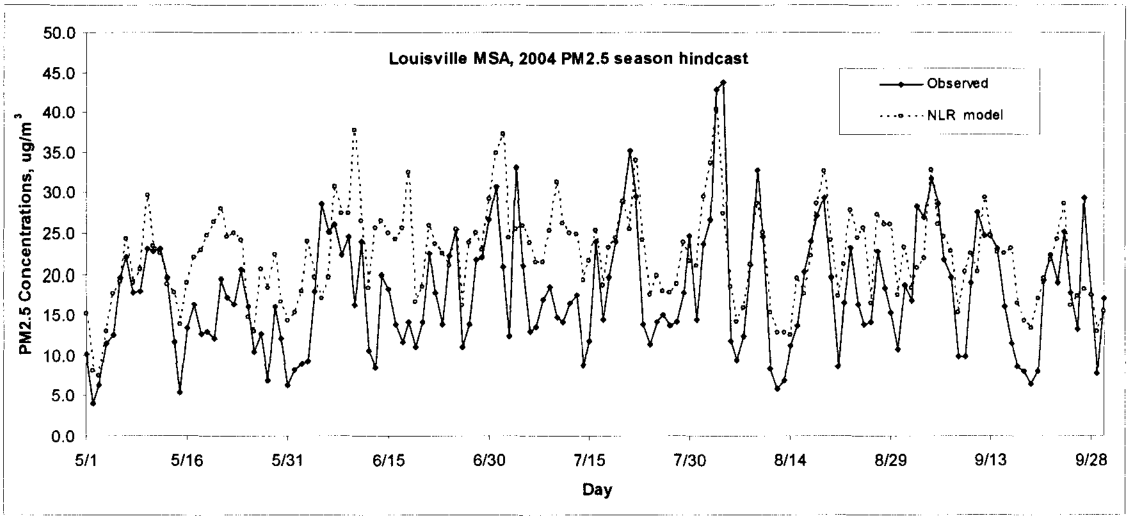


Figure 6.11 Time series of hindcasts during the 2004 summer season, for Louisville

6.4 Winter $PM_{2.5}$ Forecast Models

Nonlinear regression winter $PM_{2.5}$ forecast models were developed for the seven selected metro areas in Kentucky. The databases for winter models consisted of the air quality data and meteorological data during January, February, March, November and December, over the five year period 2000-2004. The $PM_{2.5}$ monitors in Louisville areas provided $PM_{2.5}$ observation data for each day. The Louisville database contained complete data for 748 days. In the Covington area, winter $PM_{2.5}$ data were available for each day during the period 2000-2003, but available for every three days in 2004. So the Covington database had complete data for 576 days. The effective number of days for the other areas ranged from 231 (Paducah) to 254 (Owensboro). For each of the seven areas, the operational winter $PM_{2.5}$ forecast model was the standard model fitted to the complete database. Application of the hybrid model technique did not improve performance for any of the winter models, so only the standard NLR models were developed.

6.4.1 Model development

The model building approach for the winter models was same as for the summer models. In the first step, a nonlinear term was developed considering the nonlinear function relating $PM_{2.5}$ to maximum temperature, wind speed, and relative humidity. The final form of the interactive nonlinear function that best fitted the winter $PM_{2.5}$ data was slightly different from that for the summer $PM_{2.5}$ model. It was the combination of a third order polynomial function for maximum temperature, and power functions for both wind speed and relative humidity:

$$Nonlin = (a_1 + a_2 T\,max + a_3 T\,max^2 + a_4 T\,max^3)(WS^{a_5})(RH^{a_6}) \tag{6.3}$$

The area specific coefficients a_1 to a_5 were determined through separate fitting process for each metro area (Table 6.7).

Table 6.7 Nonlinear Coefficients for Seven Winter $PM_{2.5}$ Models

Coefficient	ASH	BWG	CVG	LEX	LOU	OWE	PAH
a_1	0.937	-6.641	4.734	7.793	-1.185	-0.443	-7.369
a_2	2.876	0.968	0.683	1.954	0.991	0.550	1.206
a_3	-0.072	-0.021	-0.172	-0.055	-0.024	-0.014	-0.027
a_4	0.00058	0.00015	0.00015	0.00046	0.00018	0.00010	0.00019
a_5	-0.363	-0.160	-0.327	-0.215	-0.278	-0.276	-0.260
a_6	-0.083	0.201	0.218	0.017	0.214	0.371	0.188

In the second step, a multiple linear regression was fitted to include the linear term in the model. Both the stepwise regression method and physical reasoning based on previous model developing experience were applied to determine the parameters used in the models. In the final $PM_{2.5}$ model, each of the predictor variables selected for the multiple linear regression was statistically significant, with a t-statistic greater than or close to 2.0 in absolute value. The final equation for the forecast model, consisted of an intercept and a group of up to eleven explanatory variables,

$$\begin{aligned} PM2.5 = b_0 + b_1 Nonlin + b_2 Trend + b_3 T\,min + b_4 Dewpt + b_5 Rain \\ + b_6 Xmitt + b_7 Tmn\_dep + b_8 CC + b_9 Windrose + b_{10} Rhx + b_{11} Fri \end{aligned} \qquad (6.4)$$

For each model, terms that were not statistically significant were dropped. The input variables and their coefficients were unique for each $PM_{2.5}$ forecast model (Table 6.8). As expectation, the nonlinear term was the strongest contributor for each model with the t-value ranged from 4.3 (Covington) to 14.8 (Louisville). The trend term was statistically significant in each linear regression, with negative coefficients ranging from

-0.68 ppb/yr (Paducah) to -1.22 ppb/yr (Covington). This indicated that the winter $PM_{2.5}$ concentrations in those areas gradually declined during the period 1999-2003. The parameter Rain was included in each model, showing that precipitation tended to reduce the $PM_{2.5}$ concentrations in cold weather conditions. Minimum temperature and dew point were important winter $PM_{2.5}$ predictors for each forecast model.

Table 6.8 Model Coefficients for the Winter $PM_{2.5}$ Models

Variable	Coef.	ASH	BWG	CVG	LEX	LOU	OWE	PAH
Intercept	b0	9.09	-7.17	18.09	6.12	6.54	-0.59	-8.94
Nonlin	b1	0.89	1.11	0.60	0.98	0.93	0.84	0.97
Trend	b3	-0.89	-0.72	-1.22	-1.10	-0.86	-0.83	-0.68
Tmin	b2	-0.47	-0.33	-0.57	-0.36	-0.60	-0.56	-0.41
Dewpt	b4	0.41	0.27	0.45	0.34	0.42	0.30	0.24
Rain	b5	-2.38	-4.27	-6.07	-4.11	-4.21	-3.59	-6.58
Xmitt	b6		16.37			11.46	22.96	28.74
Tmn_dep	b7			0.16		0.17	0.25	0.16
CC	b8			0.29	0.26	0.16		
Windrose	b9			2.31	1.93	1.56	1.67	
RHx	b10	-0.05		-0.06	-0.07	-0.07		
Fri	b11			1.33		0.93		

6.4.2 Model validation on the calibration data set

Performance of the final $PM_{2.5}$ forecast models on calibration data set was evaluated by comparing the model estimates with the observed ozone concentration within the calibration periods. For each of the $PM_{2.5}$ models, the bias of the model estimates was about zero. The MAE for the seven forecast models ranged from 3.23 $\mu g/m^3$ (Paducah) to 4.61 $\mu g/m^3$ (Covington). The MAE was 26.7%-27.9% of the corresponding average $PM_{2.5}$ concentrations for each metro area. The correlation coefficients R^2 for the $PM_{2.5}$ models ranged from 0.31 (Bowling Green and Lexington) to 0.46 (Ashland). On average, the winter $PM_{2.5}$ forecast models for the seven metro areas explained about 37% of the variation in $PM_{2.5}$ concentrations, based on the correlation coefficient of the multiple linear regression.

The MAEs of the winter $PM_{2.5}$ models were lower than those of the summer models, due to the seasonal average $PM_{2.5}$ concentrations in winter were much less than the seasonal average $PM_{2.5}$ concentrations in summer. For example, the overall average MAE of the model estimates was 3.90 $\mu g/m^3$ for the seven winter models and was 6.18 $\mu g/m^3$ for the summer models. The overall average observed $PM_{2.5}$ concentrations for the seven areas was 14.5 $\mu g/m^3$ for the winter period and 19.8 $\mu g/m^3$ for the summer period. Generally, the performance of the winter $PM_{2.5}$ forecast models were inferior to that of the summer forecast models. Comparing to the summer models, the winter models had lower correlation coefficients (0.37 vs. 0.45 on overall average) and higher value of NMAE (26.8% vs. 26.2% on overall average). A probable explanation for this fact is that the primary $PM_{2.5}$ pollutants were dominant in the winter time. Meteorological parameters mostly influence secondary $PM_{2.5}$ that is formed by photochemical reactions.

Table 6.9 Statistics of the Model Fit for the NLR Winter $PM_{2.5}$ Models (2000-2004)

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
Bias (ug/m^3)	-0.005	0.001	0.001	-0.003	0.002	-0.001	-0.002	-0.001
MAE (ug/m^3)	3.85	3.40	4.61	4.14	4.08	3.99	3.23	3.90
RMSE (ug/m^3)	4.99	4.56	6.06	6.54	6.42	6.73	4.32	6.33
NMAE (%)	26.7%	27.1%	26.9%	27.9%	26.7%	26.5%	26.9%	26.8%
$[PM2.5]_{avg}$	16.0	12.6	17.1	14.8	16.4	16.0	12.0	14.5
R^2	0.46	0.31	0.40	0.31	0.40	0.32	0.36	0.37
Count	253	244	576	246	748	254	231	

The sample scatter plot of the model fits versus the observed $PM_{2.5}$ concentrations for Louisville illustrates the correspondence between model fits and observations (Figure 6.12).

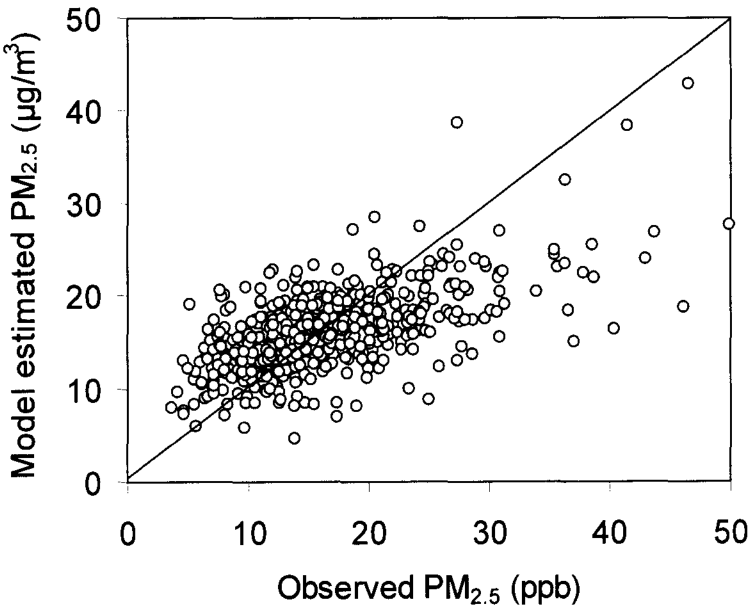


Figure 6.12 Scatter plot of model estimates against observed $PM_{2.5}$ concentrations for Louisville model (Louisville, 2000-2004). The diagonal indicates the perfect correspondence line.

The scatter plot of the residuals versus model estimated $PM_{2.5}$ demonstrated that the residuals have constant variance over the range of predicted $PM_{2.5}$ concentrations (Figure 6.13). Scatter plots for the other models a the similar pattern.

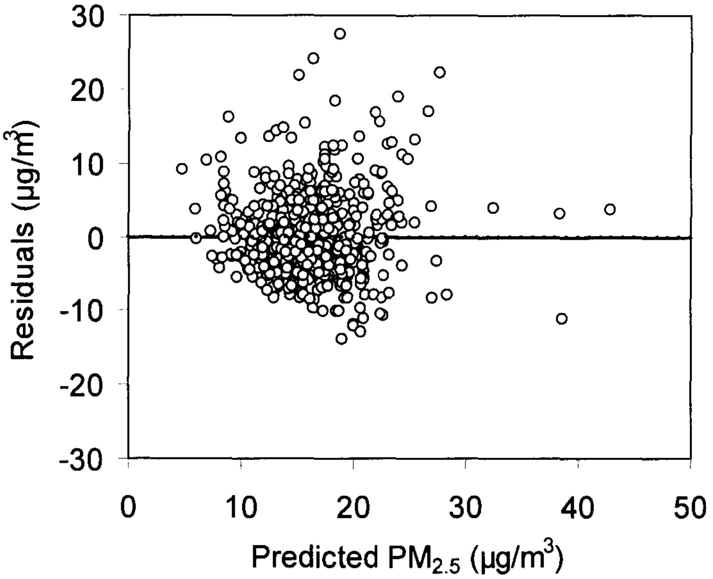


Figure 6.13 Residuals of the model estimates (winter model) versus the predicted $PM_{2.5}$ concentrations (Louisville, November-March, 2000-2004).

An example of time series plots of observed $PM_{2.5}$ concentrations versus predicted $PM_{2.5}$ concentrations for Louisville demonstrated the pattern of the winter $PM_{2.5}$ forecast model to track the variation of $PM_{2.5}$ concentrations (Figure 6.14). Most of the predicted $PM_{2.5}$ concentrations were close to the corresponding observed values (within 3.0μg/m^3). On a few days there were comparatively large errors, including over-predictings (February 4 and 16) and under-predictings (February 1, 6, 14, and 27).

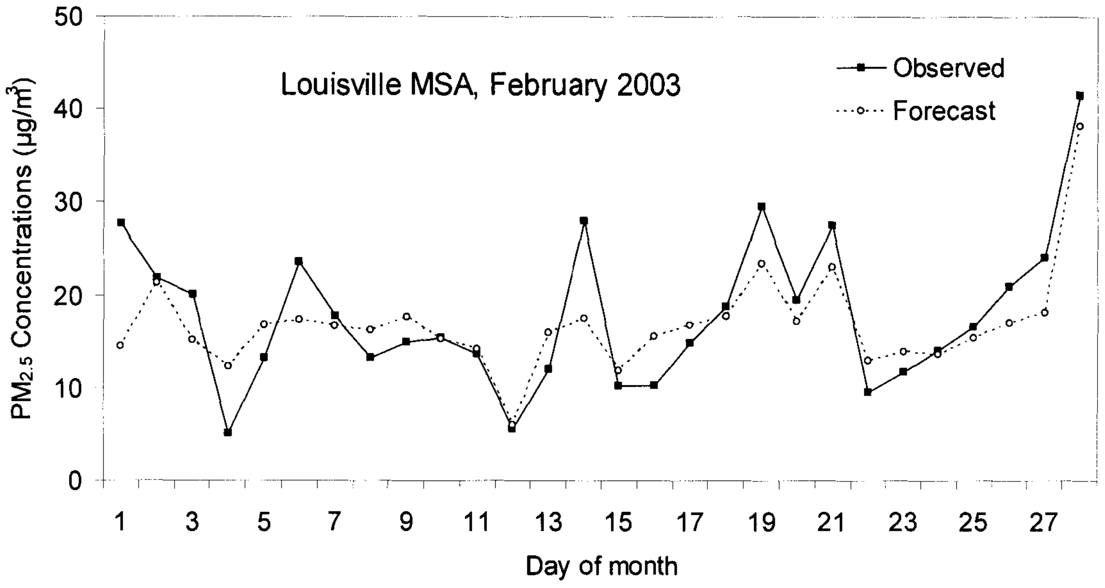


Figure 6.14 Time series of model estimates during the February 2004, Louisville

Due to the weather conditions in winter time, there were few days that exceeded the unhealthy limit of NAAQS (40 µg/m^3) during the winter calibration period 2000-2004 for each area. For example, there were 7 days out of 748 days for Louisville and 6 days out of 576 days for Covington. The indexes the DR, FAR, and CSI were not statistically significant. Therefore these model performance metrics were not applied to evaluate the winter $PM_{2.5}$ forecast models.

6.5 Fuzzy System $PM_{2.5}$ Summer Forecast Models

6.5.1 Model development

With the goal of developing more accurate $PM_{2.5}$ forecast models, NLR-Fuzzy system models were developed for the seven metro areas: Ashland, Owensboro, Bowling Green, Covington, Lexington, Louisville, and Paducah. The fuzzy system models were Takagi-Sugeno fuzzy models as described in Chapter IV, 6.4. The nonlinear term defined by Equation 6.1 was included in the fuzzy system models. Development of the NLR $PM_{2.5}$ forecast models prepared a group of input variables that were statistically correlated with local $PM_{2.5}$ at the 95% confidence level (Table 6.10). These input variables were also used in the NLR-Fuzzy fuzzy system models.

Table 6.10 Input Variables for Seven Metro Area NLR-fuzzy Summer $PM_{2.5}$ Models

Variables	ASH	BWG	CVG	LEX	LOU	OWE	PAH
x_1	Nonlin	Nonlin	Nonlin	Nonlin	Nonlin	Nonlin	Nonlin
x_2	Trend	Trend	Dewpt	Trend	Trend	Trend	Trend
x_3	Dewpt	CC	Rain	Dewpt	Dewpt	Dewpt	Windrose
x_4	Rain		CC	Rain	Rain	Windrose	Hol
x_5	CC		Windrose	Hol	Windrose	Hol	
x_6			Hol	Traj	Hol		
x_7			Sat	OZ48	Sat		
x_8					Tmn_dep		
x_9					Traj		
x_{10}					OZ48		
Count	5	3	7	7	10	5	4

The training data pairs (M) for the seven metro areas ranged from 215 (Paducah) to 760 (Louisville). Selection of the number of rules (R) and the fuzziness factor (m) was referred to the "R and m study" for the ozone fuzzy system models, based on the fact that the summer $PM_{2.5}$ concentrations are correlated to O_3 concentrations. The value of 5 for both R and m was used for the seven fuzzy system models. The error tolerance ε_c used for finding the cluster center v^i was chosen as 0.01. In summary, the pre-designed parameters were the same for each of the seven models, as follow,

- $R = 5$, number of rules
- $m = 5$, fuzziness factor
- $\varepsilon_c = 0.01$, error tolerance

The Takagi-Sugeno fuzzy models were developed with fuzzy clustering accompanied with the weighted least square method, as described in Chapter V. First, the initial cluster centers v_0^i were chosen so that the initial cluster centers were evenly distributed within the data range. As an example, the following table shows the initial cluster centers v_0^i for the Ashland model (Table 6.11).

Table 6.11 Initial Cluster Centers v_0^i for NLR-fuzzy Summer $PM_{2.5}$ Model (Ashland)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	v_0^1	9.68	16.45	23.22	29.99	36.76
Trend	v_0^2	0.40	1.20	2.00	2.80	3.60
Dewpt	v_0^3	41.96	50.16	58.36	66.56	74.76
Rain	v_0^4	2.21	1.72	1.23	0.74	0.25
CC	v_0^5	8.60	6.80	6.00	3.20	1.40

Equation 4.14 – 4.18 were applied to realize the iterative process for finding cluster centers v^i, and Equation 4.19 – 4.22 were applied to find the coefficients a_i of the fuzzy consequence. The final fuzzy system models were characterized by a group of cluster

centers v^i and coefficients a_i for each of the rules. The cluster centers and coefficients were unique for each model. As an example, the parameters for the Ashland NLR-fuzzy system model are listed in the following Table 6.12 and Table 6.13. Model parameters for other metro areas refer to Appendix D.

Table 6.12 Final Cluster centers v_f^i for NLR-fuzzy Summer $PM_{2.5}$ Model (Ashland)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	V_1^i	12.26	16.54	23.90	17.14	28.00
Trend	V_2^i	1.63	1.77	1.81	2.18	2.56
Dewpt	V_3^i	40.98	53.12	61.61	69.05	72.95
Rain	V_4^i	0.01	0.03	0.03	0.42	0.07
CC	V_5^i	1.72	1.76	1.26	6.89	2.13

Table 6.13 Coefficients a_i for NLR-fuzzy Summer $PM_{2.5}$ Model (Ashland)

Variables	Coef.	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-1.78	-14.46	2.86	-4.98	-30.38
Nonlin	a_1	0.68	0.30	0.59	0.21	0.84
Trend	a_2	0.57	-0.24	-0.41	-0.20	-0.13
Dewpt	a_3	0.06	0.50	0.14	0.35	0.61
Rain	a_4	-1.33	-6.51	-4.36	-1.78	0.98
CC	a_5	0.27	-0.07	-0.59	-0.62	-2.66

6.5.2 Model evaluation

With the parameters determined above, the fuzzy system model output can be computed using Equation 5.19 – 5.22. The flow chart in Figure 5.13 illustrates the algorithm of a Takagi-Sugeno fuzzy system. The NLR-Fuzzy system models were first evaluated on the calibration data set. For the model fit statistics (Table 6.14), the Bias of most model fits were negative but close to zero. The MAE of the model fit ranged from 4.58μg/m^3 (Lexington) to 6.41μg/m^3 (Owensboro). The RMSE ranged from 6.84μg/m^3 (Lexington) to 7.24μg/m^3 (Owensboro). The Louisville and Lexington models had slightly better performance than the other models. For example, the NMAEs of the Louisville and Lexington model fits were 23.7% and 23.3% respectively, compared to the average NMAE 26.4%. The R^2 values of 0.55 and 0.52 for the Louisville and Lexington models were slightly above the average of 0.47. This is possibly because the transport parameters Traj and OZ 48 were applied in Louisville and Lexington models.

Table 6.14 Statistics of Model Fit for NLR-fuzzy Summer $PM_{2.5}$ Models

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
Bias (ug/m^3)	-0.38	-0.06	-0.23	-0.24	0.00	-0.12	-0.03	-0.15
MAE (ug/m^3)	6.07	4.70	6.41	4.58	6.32	6.41	4.85	6.02
RMSE (ug/m^3)	6.46	6.07	7.08	6.84	7.05	7.24	6.24	6.57
MAE%	24.9%	26.5%	26.1%	23.3%	23.7%	26.4%	28.1%	26.4%
$[PM2.5]_{avg}$	20.4	17.7	21.2	19.7	22.0	20.5	17.3	19.81
R^2	0.48	0.38	0.47	0.52	0.55	0.47	0.41	0.47
Count	237	238	235	245	760	254	215	

The seven NLR-Fuzzy system $PM_{2.5}$ forecast models were also evaluated on the 2004 observed data set. The model hindcast performance statistics were close to the those of model fits (Table 6.15). The average Bias of the seven model hindcasts was 1.59μg/m^3. Except for the Ashland model and Owensboro model, the MAEs of the other five model hindcasts were slightly higher than those of the model fits. The average MAE of the model hindcasts was 6.19μg/m^3, which was 31.0% of the overall average $PM_{2.5}$

concentrations of the seven metro areas. The average RMSE of the model hindcasts was 6.33μg/m^3. An example of the time series plot of model hindcasts and observed $PM_{2.5}$ for Louisville showed performance of the NLR-Fuzzy $PM_{2.5}$ summer forecast model during May 2004 (Figure 6.15).

Table 6.15 Statistics of the 2004 Model Hindcasts for NLR-fuzzy $PM_{2.5}$ Models

Statistic	ASH	BWG	CVG	LEX	LOU	OWE	PAH	Average
Bias (ug/m^3)	0.43	0.83	1.31	0.89	4.23	1.74	1.71	1.59
MAE (ug/m^3)	4.99	4.96	6.48	4.79	6.85	4.91	6.46	6.19
RMSE (ug/m^3)	6.09	6.13	7.15	6.67	7.05	6.89	6.36	6.33
MAE%	27.2%	31.5%	29.0%	28.1%	33.1%	29.7%	38.1%	31.0%
$[PM2.5]_{avg}$	18.3	16.7	18.6	17.0	17.7	16.5	14.3	
R^2	0.28	0.25	0.47	0.30	0.95	0.55	0.44	
Count	51	50	51	51	153	51	44	

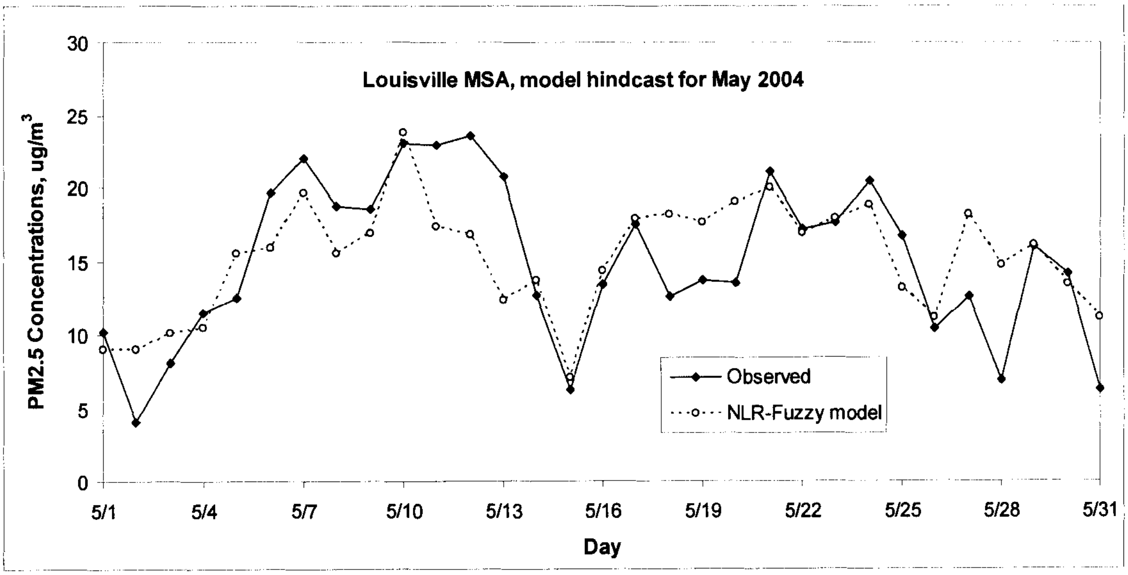


Figure 6.15 Time series of hindcasts during the 2004 summer season, Louisville

6.5.3 Comparison of NLR-Fuzzy system models and NLR models

The NLR-Fuzzy system and NLR $PM_{2.5}$ forecast models were developed for the seven metro areas based on the same databases and using the same input variables. The NLR-Fuzzy system models were compared with the NLR models by comparing the performance of these two type models on both the model fits and model hindcasts.

The model fit statistics of the NLR-Fuzzy system models was slightly better than those of the corresponding NLR forecast models (Table 6.16). For example, the MAE, RMSE, and R^2 for the Lexington NLR model fit were 4.84μg/m^3, 6.11μg/m^3, and 0.48 respectively. By applying fuzzy technique, the MAE and RMSE of NLR-Fuzzy system model decreased to 4.58μg/m^3 and 6.84μg/m^3 respectively. The R^2 was improved to 0.52. The model fit statistics of the NLR-Fuzzy system models achieved improvement for each metro area. On average, the MAE of the seven NLR-Fuzzy system models was 6.02μg/m^3, which was about 3% less than that of the NLR models (6.18μg/m^3). The average RSME and R^2 of the seven NLR-Fuzzy system models were less than those of the NLR models at the same magnitude.

Table 6.16 Comparison of Model Fit Statistics between Two Type $PM_{2.5}$ Models

Statistic	**ASH**	**BWG**	**CVG**	**LEX**	**LOU**	**OWE**	**PAH**	**Average**
NLR PM2.5 forecast models								
Bias (ug/m^3)	0.00	0.00	0.00	0.00	0.33	0.00	0.00	**0.05**
MAE (ug/m^3)	6.43	4.78	6.41	4.84	6.45	6.55	4.99	**6.18**
RMSE (ug/m^3)	6.87	6.12	7.19	6.11	7.24	7.39	6.41	**6.76**
NLR-Fuzzy PM2.5 forecast models								
Bias (ug/m^3)	-0.38	-0.06	-0.23	-0.24	0.00	-0.12	-0.03	**-0.15**
MAE (ug/m^3)	6.07	4.70	6.41	4.58	6.32	6.41	4.85	**6.02**
RMSE (ug/m^3)	6.46	6.07	7.08	6.84	7.05	7.24	6.24	**6.57**

For the model hindcasts on 2004 summer season, the NLR-Fuzzy models also had better performance statistics than those of the NLR models. Except for the hindcasts of the Lexington and Paducah models, the NLR-Fuzzy models had lower MAEs than those of the NLR models. The average MAE of the NLR-fuzzy model hindcasts for the seven metro areas was 6.19μg/m^3, which was ~3% less than that for the NLR model hindcasts. The average Bias and RMSE of the seven NLR-Fuzzy models were also less than the corresponding values of the NLR models (Table 6.17).

Table 6.17 Comparison of 2004 Model Hindcasts between Two Type PM$_{2.5}$ Models

Statistic	**ASH**	**BWG**	**CVG**	**LEX**	**LOU**	**OWE**	**PAH**	**Average**
NLR PM2.5 forecast models								
Bias (ug/m^3)	2.68	0.57	1.52	2.23	4.48	2.61	1.13	**2.17**
MAE (ug/m^3)	6.59	6.03	6.40	4.63	6.87	6.48	6.40	**6.43**
RMSE (ug/m^3)	6.74	6.14	6.86	6.65	7.19	6.37	6.24	**6.46**
NLR-Fuzzy PM2.5 forecast models								
Bias (ug/m^3)	0.43	0.83	1.31	0.89	4.23	1.74	1.71	**1.59**
MAE (ug/m^3)	4.99	4.96	6.48	4.79	6.85	4.91	6.46	**6.19**
RMSE (ug/m^3)	6.09	6.13	7.15	6.67	7.05	6.89	6.36	**6.33**

CHAPTER VII

SUMMARY AND CONCLUSIONS

The NLR ozone forecast models have been successfully applied to daily ozone forecasts for the metro areas Ashland, Bowling Green, Lexington, Louisville, Owensboro, and Paducah. In the year 2003, one more NLR ozone forecast model for the Cincinnatti-Covington metro statistical area (MSA) was developed and applied. The operational O_3 NLR forecast models were hybrid models. The input variables for each of the models were mostly the same, including a group of meteorological parameters and derived ozone predictor parameters. A nonlinear term was obtained in a nonlinear regression fitting process and was used as one of the parameters in the multiple linear regression. It was the most significant term for each of ozone forecast models.

In this study, the updated 2005 NLR ozone forecast models for these metro areas were evaluated on calibration data sets and independent data sets. The metro area MAEs for the 2000-2004 model fits varied from 5.57 ppb to 7.32 ppb (~ 11-12% NMAE). The metro area MAEs for the 2005 model hindcasts varied from 5.90 ppb to 7.20 ppb (~ 10-13% NMAE) and the metro area MAEs for the 2005 model forecasts varied from 7.90 ppb to 9.80 ppb (~ 13-17% NMAE). This level of performance was comparable or superior to other ozone forecast models reported in the literature (the range of reported NMAEs is 12% to 30%).

Based on previously developed NLR ozone forecast models for those areas, Takagi-Sugeno fuzzy system models were developed for seven metro areas in Kentucky, using the fuzzy "c-means" clustering technique coupled with the least square method. The key parameters for a Takagi-Sugeno model were the number of rules (R) and the fuzziness factor (m). The combination of R=5 and m=5 was used in the fuzzy system models, based on a sensitivity study using Louisville air quality and meteorological data. Two types of fuzzy models, basic fuzzy models and NLR-fuzzy models were developed. The basic fuzzy and NLR-fuzzy models exhibited essentially equivalent performance to the existing NLR models on 2004 ozone season hindcasts and forecasts. Both types of fuzzy models had, on average, slightly lower metro area averaged MAEs than the NLR models. For the model hindcasts, the average MAEs of the seven areas were 7.4 ppb, 7.6 ppb, and 7.7 ppb for NLR-fuzzy, basic-fuzzy, and NLR models respectively. For the model forecasts, the average MAEs were 7.8 ppb, 8.0 ppb, and 8.1 ppb for NLR-fuzzy, basic-fuzzy, and NLR models. The small differences may have been statistically significant, but for practical purposes, the models performed essentially the same. Therefore, the choice of which of these models to use for application in ozone air quality forecasting should probably be based on other factors, for example experience of the modeler and software availability.

Among the seven Kentucky metro areas Ashland, Covington, and Louisville are currently designated nonattainment areas for both ground level O_3 and $PM_{2.5}$. In this study, summer $PM_{2.5}$ forecast models were developed for providing summertime daily average $PM_{2.5}$ forecasts for the seven metro areas. The performance of the $PM_{2.5}$ forecast models was generally not as good as that of the ozone forecast models. For example, the summer 2004 model hindcasts had the metro-area average MAE of 5.33μg/m^3 (31.7% NMAE). The $PM_{2.5}$ has a longer atmospheric residence time than ozone and the local meteorological parameters have less influence on local $PM_{2.5}$ concentrations. Therefore, the lower accuracy of $PM_{2.5}$ forecast models compared to ozone forecast models was expected.

High $PM_{2.5}$ concentrations were mostly observed in the summer. However, $PM_{2.5}$ concentrations reach another peak during winter. In this study, exploratory research was conducted to find the relationship between the winter $PM_{2.5}$ concentrations and the meteorological parameters and other derived prediction parameters. Winter $PM_{2.5}$ forecast models were developed for seven selected metro areas in Kentucky. For the model fits, the MAE for the seven forecast models ranged from 3.23 μg/m^3 to 4.61 μg/m^3. The winter NLR $PM_{2.5}$ forecast models for those Kentucky metro areas had slightly higher prediction errors than the respective summer models. For example, the NMAE of the model fits for the winter models ranged from 26% - 28%, compared to 24% - 29% for the summer model. A probable explanation for this fact is that the primary $PM_{2.5}$ pollutants were dominant in the winter time. Meteorological parameters mostly influence secondary $PM_{2.5}$ that is formed by photochemical reactions.

The fuzzy technique was applied on $PM_{2.5}$ forecast models to seek more accurate $PM_{2.5}$ prediction. NLR-fuzzy system summer $PM_{2.5}$ models were developed for the seven metro areas. The NLR-fuzzy $PM_{2.5}$ forecast models had the metro area average MAE 3% less than that of the NLR $PM_{2.5}$ forecast models, for both the model fits and 2004 summer season model hindcasts.

REFERENCES

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.

Chaloulakou A., Assimakopoulos D. and Kekkas T (1999) Forecasting Daily Maximum Ozone Concentrations in the Athens Basin. *Environ Monit Assess*, 56, 97-112.

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. (1999) An Enhanced Ozone Forecasting Model Using Air Mass Trajectory Analysis. *Atmospheric Environment*, 33, 4663-4674.

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

Cobourn W. G. and Lin Y. (2004) Trends in Meteorologically Adjusted Ozone Concentrations in Six Kentucky Metro Areas, *1998-2002. Air & Waste Manage. Assoc.*, 54, 1383-1393.

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

Eder B. and Yu S. (2006) A Performance Evaluation of the 2004 Release of Models-3 CMAQ. *Atmospheric Environment*, 40, 4811-4824.

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 (1999) Overview of Emission Factors and Inventory Methods: PM2.5 Emission Inventory. *U. S. Environmental Protection Agency.* EIIP Volume 4, 1.3.1-1.3.14.

EPA (2004) The Particle Pollution Report: Current Understanding of Air Quality and Emissions through 2003. *U. S. Environmental Protection Agency.* Office of Air Quality Planning and Standards Emissions, Monitoring, and Analysis Division. North Carolina.

EPA (2005a) Six Common Air Pollutants. *U. S. Environmental Protection Agency.* Available at: http://www.epa.gov/air/urbanair/6poll.html (Accessed 2005).

EPA (2005b) National Ambient Air Quality Standards (NAAQS). *U. S. Environmental Protection Agency.* Available at: http://www.epa.gov/air/criteria.html (Accessed 2005).

EPA (2005c) Green Book: Nonattainment Areas for Criteria Pollutants. *U. S. Environmental Protection Agency.* Available at: http://www.epa.gov/oar/oaqps/greenbk/index.html (Accessed 2005).

EPA (2005d) Nitrogen Dioxide: Trend in NO_2 Levels and NOx Emissions. *U. S. Environmental Protection Agency.* Available at: http://www.epa.gov/airtrends/nitrogen.html (Accessed 2005).

EPA (2005e) 8-Hour Ground-level Ozone Designations: Ozone Trends. *U. S. Environmental Protection Agency.* Available at: http://www.epa.gov/ozonedesignations/ozonetrends.htm (Accessed 2005).

EPA (2005f) Six Common Air Pollutants: EPA's Efforts to Reduce SO_2. *U. S. Environmental Protection Agency.* Available at: http://www.epa.gov/air/urbanair/so2/effrt1.html (Accessed 2005).

EPA (2005g) Six Common Air Pollutants: EPA's Efforts to Reduce NOx. *U. S. Environmental Protection Agency.* Available at: http://www.epa.gov/air/urbanair/nox/effrt.html (Accessed 2005).

EPA (2005h) AirData - Reports and Maps. *U.S. Environmental Protection Agency.* Available at: http://www.epa.gov/air/data/reports.html (accessed 2005).

EPA (2005i) Air Emissions Trends - Continued Progress Through 2004. U.S. Environmental Protection Agency. Available at: http://www.epa.gov/airtrends/econ-emissions.html (Accessed 2005).

EPA (2005j) AIRNOW: Quality of Air Means Quality of Life. U.S. Environmental Protection Agency. Available at: http://www.airnow.gov/index.cfm?action=airnow.main.9 (Accessed 2005).

EPA (2006) Community Multi-scale Air Quality (CMAQ) Model. *U. S. Environmental Protection Agency.* Available at: http://www.epa.gov/asmdnerl/CMAQ/cmaq_model.html (Accessed 2006).

Forzatti P. (2001) Presents Status and Perspectives in De-NOx SCR Catalysis; *Appl. Catal. A: General*, 222, 221-236.

FCEAD (2005) PM2.5 Forecast FAQ's; *Forsyth County Environmental Affairs Department.* Available at: http://www.forsyth.cc/envaffairs/default.htm (Accessed 2005).

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.

Heo J. and Kim D. (2004) A New Method of Ozone Forecasting Using Fuzzy Expert and Nerual Network Systems. *Science of the Total Environment*, 325, 221-237.

Hubbard M. C. and Cobourn W. G. (1998) Development of a Regression Model to Forecast Ground-Level Ozone Concentrations in Jefferson County, Kentucky. *Atmospheric Environment*, 32, 2637-2647.

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 Y. (2004) Development of Ozone Forecast Models for Selected Kentucky Metropolitan Areas. M.S. thesis. University of Louisville.

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 (1999-2004 a) Local Climatological Data. National Climatic Data Center, Asheville, North Carolina 28801.

NCDC (1999-2004 b) Unedited Local Climatological Data. National Climatic Data Center, Asheville, North Carolina 28801. Available at: http://cdo.ncdc.noaa.gov/ulcd/ULCD (Accessed 2005).

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

NOAA (2005). Hybrid Single-Particle Lagrangian Integrated Trajectory Model. NOAA Air Resources Laboratory, Silver Spring, MD, USA Available at: http://www.arl.noaa.gov/ready/sec/hysplit4.html (Accessed 2005).

NRCS (2006) Natural Resources Conservation Service: Wind Rose Data. Available at: http://www.wcc.nrcs.usda.gov/climate/windrose.html (Accessed 2006).

Ordieres J. B, Vergara E. P, Capuz R. S, Salazar R. E (2005) Neural Network Prediction Model for Fine Particulate Matter (PM2.5) on the US-Mexico Border in Texas and Chihuahua. *Environmental Modeling & Software*, 20, 547-559.

Passina K. and Yurkovich S. (1998) Fuzzy Control. *Addison Wesley Longman, Inc.* California.

Perez P., Trier A. and Reyes J. (2000) Prediction of PM2.5 Concentrations Several Hours in Advance Using Neural Networks in Santiago, Chile. *Atmospheric Environment*, 34, 1189-1196.

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.

Revlett, G.H., 1978. Ozone Forecasting Using Empirical Modeling. *Journal of Air Pollution Control Association,* 28, 338-343.

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

Rousdseau J., Benjamin M. and Germain A. (2002) Real-time Winter Dispersion Modelling Based on PM2.5 Mass for the Greater Montreal Area in Canada. *Fourth Conference on Atmospheric Chemistry.*

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

Ryan W. (1994) Forecasting Severe Ozone Episodes in the Baltimore Metropolitan Area. *Atmospheric Environment*, 33, 2387-2398.

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.

Schwartz J. (2006) Environment News: How Ozone Is Formed, The Heartland Institute. Available at: http://www.heartland.org/Article.cfm?artId=18974 (Accessed 2006).

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.

UKAWC (2005 a) Kentucky Hourly Weather Observations, Agriculture Weather Center, University of Kentucky. Available at: http://wwwagwx.ca.uky.edu/cgi-public/hourly_www.ehtml (Accessed 2005).

UKAWC (2005 b) Kentucky Climate Data, Agriculture Weather Center, University of Kentucky. Available at: http://wwwagwx.ca.uky.edu/cgi-bin/ky_clim_data_www.pl (Accessed 2005).

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.

Wolff, G.T., Lioy, P.J., 1978. An Empirical Model for Forecasting Maximum Daily Ozone Levels in the Northeastern U.S. *Journal of Air Pollution Control Association,* 28, 1034-1038.

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. Computer Program Codes

1. Computer program used for training fuzzy system models.

```
%*******************************************************************
%This code used to determine a T-S fuzzy model, with clustering method
%and optimal output predefuzzification
%*******************************************************************

clear all;
close all;

%-----load training input and output data--------
load c:\data\o3.txt;
load c:\data\nonlin.txt;
load c:\data\xmitt.txt;
load c:\data\oz48.txt;
load c:\data\traj.txt;
load c:\data\trend.txt;
load c:\data\RHx.txt;
load c:\data\tmndep.txt;
load c:\data\ts.txt;

%overall input data
x=transpose([nonlin xmitt oz48 traj trend RHx tmndep ts]);

%overall output data
y=transpose(o3);
%------------------------------------------------

%-----design parameters-----
[N,M]=size(x);      %get "N", number of input parameters
                    %get "M", number of input-output data pairs
ac=0.001;           %set allowed error when calculating cluster centers

% R=input('Training Lex. 2004 model, Choose R: ');
%                     %number of clusters, specified by designer
% m=input('Choose m: ');
%                     %fuzzy factor, overlap of the clusters

R0=[1 3 5 10 15 20 25 30];                  %train model in group
m0=[1.5 2 3 4 5 6];
for ccR=1:8
    for ccm=1:6
        R=R0(ccR);m=m0(ccm);
%----------------------------


%********************************************************************
%The following code used to clustering the training data set
%********************************************************************
%-----setting initial cluster centers, evenly select centers for each
%-----of the parameters
```

```
    %First determine the linear coefficients for each parameter with
    %LS method. Aim to find the parameters are "+" or "-" correlated
    %with ozone concentrations

fi=transpose([ones(1,M); x]);               %input data
yi=transpose(y);                            %output data
bi=(fi'*fi)^-1*fi'*yi;                      %coefficients
%
%     %Then construct the initial cluster centers
%
for cc1=1:N
    pmax=max(x(cc1,:)); pmin=min(x(cc1,:));   %max. value for the para.
    dist=(pmax-pmin)/R;                       %distance between ini. CC
    if bi(cc1+1,:)>0
        V0(cc1,1)=pmin+dist/2;                %if coefficient of the
        for cc2=2:R                            %parameter greater than 0,
            V0(cc1,cc2)=V0(cc1,cc2-1)+dist;   %then the cluster centers
        end                                     %from min to max
    else                                      %if less than 0, from max
        V0(cc1,1)=pmax-dist/2;                  %to min
        for cc2=2:R
            V0(cc1,cc2)=V0(cc1,cc2-1)-dist;
        end
    end
end
Vold=V0;
%-----End of setting initial cluster centers---


%-----Then find the final cluster centers-----
ccc=1;
for cc3=1:2000

    %step 1. find Unew(i,j) for each input training data
for i=1:M
    for j=1:R
        uden=0;
        for k=1:R
            uden=uden+(norm(x(:,i)-Vold(:,j)))^2/(norm(x(:,i)-
Vold(:,k)))^2;
        end
        Unew(i,j)=1/uden;
    end
end

    %step 2. calculate new cluster center
for j=1:R
    num1=0;
    den1=0;
    for i=1:M
        num1=num1+x(:,i)*Unew(i,j)^m;
        den1=den1+Unew(i,j)^m;
    end
    Vnew(:,j)=num1/den1;
end

    %step 3. calculate distance between new and old cluster center
    %if the distence is greater than 'ac', then 'esum' count 1
    %when esum is zero, terminate loop

esum=0;
for j=1:R
    e(j)=norm(Vnew(:,j)-Vold(:,j));
    if e(j)<ac
        cc3=0;
    else
        cc3=1;
    end
    esum=esum+cc3;
end

if esum==0
    break
else
    Vold=Vnew;
    ccc=ccc+1;
end
%-------------------------
end

V=Vnew;             %save the new cluster centers



%****************************************************************
%the following code used to determine the parameters aj
%with least square method
%****************************************************************

%-------calculate membership function UH for each input data-----
for i=1:M
    for j=1:R
        den=0;
        for k=1:R
            den=den+norm(x(:,i)-V(:,j))^2/norm(x(:,i)-V(:,k))^2;
        end
        UH(i,j)=1/den;
    end
end
%----------------------

Mone=ones(1,M);
Xhead=transpose([Mone;x]);      %creat matrix X head

Y=transpose(y);                 %creat matrix Y


for j=1:R                       %calculate aj for each rule
    Dj=diag(UH(:,j));
    aj(:,j)=pinv(transpose(Xhead)*Dj^2*Xhead)*transpose(Xhead)*Dj^2*Y;
end

%-----save the parameters in file--------------------
if m<2
```

```
    parafile=strcat('parameterR',num2str(R),'m',num2str(m*10));
else
    parafile=strcat('parameterR',num2str(R),'m',num2str(m));
end
save(parafile,'M','N','R','V0','V','Y','ac','aj','bi','m','o3');
%----------------------

    end         %end loop for m
end             %end loop for R
```

2. Computer program used for testing fuzzy system models.

```
%*************************************************************
%This code used to test the T-S fuzzy model with testing data set
%*************************************************************

clear all;
close all;

%-----parameter setting-----
NAAQS=80.0;                 %set NAAQS
Threshold=75.0;             %set test threshold
%---------------------------


%-----load parameters of the T-S fuzzy model---
R = input('Testing model, Choose R: ');
m = input('Choose m: ');
parafile=strcat('parameterR',num2str(R),'m',num2str(m));
load (parafile);
%---------------------------

%-----load testing data set-----
load c:\data\o3.txt;
load c:\data\nonlin.txt;
load c:\data\xmitt.txt;
load c:\data\oz48.txt;
load c:\data\traj.txt;
load c:\data\trend.txt;
load c:\data\RHx.txt;
load c:\data\tmndep.txt;
load c:\data\ts.txt;

%overall input data
x=transpose([nonlin xmitt oz48 traj trend RHx tmndep ts]);

%overall output data
y=transpose(o3);

[N,M]=size(x);      %get "N", number of input parameters
                    %get "M", number of input-output testing data pairs
```

```
%-------------------------------


%-----test fuzzy model with training data set-----

for i=1:M

     for j=1:R                   %find values of the membership functions
        temp=0;
        for k=1:R
            temp=temp+(norm(x(:,i)-V(:,j)))^2/(norm(x(:,i)-V(:,k)))^2;
        end
        UH(j)=1/temp;
    end

    num=0;                      %calculate output of fuzzy model
    den=0;
    for j=1:R
        g=[1 transpose(x(:,i))]*aj(:,j);
        num=num+g*UH(j);
        den=den+UH(j);
    end
    yt(i)=num/den;

    T(i)=i;

end
%-----------------------

% %-----time series plots-----
% pn=100;
% plot(T(1:pn),y(1:pn),T(1:pn),yt(1:pn),'r');
% xlabel('number of day'),ylabel('model pred. and observation(ppb)');
% grid;
% %----------

%-----statistics calculations-----

    %bias, MAE, Rsquare :
bias=0;
mae=0;
Rnum=0;
Rden=0;
avgY=norm(y,1)/M;
rmse=0;

for cc1=1:M
    error=yt(cc1)-y(cc1);
    bias=bias+error;
    mae=mae+abs(error);
    Rnum=Rnum+(yt(cc1)-avgY)^2;
    Rden=Rden+(y(cc1)-avgY)^2;
    rmse=rmse+error^2;

end

Bias=bias/M;
```

```
MAE=mae/M;
RMSE=sqrt(rmse/M);
Rsquare=Rnum/Rden;

    %model performance parameter :
DE=0;               %number of detected exceedances
EX=0;               %number of observed exceedances
FA=0;               %false alarms
AL=0;               %total alarms

for cc2=1:M
    if y(cc2)>=NAAQS
        EX=EX+1;
        if yt(cc2)>=Threshold
            DE=DE+1;
        end
    end

    if y(cc2)<Threshold
        if yt(cc2)>=Threshold
            FA=FA+1;
        end
    end

    if yt(cc2)>=Threshold
        AL=AL+1;
    end
end

DR=DE/EX;                 %detection rate
FAR=FA/AL;                %false alarm rate
CSI=(AL-FA)/(AL+EX-DE); %critical success index


%------results output-----

sheetname=strcat('R',num2str(R));   %creat Excel filename

    %Testing data set and fuzzy model predictions
tdata0={'O3_obs', 'O3_pred'};
xlswrite('results.xls', tdata0, sheetname,'A1');
xlswrite('results.xls', y', sheetname,'A2');
xlswrite('results.xls', yt', sheetname,'B2');

  %Fuzzy Model design parameters
 par={'R(rules)','m(overlap)','N(inputs)','M(datapair)','ac'; R m N M
ac}';
 Std={'NAAQS', 'Threshold'; NAAQS Threshold}';
 xlswrite('results.xls', par, sheetname,'d2');
 xlswrite('results.xls', Std, sheetname,'d8');

    %Error statistics
Serror0 = {'Error_statistics'};
Serror = {'O3_avg','Rsquare','Bias','MAE','RMSE'; avgY Rsquare Bias MAE
RMSE}';
xlswrite('results.xls', Serror0, sheetname,'g2');
xlswrite('results.xls', Serror, sheetname, 'g4');
```

```
    %Performance statistics
Sperf0={'Performance_statistics'};
Sperf={'DR','FAR','CSI','DE','AL','FA','EX'; DR FAR CSI DE AL FA EX}';
xlswrite('results.xls', Sperf0, sheetname,'j2');
xlswrite('results.xls', Sperf, sheetname,'j4');
```

Appendix B. Parameters for Fuzzy System Ozone Forecast Models

Table A.1 Final cluster centers v^i for NLR-fuzzy system model (Ashland)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	V_1^i	41.55	54.79	70.39	54.79	85.04
Xmitt	V_2^i	0.64	0.64	0.65	0.64	0.65
Trend	V_3^i	2.58	2.19	2.29	1.69	0.86
RH	V_4^i	92.40	72.07	60.76	43.92	40.98
CC	V_5^i	7.93	4.06	1.59	1.79	0.88
WS	V_6^i	5.94	6.58	5.86	8.16	5.31

Table A.2 Coefficients a_i for NLR-fuzzy system model (Ashland)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-93.16	-134.20	-291.60	-260.61	-142.43
Nonlin	a_1	0.45	0.57	0.77	0.73	0.71
Xmitt	a_2	238.32	287.00	494.86	426.01	266.75
Trend	a_3	-0.20	-0.79	-0.07	0.36	-0.10
RH	a_4	-0.30	-0.27	-0.10	0.04	-0.08
CC	a_5	-1.00	-0.69	-0.81	-0.74	-2.25
WS	a_6	-0.73	-0.63	-0.69	0.28	0.19

Table A.3 Final cluster centers v^i for basic-fuzzy system model (Ashland)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Tmax	V_1^i	73.23	78.92	83.80	88.01	80.88
WS	V_2^i	5.25	5.27	6.41	6.10	6.45
RHx	V_3^i	84.82	70.26	59.29	48.06	36.64
Xmitt	V_4^i	0.64	0.64	0.65	0.65	0.64
Trend	V_5^i	2.71	2.14	2.03	2.12	0.81
RH	V_6^i	93.67	79.21	65.73	51.08	35.99
CC	V_7^i	8.00	5.14	2.80	1.31	0.62

Table A.4 Coefficients a_i for basic-fuzzy system model (Ashland)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-44.34	-122.78	-223.56	-236.79	-141.23
Tmax	a_1	0.57	0.61	0.81	0.99	1.42
RHx	a_2	0.08	-0.16	-0.26	-0.42	-0.43
WS	a_3	-1.00	-0.76	-1.70	-1.94	-1.22
Xmitt	a_4	137.91	265.88	410.88	422.69	199.30
Trend	a_5	-0.97	-0.57	-0.45	0.83	-0.71
RH	a_6	-0.44	-0.33	-0.25	-0.31	-0.15
CC	a_7	-0.60	-0.75	-0.96	-0.78	1.07

Table A.5 Final cluster centers v^i for NLR-fuzzy system model (Bowling Green)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	V_1^i	39.11	51.09	65.58	61.09	86.67
Xmitt	V_2^i	0.65	0.65	0.65	0.65	0.64
Trend	V_3^i	2.73	2.50	2.02	2.13	0.34
RH	V_4^i	87.14	71.69	56.57	45.13	28.33
Tmn_dep	V_5^i	3.07	4.31	3.52	-7.44	-2.85
WS	V_6^i	8.46	7.89	7.58	7.93	6.17

Table A.6 Coefficients a_i for NLR-fuzzy system model (Bowling Green)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-193.57	-144.55	-141.95	-268.83	-197.06
Nonlin	a_1	0.55	0.51	0.65	0.78	0.96
Xmitt	a_2	356.71	316.00	300.07	459.15	313.56
Trend	a_3	-0.22	0.02	-0.16	-0.09	0.20
RH	a_4	-0.20	-0.44	-0.45	-0.20	-0.05
Tmn_dep	a_5	-0.01	-0.08	0.06	0.31	0.71
WS	a_6	-0.50	-0.72	-0.54	-0.22	0.47

Table A.7 Final cluster centers v^i for basic-fuzzy system model (Bowling Green)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Tmax	V_1^i	77.98	85.07	90.25	79.47	93.04
WS	V_2^i	8.10	7.88	7.91	7.20	6.62
RHx	V_3^i	80.68	67.90	53.95	41.71	35.73
Xmitt	V_4^i	0.65	0.65	0.65	0.65	0.64
Trend	V_5^i	2.61	2.51	2.02	2.29	0.27
RH	V_6^i	85.53	71.29	55.74	47.15	27.23
Tmn_dep	V_7^i	4.64	4.87	3.76	-7.42	-3.95

Table A.8 Coefficients a_i for basic-fuzzy system model (Bowling Green)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-165.73	-125.05	-114.82	-237.90	-208.53
Tmax	a_1	0.48	0.44	0.56	0.72	1.10
WS	a_2	-0.80	-0.86	-0.94	-0.45	-0.46
RHx	a_3	-0.18	-0.17	-0.52	-0.60	-1.01
Xmitt	a_4	326.05	301.60	296.33	444.80	356.44
Trend	a_5	-0.11	0.06	-0.25	0.17	-0.23
RH	a_6	-0.27	-0.55	-0.47	-0.25	0.14
Tmn_dep	a_7	-0.13	-0.25	0.13	0.47	0.85

Table A.9 Final cluster centers v^i for NLR-fuzzy system model (Covington)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	V_1^i	40.78	57.30	53.35	75.63	87.89
Xmitt	V_2^i	0.64	0.64	0.64	0.65	0.64
Trend	V_3^i	2.26	2.05	1.98	1.79	1.12
RHx	V_4^i	91.48	72.85	45.53	58.59	35.57
Tmn_dep	V_5^i	2.58	3.94	-7.06	2.46	-2.97

Table A.10 Coefficients a_i for NLR-fuzzy system model (Covington)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-152.15	-140.59	-267.55	-368.36	-409.12
Nonlin	a_1	0.53	0.69	0.73	0.95	0.82
Xmitt	a_2	311.40	285.13	453.66	587.97	684.97
Trend	a_3	-0.39	0.30	0.43	0.66	1.34
RHx	a_4	-0.32	-0.39	-0.13	-0.11	-0.38
Tmn_dep	a_5	0.08	0.06	0.37	-0.08	0.46

Table A.11 Final cluster centers v^i for basic-fuzzy system model (Covington)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Tmax	V_1^i	71.29	82.78	87.49	70.95	87.20
WS	V_2^i	9.37	7.88	8.00	9.59	7.75
RHx	V_3^i	92.67	74.92	58.61	49.04	34.59
Xmitt	V_4^i	0.64	0.64	0.65	0.64	0.64
Trend	V_5^i	2.44	2.12	1.88	2.26	1.27
Tmn_dep	V_6^i	2.56	4.12	3.13	-8.78	-2.68

Table A.12 Coefficients a_i for basic-fuzzy system model (Covington)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-135.52	-128.88	-377.70	-281.32	-314.42
Tmax	a_1	0.60	1.01	1.40	1.04	1.60
WS	a_2	-0.11	-0.87	-1.75	-0.59	-1.59
RHx	a_3	-0.48	-0.62	-0.39	-0.27	-0.40
Xmitt	a_4	276.41	242.22	569.97	436.19	443.17
Trend	a_5	-0.37	-0.07	0.17	0.62	0.17
Tmn_dep	a_6	-0.11	-0.35	-0.36	0.05	0.25

Table A.13 Final cluster centers v^i for NLR-fuzzy system model (Lexington)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	V_1^i	37.68	48.56	58.78	67.46	84.69
Xmitt	V_2^i	0.64	0.65	0.65	0.65	0.64
Trend	V_3^i	2.15	2.32	2.07	1.63	1.28
RH	V_4^i	88.55	72.30	58.45	42.87	32.20
OZ48	V_5^i	0.09	0.10	0.09	0.29	0.47
Traj	V_6^i	0.14	0.23	0.23	0.28	0.49
CC	V_7^i	7.90	4.31	2.02	1.07	0.73

Table A.14 Coefficients a_i for NLR-fuzzy system model (Lexington)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-83.85	-105.66	-179.73	-239.04	-226.34
Nonlin	a_1	0.19	0.56	0.61	0.86	0.71
Xmitt	a_2	218.65	229.11	331.70	391.46	400.36
Trend	a_3	-0.62	-1.08	-0.92	-1.11	-0.72
RH	a_4	-0.27	-0.26	-0.15	-0.09	-0.29
OZ48	a_5	6.22	6.67	7.94	10.18	8.91
Traj	a_6	3.73	2.89	2.69	2.35	-2.05
CC	a_7	-0.58	-0.48	-1.11	-0.18	0.44

Table A.15 Final cluster centers v^i for basic-fuzzy system model (Lexington)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Tmax	V_1^i	75.93	82.83	87.75	76.77	89.51
WS	V_2^i	8.47	8.51	8.20	8.09	7.77
RHx	V_3^i	82.82	68.42	56.56	46.56	41.49
Xmitt	V_4^i	0.64	0.65	0.65	0.64	0.64
Trend	V_5^i	2.16	2.41	2.01	2.04	1.38
RH	V_6^i	88.00	72.59	59.47	47.17	32.67
OZ48	V_7^i	0.11	0.12	0.14	0.06	0.37
Traj	V_8^i	0.15	0.22	0.28	0.15	0.30
CC	V_9^i	7.64	4.39	2.01	1.27	0.70

Table A.16 Coefficients a_i for basic-fuzzy system model (Lexington)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-23.21	-87.72	-130.23	-254.60	-164.64
Tmax	a_1	0.31	0.36	0.47	0.76	1.03
WS	a_2	-0.19	-0.64	-0.69	-0.24	-0.20
RHx	a_3	-0.01	-0.21	-0.32	-0.43	-0.56
Xmitt	a_4	113.55	243.40	297.52	427.81	274.76
Trend	a_5	-0.53	-1.23	-1.59	0.08	-1.22
RH	a_6	-0.35	-0.38	-0.23	0.00	-0.15
OZ48	a_7	5.93	7.25	8.25	10.01	10.85
Traj	a_8	4.07	1.65	0.88	3.77	-0.93
CC	a_9	-0.51	-0.58	-1.18	-0.48	0.88

Table A.17 Final cluster centers v^i for NLR-fuzzy system model (Owensboro)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	V_1^i	41.27	56.87	52.82	72.01	99.17
Xmitt	V_2^i	0.65	0.65	0.64	0.65	0.64
Trend	V_3^i	2.07	1.95	2.04	2.30	0.61
Tmn_dep	V_4^i	4.78	5.00	-8.31	1.21	-2.23
CC	V_5^i	6.70	2.70	1.58	1.28	0.62
Dewpt	V_6^i	68.10	72.16	47.22	68.50	54.10
WS	V_7^i	8.03	9.30	9.23	6.97	4.74

Table A.18 Coefficients a_i for NLR-fuzzy system model (Owensboro)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-154.70	-207.83	-287.32	-292.12	-179.63
Nonlin	a_1	0.83	0.88	0.95	0.96	0.90
Xmitt	a_2	304.87	401.49	470.80	507.00	338.38
Trend	a_3	-0.56	-0.95	0.23	-0.71	-2.37
Tmn_dep	a_4	0.05	0.34	0.36	0.27	1.07
CC	a_5	-0.79	-1.67	-0.12	-1.10	-2.27
Dewpt	a_6	-0.40	-0.51	-0.18	-0.44	-0.40
WS	a_7	-0.27	-0.62	-0.19	0.08	0.66

Table A.19 Final cluster centers v^i for basic-fuzzy system model (Owensboro)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Tmax	V_1^i	77.87	87.02	91.59	84.14	72.12
WS	V_2^i	7.23	8.62	7.25	6.96	8.66
RHx	V_3^i	82.35	65.85	51.97	45.55	44.59
Xmitt	V_4^i	0.64	0.65	0.65	0.65	0.64
Trend	V_5^i	2.00	1.91	2.21	1.97	2.00
Tmn_dep	V_6^i	5.51	5.22	3.84	-5.78	-9.59
CC	V_7^i	6.80	3.09	1.29	1.27	1.42
Dewpt	V_8^i	69.66	72.00	71.40	60.12	45.14

Table A.20 Coefficients a_i for basic-fuzzy system model (Owensboro)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-194.85	-158.97	-312.75	-247.86	-265.10
Tmax	a_1	1.16	0.82	1.43	1.67	0.98
WS	a_2	-0.62	-0.98	-1.01	-0.51	-0.16
RHx	a_3	-0.27	-0.46	-0.56	-0.57	-0.67
Xmitt	a_4	361.58	357.27	530.71	370.17	457.46
Trend	a_5	-0.15	-0.99	-1.15	-0.31	0.43
Tmn_dep	a_6	-0.06	0.17	0.24	0.35	0.51
CC	a_7	-0.71	-1.58	-0.72	0.60	0.47
Dewpt	a_8	-0.84	-0.62	-0.72	-0.54	-0.25

Table A.21 Final cluster centers v^i for NLR-fuzzy system model (Paducah)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	V_1^i	41.59	58.46	52.22	75.09	86.82
Xmitt	V_2^i	0.65	0.65	0.64	0.65	0.64
Trend	V_3^i	2.25	2.09	2.17	2.16	1.00
Tmn_dep	V_4^i	6.25	4.91	-7.14	3.71	-7.15
CC	V_5^i	6.04	2.27	1.56	1.37	0.54
Dewpt	V_6^i	68.82	71.47	47.62	71.17	52.93
WS	V_7^i	7.07	8.04	9.07	6.24	5.19

Table A.22 Coefficients a_i for NLR-fuzzy system model (Paducah)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-124.89	-121.62	-291.05	-242.08	-161.47
Nonlin	a_1	0.66	0.85	0.83	1.01	1.06
Xmitt	a_2	287.27	250.78	473.59	447.69	251.85
Trend	a_3	0.73	-0.15	-0.47	-1.09	-0.66
Tmn_dep	a_4	0.09	0.02	0.26	0.37	0.31
CC	a_5	-1.48	-1.72	-0.43	-1.75	-0.34
Dewpt	a_6	-0.61	-0.35	0.01	-0.68	0.01
WS	a_7	-0.10	-0.50	-0.17	0.10	-0.16

Table A.23 Final cluster centers v^i for basic-fuzzy system model (Paducah)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Tmax	V_1^i	77.73	88.72	93.24	84.92	71.45
WS	V_2^i	6.01	8.13	6.92	6.65	9.91
RHx	V_3^i	83.67	63.92	48.41	37.58	49.32
Xmitt	V_4^i	0.65	0.65	0.65	0.64	0.64
Trend	V_5^i	2.07	2.16	2.14	1.88	2.22
Tmn_dep	V_6^i	5.53	6.21	3.63	-7.82	-7.25
CC	V_7^i	6.72	2.79	1.37	0.92	2.38
Dewpt	V_8^i	69.42	72.26	70.30	54.91	45.36

Table A.24 Coefficients a_i for basic-fuzzy system model (Paducah)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-115.64	-148.41	-146.51	-223.24	-274.81
Tmax	a_1	0.39	1.30	1.34	1.36	0.95
WS	a_2	-0.16	-1.23	-0.89	-1.04	-0.49
RHx	a_3	-0.23	-0.35	-0.36	-0.64	-0.39
Xmitt	a_4	300.99	281.93	276.26	345.92	446.06
Trend	a_5	1.12	0.14	-0.94	-0.79	-0.05
Tmn_dep	a_6	-0.04	-0.15	0.03	0.37	0.16
CC	a_7	-1.80	-1.08	-1.94	-0.49	-0.25
Dewpt	a_8	-0.63	-0.83	-0.84	-0.15	-0.13

Appendix C. Time Series Plots for NLR-fuzzy Ozone forecast Models

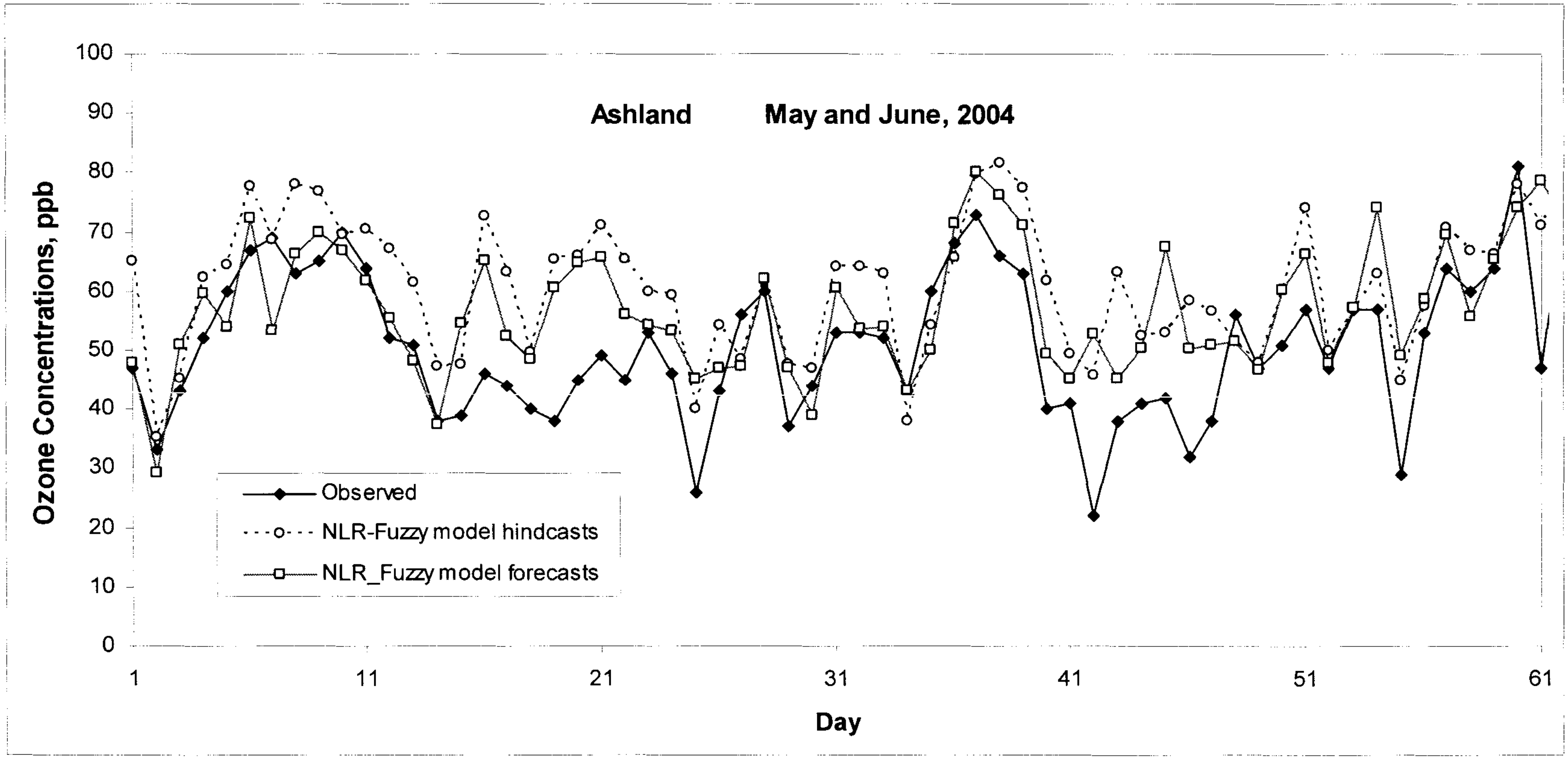


Figure A1 Time series of observed 8-hr ozone concentrations and NLR-fuzzy model hindcasts and forecasts for May and June, 2004. (Ashland)

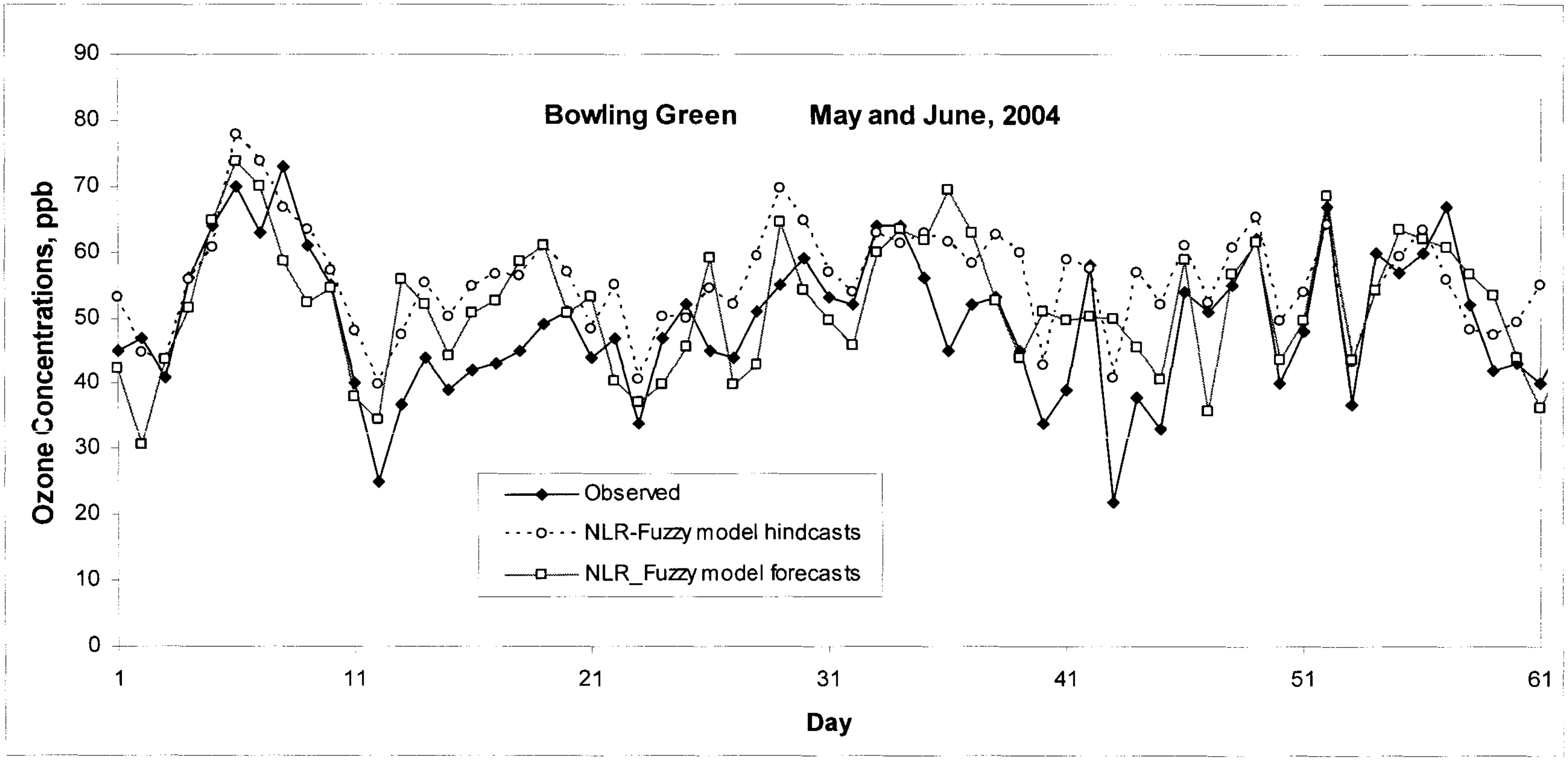


Figure A2 Time series of observed 8-hr ozone concentrations and NLR-fuzzy model hindcasts and forecasts for May and June, 2004. (Bowling Green)

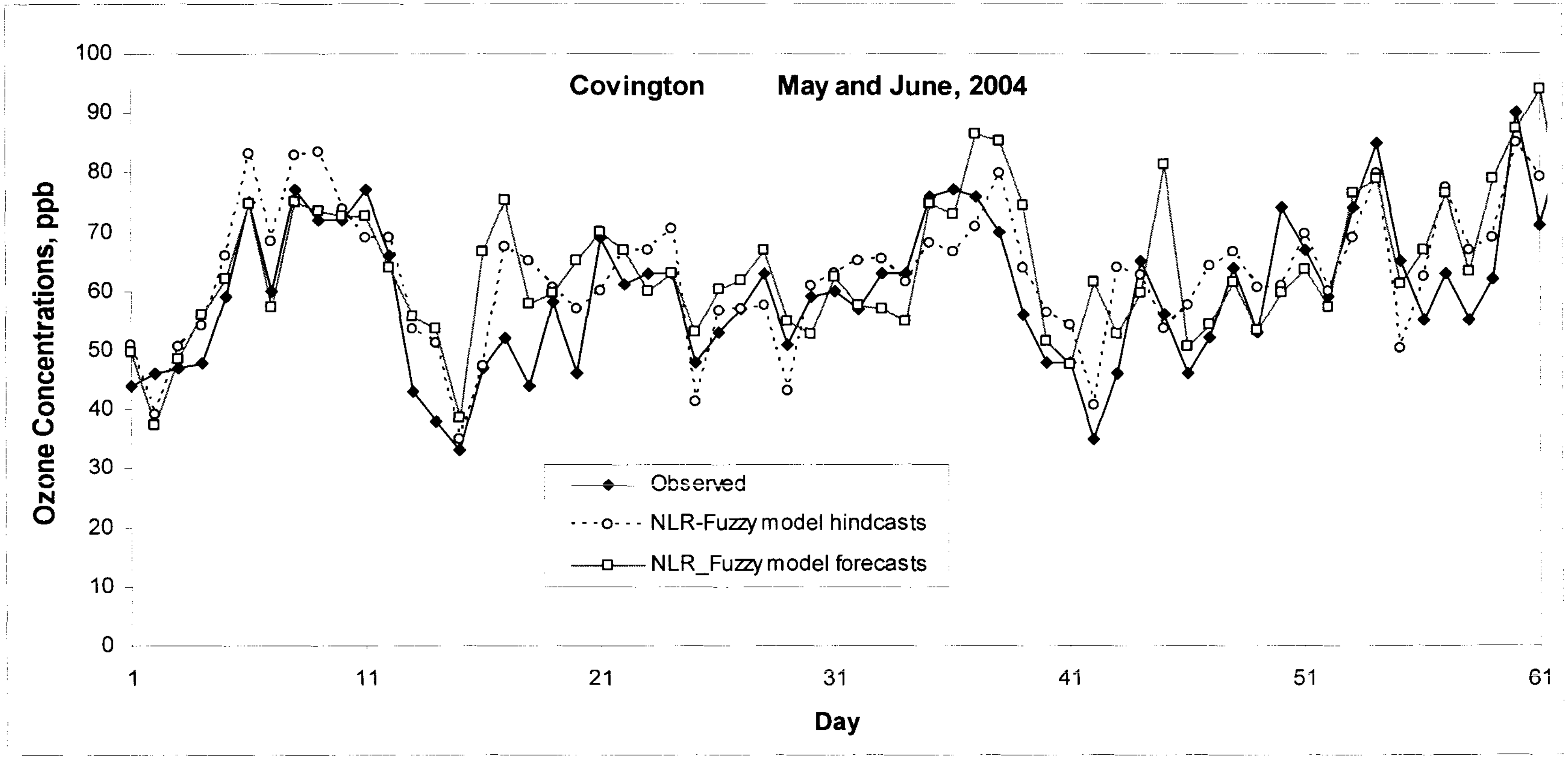


Figure A3 Time series of observed 8-hr ozone concentrations and NLR-fuzzy model hindcasts and forecasts for May and June, 2004. (Covington)

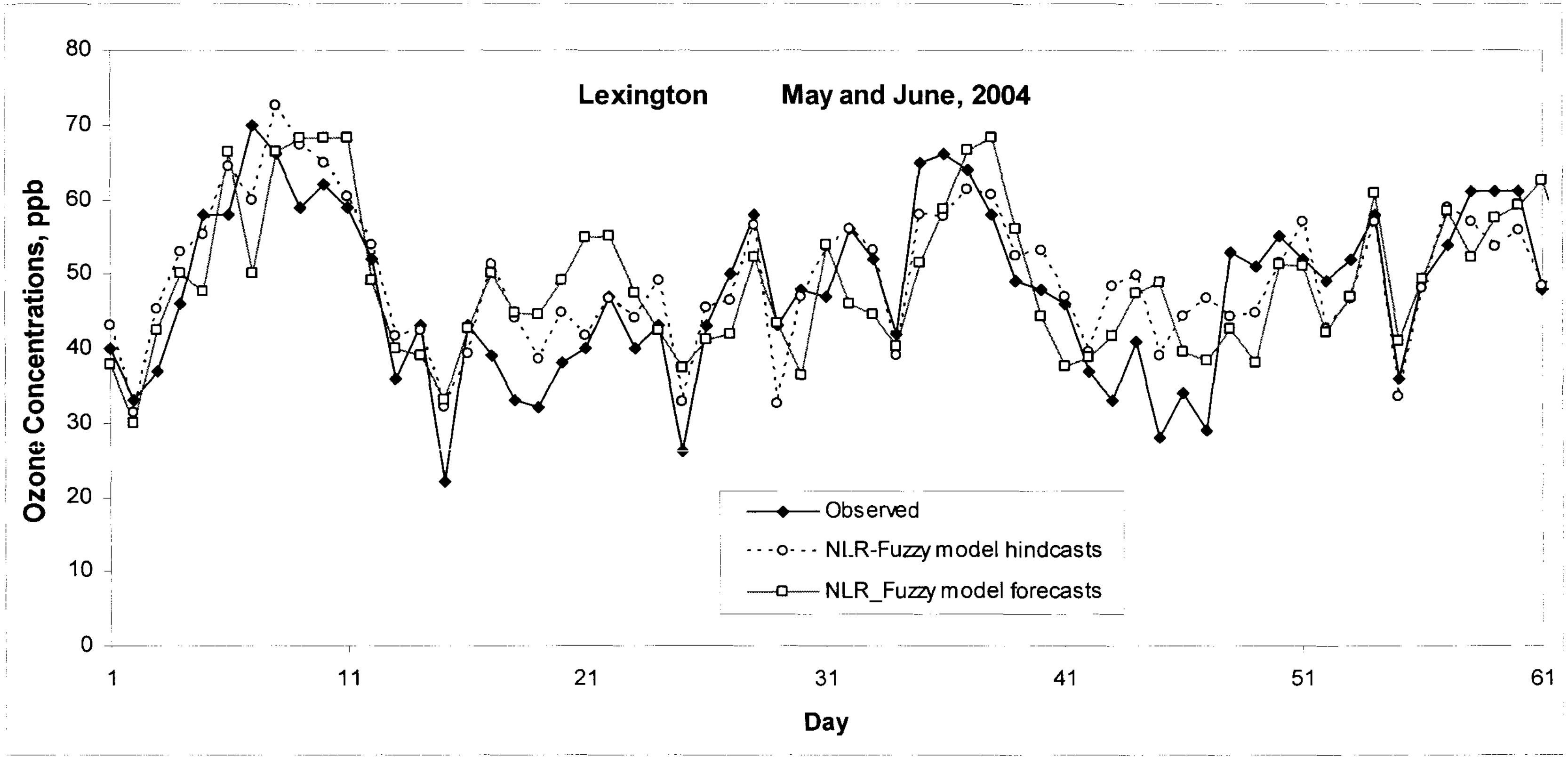


Figure A4 Time series of observed 8-hr ozone concentrations and NLR-fuzzy model hindcasts and forecasts for May and June, 2004. (Lexington)

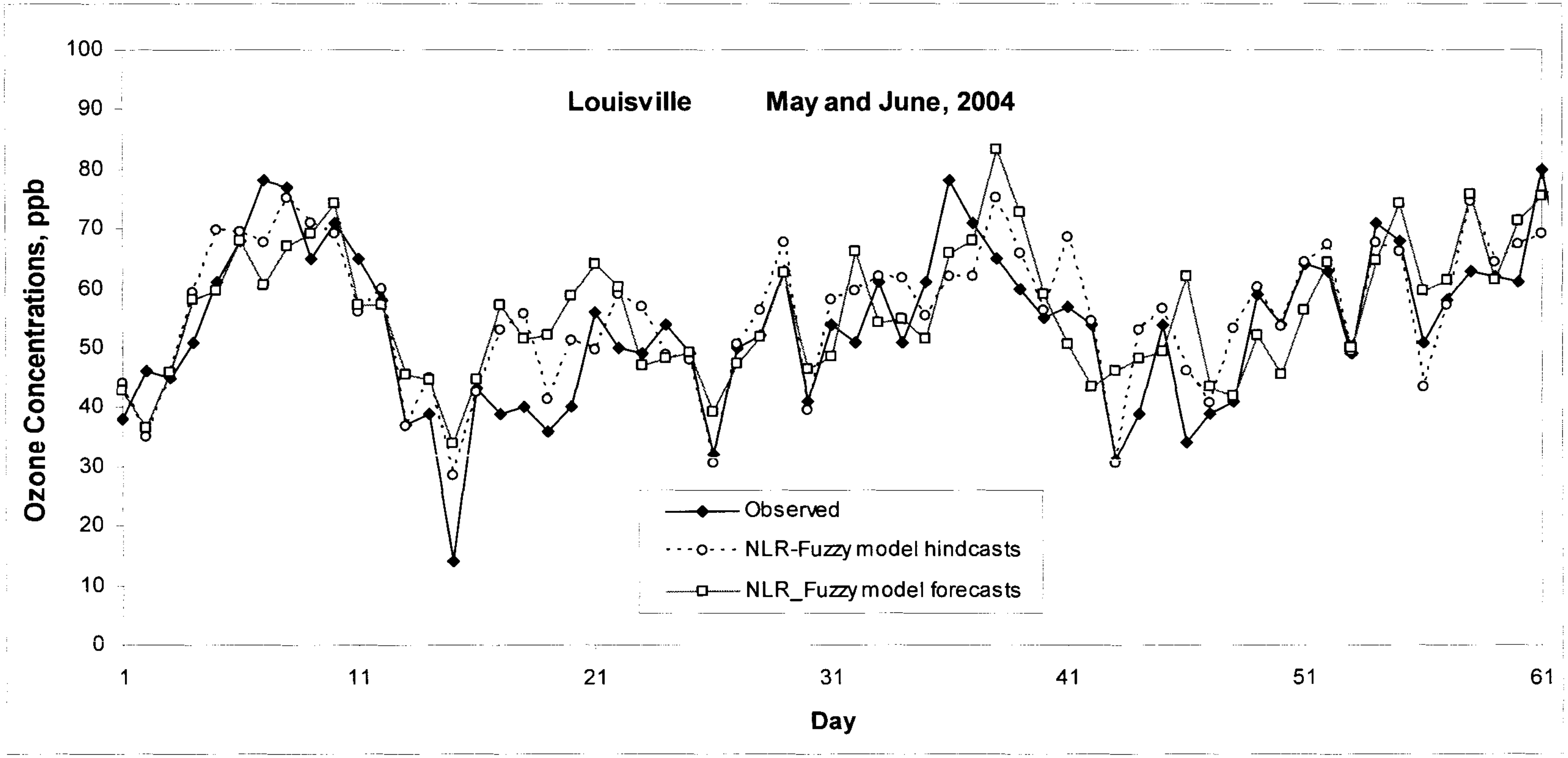


Figure A5 Time series of observed 8-hr ozone concentrations and NLR-fuzzy model hindcasts and forecasts for May and June, 2004. (Louisville)

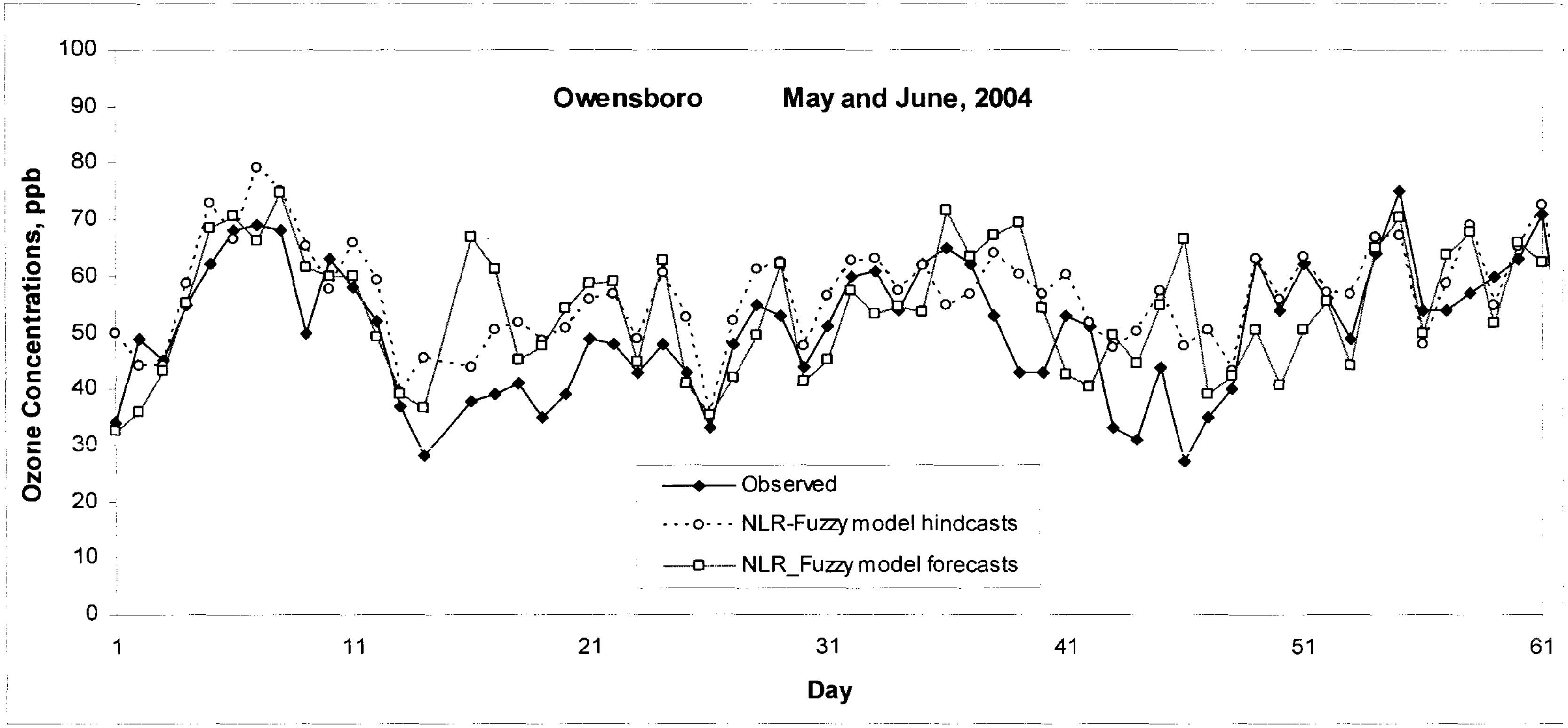


Figure A6 Time series of observed 8-hr ozone concentrations and NLR-fuzzy model hindcasts and forecasts for May and June, 2004. (Owensboro)

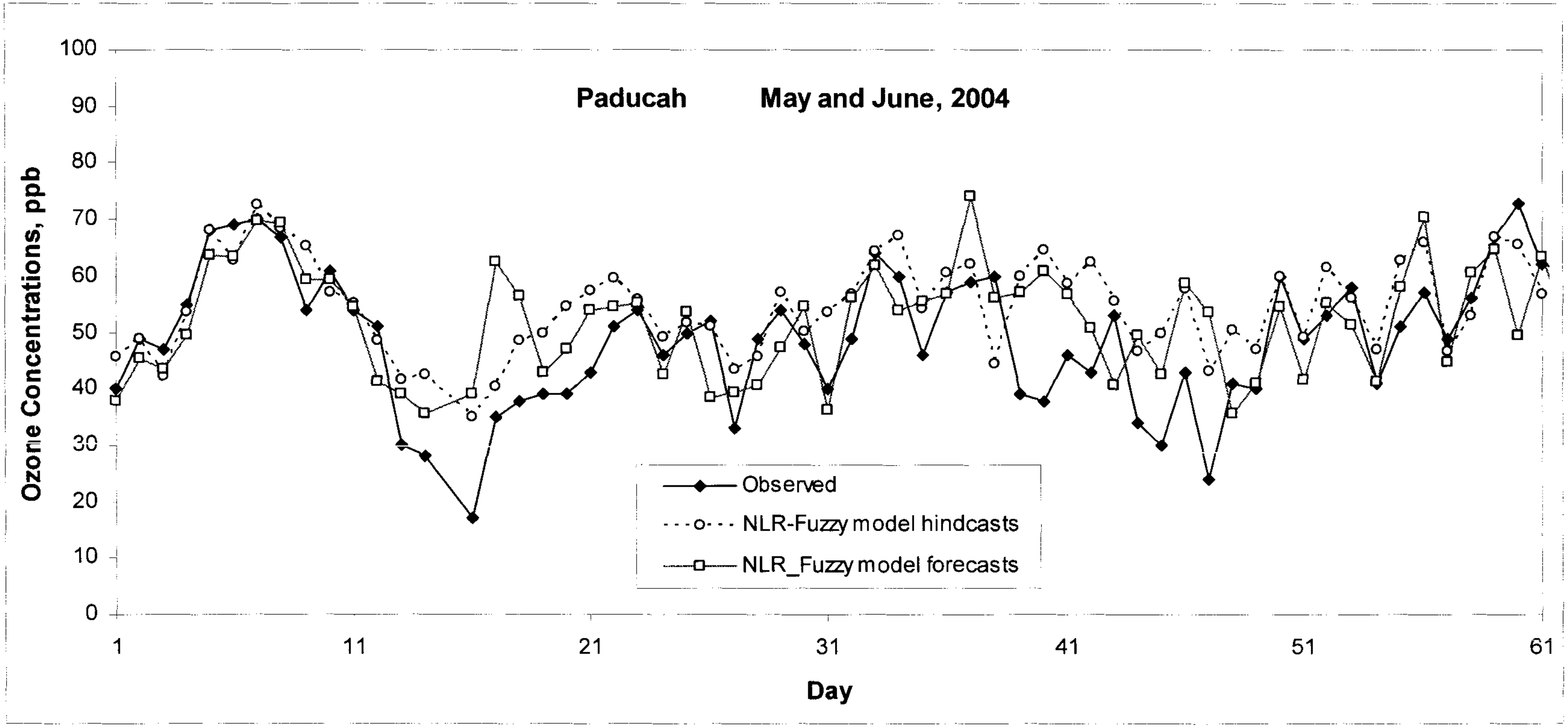


Figure A7 Time series of observed 8-hr ozone concentrations and NLR-fuzzy model hindcasts and forecasts for May and June, 2004. (Paducah)

Appendix D. Parameters for NLR-fuzzy $PM_{2.5}$ Forecast Models

Table A.25 Final cluster centers v^i for NLR-fuzzy system model (Bowling Green)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	v^1	8.01	13.21	18.12	22.78	24.86
Trend	v^2	2.75	1.17	2.53	3.24	0.11
CC	v^3	4.47	1.85	1.61	1.29	1.93

Table A.26 Coefficients a_i for NLR-fuzzy system model (Bowling Green)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	0.68	3.25	0.71	-11.78	-1.98
Nonlin	a_1	0.80	0.75	0.96	1.49	1.09
Trend	a_2	0.20	-0.55	-0.06	-0.03	0.62
CC	a_3	0.44	0.58	-0.18	0.47	0.30

Table A.27 Final cluster centers v^i for NLR-fuzzy system model (Covington)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	v^1	12.73	13.55	22.34	19.03	30.13
Dewpt	v^2	36.71	48.58	58.35	68.13	71.13
Rain	v^3	0.01	0.11	0.01	0.39	0.12
CC	v^4	3.31	6.08	3.89	8.11	5.75
Windrose	v^5	-0.29	-0.16	0.03	0.42	-0.02
Hol	v^6	0.00	0.00	0.00	0.01	0.04
Sat	v^7	0.15	0.20	0.16	0.16	0.14

Table A.28 Coefficients a_i for NLR-fuzzy system model (Covington)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-0.39	-10.93	-14.12	2.25	-10.56
Nonlin	a_1	0.63	0.55	0.73	0.59	1.07
Dewpt	a_2	0.14	0.31	0.29	0.09	0.14
Rain	a_3	1.73	-3.16	-2.06	-3.67	-3.74
CC	a_4	-0.15	0.37	0.37	0.26	0.06
Windrose	a_5	2.87	1.82	2.92	2.34	2.98
Hol	a_6	15.99	13.39	12.76	3.19	5.11
Sat	a_7	-2.94	-0.76	-2.40	-2.54	-1.96

Table A.29 Final cluster centers v^i for NLR-fuzzy system model (Lexington)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	v^1	16.81	12.75	17.72	15.76	25.25
Trend	v^2	0.63	1.86	1.76	2.00	1.91
Dewpt	v^3	32.31	44.90	56.90	68.50	70.74
Rain	v^4	0.00	0.02	0.02	0.37	0.08
Hol	v^5	0.00	0.00	0.00	0.00	0.00
Traj	v^6	0.69	0.08	0.17	0.06	0.25
OZ48	v^7	0.00	0.03	0.09	0.08	0.13

Table A.30 Coefficients a_i for NLR-fuzzy system model (Lexington)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-5.97	-10.75	-5.87	13.82	-20.40
Nonlin	a_1	1.17	0.45	0.48	0.90	1.24
Trend	a_2	1.52	0.35	0.57	-0.29	-0.01
Dewpt	a_3	-0.07	0.35	0.22	-0.17	0.20
Rain	a_4	3.25	-5.98	-3.08	-0.52	-8.09
Hol	a_5	24.24	28.15	29.91	24.81	23.50
Traj	a_6	-2.51	6.82	1.56	1.57	2.39
OZ48	a_7	5.40	-6.64	7.82	2.93	7.25

Table A.31 Final cluster centers v^i for NLR-fuzzy system model (Louisville)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	v^1	10.17	12.87	22.64	21.33	33.01
Trend	v^2	2.17	1.93	1.94	2.08	1.98
Dewpt	v^3	37.57	48.96	61.10	69.81	70.68
Rain	v^4	0.01	0.07	0.02	0.28	0.06
Windrose	v^5	0.02	-0.19	0.03	0.43	0.05
Hol	v^6	0.00	0.00	0.00	0.00	0.01
Sat	v^7	0.12	0.13	0.14	0.16	0.15
Tmn_dep	v^8	-13.16	-6.24	-3.41	5.25	6.14
Traj	v^9	0.21	0.11	0.29	0.12	0.46
OZ48	v^{10}	0.00	0.01	0.08	0.10	0.19

Table A.32 Coefficients a_i for NLR-fuzzy system model (Louisville)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-7.81	-5.77	-8.87	10.23	-7.86
Nonlin	a_1	0.52	0.27	0.43	0.76	0.64
Trend	a_2	0.78	0.23	0.34	-0.15	1.03
Dewpt	a_3	0.23	0.30	0.31	-0.10	0.20
Rain	a_4	0.15	-2.40	-5.72	-2.37	-7.41
Windrose	a_5	1.21	1.86	3.92	1.86	1.10
Hol	a_6	32.35	29.93	18.76	16.31	30.60
Sat	a_7	-2.71	-1.07	-2.43	-1.24	-1.43
Tmn_dep	a_8	0.02	0.06	0.22	0.08	0.13
Traj	a_9	3.88	4.75	3.16	5.24	7.16
OZ48	a_{10}	7.57	4.21	6.89	6.82	3.88

Table A.33 Final cluster centers v^i for NLR-fuzzy system model (Owensboro)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	v^1	8.40	13.00	23.88	14.90	27.13
Trend	v^2	1.38	2.42	1.89	1.96	1.94
Dewpt	v^3	33.83	48.37	60.33	68.83	72.71
Windrose	v^4	-0.40	-0.29	-0.08	0.17	-0.09
Hol	v^5	0.00	0.00	0.00	0.00	0.00

Table A.34 Coefficients a_i for NLR-fuzzy system model (Owensboro)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	-5.49	-10.73	-18.60	8.72	-7.75
Nonlin	a_1	0.55	0.65	1.02	0.65	1.16
Trend	a_2	0.41	-0.08	0.11	-1.43	-0.36
Dewpt	a_3	0.20	0.31	0.29	0.02	0.05
Windrose	a_4	0.77	0.93	0.60	0.42	1.76
Hol	a_5	36.01	31.47	26.47	36.45	28.76

Table A.35 Final cluster centers v^i for NLR-fuzzy system model (Paducah)

Variables	Cluster Center	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Nonlin	v^1	7.20	13.48	18.08	22.51	29.30
Trend	v^2	1.91	2.37	2.56	0.85	2.73
Windrose	v^3	0.13	0.05	0.16	0.19	0.03
Hol	v^4	0.00	0.00	0.00	0.00	0.23

Table A.36 Coefficients a_i for NLR-fuzzy system model (Paducah)

Variables	Coefficient	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
Intercept	a_0	2.00	3.25	7.48	-6.42	0.04
Nonlin	a_1	0.89	0.76	0.56	1.31	1.03
Trend	a_2	-0.04	-0.29	0.27	-0.55	0.02
Windrose	a_3	1.75	1.70	-0.34	2.55	-5.53
Hol	a_4	16.69	19.77	19.70	14.32	14.34

CURRICULUM VITAE

NAME: Yiqiu Lin

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

DOB: Pingdingshan, China – September 5, 1971

EDUCATION
& TRAINING: B.S., Mechanical Engineering
Tianjin University
1989 – 1993

M.S., Mechanical Engineering
University of Louisville
2001 – 2004

Ph.D., Mechanical Engineering
University of Louisville
2004 – 2007

PUBLICATIONS:

Cobourn W. G. and Lin Y. (2004) Trends in Meteorologically Adjusted Ozone Concentrations in Six Kentucky Metro Areas, *1998-2002. Air & Waste Manage. Assoc.*,

Lin Y. and Cobourn W.G. (2007) Fuzzy System Models Combined with Nonlinear Regression for Daily Ground-level Ozone Predictions. *Air & Waste Manage. Assoc.*,