UNIVERSITY OF CALGARY

Statistical Analyses of Ozone Temporal Trends in Calgary, Alberta: an Application of Multivariate Geostatistics

by

Nooryusmiza Yusoff

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE IN CHEMICAL ENGINEERING

DEPARTMENT OF CHEMICAL AND PETROLEUM ENGINEERING

CALGARY, ALBERTA

SEPTEMBER, 2001

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Statistical Analyses of Ozone Temporal Trends in Calgary, Alberta: an Application of Multivariate Geostatistics" submitted by Nooryusmiza Yusoff in partial fulfillment of the requirements for the degree of Master of Science in Chemical Engineering.

Supervisor, Dr. S. Srinivasan
Department of Chemical and Petroleum Engineering

Dr. H. Baheri
Department of Chemical and Petroleum Engineering

Dr. B.Young
Department of Chemical and Petroleum Engineering

Dr. M.A. Maes
Department of Civil Engineering

Date

ABSTRACT

The prediction of tropospheric (surface) ozone episodes is a challenging task that requires the integration of physicochemical and statistical techniques. Governmental agencies such as the U.S. Environmental Protection Agency (EPA) and Alberta Environment favor physicochemical modeling in order to capture the complexity of the underlying physical processes. Unlike physicochemical models, statistical techniques are usually based on spatial and/or temporal correlations between relevant variates. The statistical models also require less exhaustive data sets for accurate predictions; this major advantage is perhaps more obvious when ozone prediction is performed for a longer period of interest.

The primary objective of this research is to investigate statistical techniques for modeling ozone and/or other pollutant concentrations given only sparse environmental records at the monitoring stations. Straightforward linear regression based techniques are implemented initially but the inadequacy of these approaches for predicting detailed temporal ozone variations is verified by the results. Then geostatistical paradigms of kriging and sequential stochastic simulation are implemented to incorporate temporal correlation in the form of variogram. Secondary variables (covariates) can also be useful for providing extra information and their influence is accounted for in cokriging and co-simulation. The positive-definiteness of auto and cross-covariances are ensured via a linear model of coregionalization (LMC). The "two-point" statistic (variogram) is found to be insufficient and hence this thesis strives to explore methodologies for modeling the highly fluctuating temporal profiles with a view to providing a sound framework for subsequent extensions to spatiotemporal modeling.

ACKNOWLEDGEMENTS

Among the most difficult tasks in life is to make the 'right' decision at the 'right' moment, especially when your future depends on it. In the summer 2000, I reached a crossroad: the traditional approaches in Chemical Engineering are on the one route, and on the other lies the random paths, or more specifically, those based on stochastic simulation paradigms.

Professor Sanjay Srinivasan is the one who first prompted and developed my interests in applying statistical methodologies for solving engineering problems, and thus convinced me to endure the challenge of the latter route. Taking his course (Geostatistics for Reservoir Characterization) is indeed an unforgettable experience. Countless hours were spent for 'kriging,' a sophisticated methodology that can be implemented for estimating space-time phenomena. The statistical concepts I learned during the 'long' summer would have been wasted if not extended to practical applications. Thus together, Sanjay and I instigated an interesting research project for analyzing ozone episodes in Calgary, Alberta via stochastic approaches. From that moment on, I have gained much more knowledge than I initially bargained for. Likewise, his vision in planning and directing the research topics, his dedication especially in spending two to three hours on the weekends to assist me with NNFIT and GSLIB, as well as generosity in buying coffee, lunch and dinner are forever cherished.

Throughout my (almost) two-year stay at the University of Calgary, Professors Ayo Jeje and Caroline Hyndman diligently polished my analytical skills in the 'transport' courses. I am truly grateful for having received their insights and constructive criticisms, especially on my class projects. I also consider myself fortunate to have had opportunities to conduct invigorating academic discussions with Arasy Boustani in Mass Transfer, Sergio Guillen in Polymer Engineering, Sergio Merchan in Geostatistics, and Syed Ahmad Imtiaz in Macrotransport Processes. These colleagues of mine eagerly offered invaluable assistances whenever I required them. To my delight, the arrival of the next generation of fellow graduate students, particularly Dhamendra Tiwary, Hamid Bidmus,

Xiaohuan (Peter) Liu, Tarun Kashib and Tao (Tom) Zhang in the Fall 2000 brought about a fresh wave of excitement in the Chemical and Petroleum Engineering Department. One deed I will always remember is the willingness of Huifang Hong, Na Jia and Grace Guo in helping me with grocery shopping at the Real Canadian Superstore. Also not forgotten are two (among others) of my travel companions, i.e., Banky Djamasi and Naveen Jain, with whom I enjoyed a four-day road trip to Vancouver, Victoria and Seattle.

My international journey to Calgary would not have been possible without the financial support from PETRONAS, a Malaysian-owned conglomerate venturing in oil and gas. The advice and encouragements from Mr. Muhamad Jurimi, Dr. Ahmad-Fadzil Mohamad-Hani, Dr. Mohamad-Ibrahim Abdul-Mutalib and Dr. Adilah Abdul-Hamid as well as the assistances from Mr. Mohd-Azminuddin Afendi, Ms. Emy-Hadida Mohd-Nor and Mr. Azhari Abdul-Aziz are greatly appreciated. During the first one-year vocation at the Universiti Teknologi PETRONAS (UTP) in Malaysia, I treasured the friendships of Hasrin Md-Sihap (my ex-roommate), Marfauza Mat-Jusoh and many others, especially those with whom I experienced jungle adventures in Lembah Belum, Perak.

Most importantly, I would like to express the highest gratitude to my family in Malaysia, especially my father Yusoff, mother Aminah and younger brother Nooryusazli, as well as my other siblings and relatives for their motivational support and guidance.

Last but certainly not least, let us begin our quest for knowledge by remembering the wisdom of Imam Ahmad al-Hambali (780-855 A.D.) who contemplated that:

> *"A knowledgeable person who is aware of his knowledge and practices it, follow him.*
>
> *A knowledgeable person who does not practice his knowledge, remind him.*
>
> *A person who realizes that he lacks of knowledge but relentlessly seeking for it, guide him.*
>
> *Someone who is ignorant but pretending to be a knowledgeable person, curse be upon him."*

DEDICATION

"To my beloved fiancée,

Noorasidah Abdul-Rahman,

whose tenderness and patience

I'll always adore and admire ..."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Tropospheric ozone has been determined to be detrimental to public health and welfare. As a secondary pollutant, ozone is not directly released into the lower atmosphere but is formed via complex photochemical reactions of the two main pre-cursors: nitrogen oxides (NO_x) and volatile organic compounds (VOCs), produced by anthropogenic activities such as fossil-fuel combustion and open biomass burning. Regulating the emissions of these pre-cursors may effectively reduce the formation of ozone or more importantly photochemical smog, a well-known agent for respiratory diseases especially in the major urban areas.

Since the inception of an Index Quality of Air (IQUA) by the Federal-Provincial Committee on Air Pollution (1980), ozone has been identified as one of the "markers" or *criteria pollutants* used for assessing the conditions of life and ecosystem in Canada. Understanding the spatiotemporal variations of ozone is an important issue in order to quickly alert the public of the high ozone levels; thus serving as an early warning for public safety and for monitoring the impact of environmental regulations in reducing the concentration level of atmospheric pollutants. However the occurrence of sudden rise in ozone concentrations, termed an ozone episode, is often difficult to predict. Early simulation packages such as the Empirical Kinetics Modeling Approach (EKMA) and Urban Airshed Model (UAM) were developed to tackle this problem. The former is a Lagrangian based model and hence deemed inappropriate for regional ozone modeling. The latter, a 3D Eulerian based model, is also not without a few shortcomings: (1) it requires very small (~hourly) time steps for accurately solving the atmospheric transport equations due to the complex reaction kinetics of the pollutant species, and (2) exhaustive meteorological and environmental data are needed for simulating at the most five-day

ozone episodes. Physicochemical models are therefore inappropriate for modeling annual ozone concentration profiles involving arbitrarily sampled data.

Statistical methods provide an alternative to detailed physicochemical modeling and can eliminate some of the drawbacks of the process based modeling approaches. Here the ozone concentration at a particular instant in time and location in space is considered as an outcome of a spatiotemporal random variable (RV). Secondary information in the form of covariates is treated as auxiliary information for modeling the probability distribution underlying the RV (ozone concentration). Once historical data have been matched and the parameters of the statistical model calibrated, the resultant model can be utilized to predict ozone concentrations given only "a few" sample data at different time instants and/or locations. The simplest statistical technique would be to apply multivariate linear regression on the relevant variables, i.e., ozone (O3), dust and smoke (COH), carbon monoxide (CO), nitric oxide (NO), nitrogen dioxide (NO2), total hydrocarbon (THC), wind speed (WSPD), daily average temperature (Tavg), relative humidity (RHavg), and also bright sunshine hours (bSUN), jointly occurring in space and time. Nevertheless the non-linearity of the physicochemical phenomena influencing ozone concentrations deems the linear regression technique inappropriate. On the other hand, nonlinear and tree-based regression approaches can be adopted but they require *a priori* knowledge of the multivariate probability distribution underlying the stochastic RV and therefore may suffer from lack of generality when applied to the "uncharted" regions, especially those corresponding to unusual meteorological conditions.

The traditional statistical notion of linear regression neglects redundancy between data. The data-to-estimate covariance is accounted for in regression but not the data-to-data covariance that measures redundancy between data. This shortcoming of traditional linear regression is rectified by using the geostatistical paradigm of kriging. Variable interactions in space and/or time are accounted for by a "two-point" statistic known as a variogram. The kriged estimates represent the mean or expected value of the conditional probability distribution underlying the random variable at time instant t and location $\mathbf{u}$. Similar to the linear regression estimate, the kriged profile is a smooth representation of the ozone trend. The kriged estimates reproduce the data-to-unknown covariance and the

histogram of the samples. In addition, kriging is data-exact. However kriging can at very best predict the smooth ozone trends. Patterns of ozone variation in the form of ozone concentration fluctuations cannot be reproduced by kriging. The remedy is to employ stochastic simulation that adds a spatiotemporal residual component to the kriged estimate to correct for the reduction in variance and to impart the "true" pattern of spatiotemporal variation.

The stochastic simulation approach has been employed by many researchers, including Kyriakidis (1999) who investigated the space-time phenomena of monthly average sulfate deposition over several European countries. Following a two step approach, the temporal profiles were initially parameterized using a deterministic model in the form of Fourier series, and in the second step the parameters were regionalized in space to probabilistically obtain the corresponding coefficients of the Fourier series at unsampled locations. The accuracy of prediction using this technique requires good knowledge of the temporal variability at a number of monitoring stations. The appropriateness of the models of temporal variability (such as the Fourier series approach in the previous works) has to be verified prior to embarking on the subsequent step of spatial modeling. For this reason, a detailed investigation of the temporal ozone phenomena in Calgary, Alberta is the focus of the current research. The research will endeavor to identify a suitable methodology for modeling temporal variability of ozone concentration and then make recommendations for extending the temporal model to the selected spatiotemporal problems.

This thesis begins with extensive reviews of the available literature in Chapter 2. The background of the problem and overview of ozone photochemistry are initially discussed to gain familiarity with the subject. Then the traditional approaches of ozone modeling based on physicochemical models employed by regulatory agencies are summarized. Statistical methodologies for modeling atmospheric phenomena are assessed for their suitability of implementation in this work.

In Chapter 3, the CASA data sets are introduced, and the environmental and meteorological variables are presented. The annual trends of these predictor variables

(covariates) as well as ozone are explored in the form of time-series plots. Based on the suggestions from literature and data availability in Calgary, the rationales for selecting covariates and temporal scale are justified. Then the method of data standardization is elaborated, and using these standardized data, the distributions of all variables are illustrated in the form of box plots.

In Chapter 4, the annual trends of ozone are preliminarily examined via simple linear regression using historical ozone data. Since this approach is clearly inadequate to emulate the trends based on sparse ozone data, secondary information is introduced into a multivariate regression framework. The bivariate relations between the selected covariates and ozone are evaluated in the form of correlation coefficients $\rho_{\alpha\omega}$, and their similarities with ozone are presented via scatter and quantile-quantile (Q-Q) plots.

In Chapter 5, the basic concepts of geostatistics are discussed and the stochastic framework for temporal modeling of ozone concentration is clarified. As a start, the theoretical aspects of variogram, a measure of temporal correlation, and the techniques for ensuring positive-definite variogram modeling are described. This "two-point" statistic is later applied to a generalized regression technique, commonly known as kriging, for ozone prediction. To investigate the effects of covariates on ozone formation, cokriging is performed using the variogram model obtained via the linear model of coregionalization (LMC).

In Chapter 6, the paradigm of stochastic simulation is introduced. In particular, sequential Gaussian simulation is implemented corresponding to different data sampling scenarios. Several realizations of the simulated results using evenly spaced data are compared with those randomly selected between the 25^{th} and 30^{th} day of the month. The periodicity in the long-term variations of ozone concentration is modeled using the hole-effect variogram model, which is subsequently used for simulating the ozone variations over a four-year period. The assessment of covariate influence on ozone phenomena is performed via co-simulation utilizing the LMC model obtained in the similar manner as the cokriging case.

In Chapter 7, other statistical methods for noise filtration are implemented. Here the random fluctuations of ozone signals are filtered out via Fourier series analysis (FSA) to obtain a smooth trend for each year. In order to solve for a system of large matrices containing sine and cosine series as well as the Fourier coefficients, a technique called singular value decomposition (SVD) is utilized. The temporal variability of ozone concentrations over four years is compared by examining the similarity between the corresponding coefficients. Next the applicability of a neural network for modeling the highly nonlinear and complex interactions between ozone and its covariates is explored.

In Chapter 8, the general conclusions of this work are presented. The limitations and potential applications of the previously discussed approaches are highlighted. The results of the current research are discussed in the context of temporal modeling of the ozone phenomena. New methodologies for improving ozone prediction are suggested as part of future research avenues.

Finally, two appendices are included for showing examples of GSLIB parameter files (Appendix A), and discussing the theoretical and practical aspects of the physico-chemical model, in particular the Urban Airshed Model (UAM) (Appendix B). The detailed explanation of the latter appendix is intended for addressing its importance in ozone simulation as well as its feasibility to be coupled with statistical approaches as a mechanism for training statistical models.

CHAPTER 2

LITERATURE REVIEW

This chapter begins with general discussions on the background of tropospheric ozone problem and relevant photochemistry. A brief overview of physicochemical (box and trajectory) models, previously employed by regulatory agencies (e.g., US EPA) in air pollution abatement, is highlighted to recognize their importance in air quality modeling. The use of the above physicochemical models for ensuring "attainment" of the environmental policies has been superseded by the more versatile Urban Airshed Model (UAM: SAI, 1999, Appendix B). More recently, statistical modeling has become popular due to lesser requirement with regards to the input data sets. Advanced statistical analyses, if used correctly, can complement the physicochemical models in a way so as to reduce the computing time without sacrificing the accuracy of ozone prediction. As the focal point of this thesis, statistical approaches, or more specifically regression and space-time modeling, are also reviewed extensively.

2.1 Background

Smog (derived from **sm**oke and f**og**) has been the cause of adverse health effects since it was first recorded in Los Angeles (ca. 1942) and then London (ca. 1952); thousands of people were hospitalized due to nose, eye and throat irritations. At higher concentrations, smog may cause severe respiratory problems, e.g., asthma and bronchitis (British Columbia Ministry of Environment, Lands and Parks, 1992). In Canada, smog and other air pollution problems commonly occur in major cities like Vancouver and the urban belt ranging from Windsor to Quebec City due to dense population and industrialization. Recognizing the need to increase public awareness on environmental issues, the Federal-

Provincial Committee on Air Pollution (Environment Canada, 1980) introduced the Index Quality of Air (IQUA), calculated using the following equation (Nkemdirim, 1988):

$$IQUA = [O_3]^{1.3} + [NO_2] + [CO]^{1.05} + 10[COH]^{1.2} + 0.33[SO_2]^{1.55} \quad (2.1)$$

The concentrations (denoted by the square brackets) of gaseous pollutants are expressed in parts per million by volume (ppmv) and coefficient of haze (COH) measures the amount of dust and smoke (i.e., particulate matter) in the unit of fractional transmittance of light. In plain words, the IQUA indicates the seriousness of air pollution due to the presence of five so-called *criteria pollutants*: ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO), dust & smoke (COH) and sulfur dioxide (SO_2). The concentrations and amount of those pollutants are converted to a single number with a matching description; an IQUA rating of 0-25 indicates **Good**, 26-50 **Fair**, 51-100 **Poor**, and greater than 100 **Very Poor** air quality. Table 2.1 illustrates more descriptive meanings of these numbers.

Table 2.1
Index Quality of Air (IQUA).

IQUA Rating	Frequency in Alberta	Effects
Good	Almost all the time	Desirable range: No known harmful effects to soil, water, vegetation, animals, materials, visibility or human health.
Fair	Occasional (typical when weather conditions prohibit pollution dispersion	Acceptable range: Adequate protection against harmful effects to soil, water, vegetation, animals, materials, visibility and human health.
Poor	Very seldom	Tolerable range: Not all aspects of the environment are protected from possible adverse effects. Long-term control action may be necessary depending on the frequency, duration and circumstances of the readings.
Very Poor	Very rare	Intolerable range: In this range, further deterioration of air quality and continued high readings could pose a risk to public health.

Adapted from Environment Canada (1993).

The first *criteria pollutant* (i.e., O_3) instigates high interest in further research studies because of two reasons: (1) O_3 is one of the main elements (beside hydrocarbons) in the smog formation, and (2) O_3 is also a *secondary pollutant*, i.e., it is not directly emitted from anthropogenic or biogenic sources and hence its formation may be avoided by curbing the emissions of its precursors. In fact, O_3 only appears in the lower atmosphere (troposphere) as a result of photochemical reaction between nitrogen oxides (NO_x), volatile organic compounds (VOCs) and their derivatives. In an effort to control the smog problem, the Alberta Environment (a provincial regulatory body), under Section 14 of the Environmental Protection and Enhancement Act (EPEA), sets mandatory guidelines for the maximum ambient concentrations of O_3 and various other pollutants. The one-hour average concentration limit for tropospheric O_3 is 160 micrograms per cubic meter ($\mu g/m^3$), or 82 parts per billion by volume (ppbv) after approximate conversion at standard conditions of 25°C and 101.325 kPa (Alberta Ambient Air Quality Guidelines: AAAQG, 2000). For convenience, the following equations describe the standard procedure for unit conversion (Flagan and Seinfeld, 1988; pg. 5):

$$[\text{ppmv}] = \frac{C_i}{C_{air}} \times 10^6 \quad \text{O} \quad [\text{ppbv}] = \frac{C_i}{C_{air}} \times 10^9 \qquad (2.2)$$

r

where C_i and C_{air} are the respective concentrations of species i and air in moles per volume (molar), at specific temperature T and pressure P. The pollutant concentration in ppmv (or ppbv) can be easily converted to $\mu g/m^3$ using the ideal gas law, and assuming standard conditions (T and P) as specified by the regulatory agencies,

$$\left[\frac{\mu g}{m^3}\right] = \frac{PM_i}{RT} \times [\text{ppmv}] \qquad (2.3)$$

where M_i is the molecular weight (in grams per mole or g/mol) of species i and R is the universal gas constant in appropriate units (e.g., R = 8.3144 Joules per mole per Kelvin or J/mol·K).

As noted above, ozone is produced in the troposphere from the photochemical reactions between NO_x and VOCs. The complexity of the nonlinear associations between ozone and its precursors (i.e., NO_x and VOCs) poses a challenging research project, at least from the statistical point of view. By applying stochastic analyses (e.g., sequential Gaussian simulation), we may improve understanding of urban ozone episodes from such studies as (1) characterizing the temporal ozone variations and (2) exploring the influence of precursors on ozone formation before proceeding to study the space-time phenomena. In addition, the underlying physicochemical mechanisms need to be conceptually grasped before we can make any sensible interpretation using statistical methodologies.

2.2 An Overview of Ozone Photochemistry

Although the smog problem has been drastically reduced through strict regulatory measures and technological advancements, it still persists especially in large urban areas mainly due to high pollutant emissions from the increased number of fossil-fuel-powered vehicles. It is well known that the combustion of fossil fuels, e.g., gasoline and diesel, produces among others nitrogen oxides NO_x (mainly NO and NO_2) and volatile organic compounds (VOCs). Previous studies (e.g., Fishman et al., 1979; Campbell, 1986; Liu et al., 1987; NRC, 1991) have shown that in the presence of "bright" sunlight, ozone is formed from the complex photochemical reactions among NO_x, VOCs (the ozone precursors), and their derivatives. As simplified by de Nevers (1995), the photochemical reactions start with the decomposition (also termed photolysis) of nitrogen dioxide (NO_2) by the high intensity of solar radiation ($h\nu$) to nitric oxide (NO) and oxygen radical (O; note that the omission of 'dot' for the radical is common in the air pollution literature),

$$NO_2 \xrightarrow{h\nu} NO + O \tag{2.4}$$

The highly reactive oxygen radical (O) immediately attacks an oxygen molecule (O_2) in the presence of another molecule (M; usually in the form of nitrogen N_2 or another oxygen O_2), which absorbed some of the energy released from the following reaction,

$$O + O_2 \xrightarrow{M} O_3 \tag{2.5}$$

To complete the cycle or restore the equilibrium of natural processes, another product (i.e., NO) from the decomposition of nitrogen dioxide (see reaction 2.4) reacts with ozone to reproduce the starting materials,

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{2.6}$$

At steady state, the rates of all three reactions (2.4-6) are equal and the O_3 concentration (denoted by square brackets $[O_3]$) can be solved; after simplifying the kinetics (i.e., reaction rates, orders, etc.), the final result is:

$$[O_3] = \frac{[h\nu][NO_2]}{k_o[NO]} \tag{2.7}$$

where $[h\nu]$ is the solar intensity in appropriate unit and k_o is the rate constant for reaction (2.6); [NO] and $[NO_2]$ are the concentrations of nitric oxide and nitrogen dioxide, respectively. This preliminary result suggests that, in the absence of VOCs, ozone formation is strongly dependent on the intensity of solar radiation due to the canceling off effect of the NO_2 and NO (i.e., NO_x).

Nowadays, more VOCs are emitted from anthropogenic sources, mainly from the combustion of fossil fuel. The VOCs interfere with the above reactions in such a way that NO is oxidized to NO_2 without depleting O_3. Hence the level of tropospheric O_3 increases, which subsequently induces the formation of photochemical smog. To further understand this process, a mechanism on how the VOCs interfere with these reactions was proposed by Campbell (1986):

$$OH + VOC \rightarrow RO_2 + H_2O \tag{2.8a}$$

$$RO_2 + NO \rightarrow NO_2 + RO \tag{b}$$

$$RO + O_2 \rightarrow RCHO + HO_2 \tag{c}$$

$$HO_2 + NO \rightarrow NO_2 + OH \tag{d}$$

where R represents any carbon-hydrogen bonds. Clearly, the free radicals (in the forms of hydroxy OH and its derivatives RO, RO_2 and HO_2) play important roles in the VOC chemistry. Without dwelling too much into the complexity of reaction kinetics, these four reactions (2.8a-d) can be summarized as a single overall reaction,

$$VOC + 2NO + O_2 \rightarrow H_2O + RCHO + 2NO_2 \qquad (2.9)$$

where O, C and H are customarily left unbalanced due to the complex nature of VOC and aldehydes (RCHO). It is obvious that NO is consumed in the VOC-oxidation process and therefore cannot be an effective agent for ozone reduction. Later on, other researchers (e.g., Milford et al., 1992; Gao et al., 1995, 1996; Yang et al., 1995, 1996; Vuilleumeir et al., 1997; Bergin et al., 1998) extended research on the reaction mechanisms and went a step further by performing uncertainty/sensitivity analysis on the photochemistry. They identified nitrogen dioxide (NO_2) and aldehydes (RCHO) as the two most important chemical species in ozone formation.

In a related work, Dickerson et al. (1997) showed that the UV-scattering aerosols (e.g., sulfates) could increase the rate of ozone formation whereas the UV-absorbing particles (e.g., soot) decrease it. Vuilleumier et al. (2001) studied how solar irradiance (i.e., uncalibrated quantity of radiation) relates to total atmospheric optical depths (i.e., the solar direct beam differential extinction rate per unit of vertical path length) in order to improve simulation of photochemistry in air quality (physicochemical) modeling. To localize the factors influencing the optical depth variability, the authors applied principal component analysis (PCA) on the simultaneous measurements of optical depths at seven wavelengths (λ = 300, 306, 312, 318, 326, 333 and 368 nanometers, nm); they found that "light absorption and scattering by aerosols as the major factor, and absorption by ozone as the minor factor." In the long run, this contribution was hoped to enrich knowledge on the air quality modeling treatment of the solar actinic[1] flux, which is one of the governing factors in the photochemical reactions.

[1] Actinic = causing chemical changes.

2.3 Physicochemical Models

Air pollution, if left uncontrolled, can cause detrimental effects to the public (health and welfare), materials and vegetation. Initially it was thought of as a local phenomenon, thus not treated seriously. Since the late 1960s, however, air pollution has been identified as both a regional and a global problem, and thus requires long-term solutions, especially in controlling the emissions of primary pollutants and processes responsible for producing secondary pollutants. Regulatory agencies (e.g., Alberta Environment) are well aware of these problems; therefore, strict environmental policies (e.g., AAAQG) were introduced to curb excessive emissions of the pollutants from point (e.g., tall stacks), line (e.g., highways) and area sources (e.g., a cluster of gas stations in the downtown area). On many occasions, the regulatory policies are formulated partly based on the results from the physicochemical models, which are the mathematical descriptions of the atmospheric transport (dispersion, diffusion and/or surface removal) and the relevant chemical reactions. In general, there are three types of air quality models used in the practice of air pollution control: (1) box, (2) trajectory, and (3) grid models.

2.3.1 Box Models

These models are the simplest of all. Basically, a region to be simulated is treated as a single cell (box) bounded by the ground at the bottom, the upper boundary layer (usually mixing height) on the top and arbitrarily fixed lateral boundaries. The mixing height is often less than a kilometer and the area that the box encloses may be a few hundred square kilometers (Seinfeld, 1988). The fundamental theory of these models relies on the 'perfectly stirred reactor' assumption where the emissions of primary pollutants at various locations are treated as spatially uniform due to instantaneous mixing; that is the pollutant concentrations in the air within the box volume are homogeneous. The only inward and outward flow mechanisms, i.e., bulk transport, are represented by the horizontal wind components and rate of convection to the mixing height. One typical way to describe the model is by the following equation (Sportisse, 2001), which is rewritten using the familiar notation of this thesis,

$$\frac{\partial \overline{C}_i}{\partial t} = \frac{1}{H(t)}\left[S_i(t) - u_{dep}^i \overline{C}_i\right] + R_i + E_i(t), \qquad \forall i = 1, ..., N \tag{2.10}$$

with an initial condition $\overline{C}_i(t) = \overline{C}_{io}$; the bar above C_i represents the spatially average concentration of species i due to the instantaneous mixing assumption. The height of mixing H is spatially homogeneous and only varies with time. E_i is the emission flux from, say, industrial plants and S_i describes the injection rate from elevated surfaces (layers), which depends on the growth of atmospheric boundary layer and only a function of time (Zannetti, 1990, Van Loon, 1996; Stull, 1988;). The dry deposition velocity (also termed surface removal) is denoted by u_{dep}^i, and R_i is the rate of generation (or depletion) by chemical reactions, which is a function of concentration ($\overline{C}_i$), temperature (T) and time (t).

In more realistic processes, pollutant concentrations vary with space and time. For example, vehicular emissions are higher in downtown than in the suburban areas or industrial zones are more polluted than family residence, and more pollutants are released or produced during the day than at nighttime. Thus the treatment of pollutant emissions as spatially homogeneous by the box models is clearly misleading. At best, box models can forecast the average temporal phenomena of each pollutant species i. In other words, box models cannot effectively simulate the complex air pollution episodes, e.g., predicting the location-specific profiles of the maximum ozone concentrations, due to simplistic assumptions of real life processes. Hence they are not generally implemented in air pollution controlling strategies.

2.3.2 Trajectory Models

A better approach in physicochemical modeling is to apply trajectory models. Based on the theory of mass conservation, the transport of air pollutants and relevant chemical reactions are treated in moving coordinates (hence termed Lagrangian), which are anchored to a hypothetical or fictitious vertical air column bounded by ground at the bottom and inversion base, if exists, on the top. Usually, the starting point is specified at

the city limit and the air column moves under the influence of horizontal wind fields. When it passes through emission sources, e.g., coal-fired utilities, the pollutants are "injected" into the column and chemical reactions occur simultaneously.

However, it is perhaps easier to understand the mass conservation problems in Cartesian coordinates before we proceed to the Lagrangian frame of reference. In a simple case, Walters (1969) studied the continuous surface line source by assuming uniform wind speed U in the x-direction. The corresponding horizontal and vertical dispersion coefficients K_H and K_V were chosen to be invariant with the wind direction and only a function of altitude (z); that is $K_H = \overline{K}_o z$ and $K_V = \overline{K}_1 z$ where $\overline{K}_o$ and $\overline{K}_1$ are average values associated with the horizontal and vertical directions, respectively. The governing mass balance equation was given as the following:

$$U\frac{\partial C_i}{\partial x} = K_H(z)\frac{\partial^2 C_i}{\partial x^2} + \frac{\partial}{\partial z}\left[K_V(z)\frac{\partial C_i}{\partial z}\right], \quad \forall i = 1, \ldots, N \tag{2.11}$$

where $C_i(x, z)$ denotes the concentration of species i associated with the spatial process; the solution to this problem was obtained as:

$$C_i(x, z) = \frac{q_L}{\overline{K}_1(1 + e^{-\lambda\pi})[x^2 + (\mu z)^2]^{1/2}} \exp\left[-\lambda \tan^{-1}\left(\frac{\mu z}{x}\right)\right] \tag{2.12}$$

where q_L denotes the pollutant flux for a continuous line source; $\lambda = U/(\overline{K}_o \overline{K}_1)^{1/2}$ and $\mu = (\overline{K}_o / \overline{K}_1)^{1/2}$.

In another study, Liu and Seinfeld (1975) improved the model by including the temporal component of the concentration, i.e., $C_i(t)$, and derived the analytical solution in terms of gamma (Γ) and exponential functions. If K_H is considered to be constant and $K_V(z)$ is represented as a power law function of altitude (z), i.e., $K_V = \overline{K}_1 z^m$, where $\overline{K}_1$ is an average value associated with the vertical dispersion coefficient and m $\in$ (0, 2), the solution can be obtained as:

$$C_i(x,z,t) = \frac{1}{(2-m)^{m/(2-m)}\Gamma[1/(2-m)]} \int_0^t \int_0^L \frac{[4\pi K_H(t-\beta)]^{-1/2}}{[\overline{K}_1(t-\beta)]^{1/(2-m)}} \cdot$$

$$\exp\left[\frac{z^{2-m}}{(2-m)^2 \overline{K}_1(t-\beta)}\right] \exp\left\{-\frac{[x-U(t-\beta)-\alpha]^2}{4K_H(t-\beta)}\right\} q_A \, d\alpha \, d\beta \qquad (2.13)$$

where $q_A(\alpha, \beta)$ is the space-time area source flux; α and β are the spatial and temporal dummy variables, respectively.

In a similar case but now considering the trajectory model, the solution requires prior coordinate transformation, and assumptions that the horizontal dispersion and wind shear (vertical component) can be neglected. The governing equation (2.11) may be rewritten in the Lagrangian coordinate as:

$$\frac{\partial \widetilde{C}_i}{\partial \tau} = \frac{\partial}{\partial \rho}\left(K_V \frac{\partial \widetilde{C}_i}{\partial \rho}\right) \qquad (2.14)$$

Note that horizontal dispersive term is eliminated from the mass balance equation due to neglecting horizontal mixing across the boundaries of the air column (i.e., $K_H = 0$). The downwind distance x in the convective term is converted into traveling time τ, i.e., $x = U\tau$ since the velocity U is assumed uniform. Hence the pollutant concentration associated with trajectory model, denoted by the tilde (~), only varies with the distance ρ above the ground (i.e., vertically) and traveling time τ. Eqn. (2.14) is subject to the initial condition:

$$\widetilde{C}_i(\rho, 0) = Q_A \delta(\rho) \qquad (2.15)$$

where Q_A is the instantaneous emission source, which value exists only at altitude ρ as ensured by the Dirac delta function δ. When coupled with the boundary conditions of no flux at the ground ($\rho = 0$) and zero concentration aloft, i.e.,

$$-K_V \frac{\partial \widetilde{C}_i}{\partial \rho} = 0, \quad \rho = 0 \tag{2.16a}$$

$$\widetilde{C}_i(\rho, \tau) = 0, \quad \rho \to \infty \tag{b}$$

the solution for the trajectory model (area source) is obtained as

$$\widetilde{C}_i(\rho, \tau) = \frac{Q_A}{(2-m)^{m/(2-m)}\Gamma[1/(2-m)](\overline{K}_1\tau)^{1/(2-m)}} \exp\left[-\frac{\rho^{2-m}}{(2-m)^2 \overline{K}_1 \tau}\right] \tag{2.17}$$

after simplifying the vertical dispersivity $K_V(z)$ with the power law representation as in the case of Cartesian coordinate above.

The validity of implementing trajectory models in a full 3D simulation may be questioned because three major assumptions were made: (1) the horizontal mixing across the air column is neglected, (2) the movement of the column is treated only in 2D (i.e., horizontally) by omitting the influence of vertical wind component (w), and (3) the whole parcel of air moves with a wind speed that is invariant with altitude, which means that the air column is assumed vertically straight at all time. These simplifications may cause large error, e.g., when the air column passes in the proximity of a major area source but not over it, the horizontal pollutant dispersion due to such source will be neglected in the model calculation. A poor result is expected, especially in the direction orthogonal to the movement of the air column.

Under optimal conditions (sufficiently high horizontal but low vertical wind speeds), Liu and Seinfeld (1975) found the absolute error of the trajectory model to be less than 10 percent when it was compared with the "exact" solution. Consequently, in the mid 1970s, the U.S. Environmental Protection Agency (EPA) proposed the use of Empirical Kinetics Modeling Approach (EKMA), a Lagrangian model, for estimating the effects of volatile organic compounds (VOCs) and nitrogen oxides (NO_x) on urban ozone episodes. However, 3D photochemical grid models (termed Eulerian due to assumption of fixed coordinates) such as the Urban Airshed Model (UAM) have superseded the

Lagrangian models for regulatory purposes. A detail description of the UAM is included in Appendix B of this thesis.

2.4 Statistical Approaches

Besides the physicochemical modeling approaches, the forecasting (in temporal sense) of ozone phenomena in the urban area can be accomplished by means of statistical methodologies. Recent advancements in theory as well as computing technologies encourage researchers to seek statistical representation of ozone data. For example, the analysis of complex nonlinear relationships between ozone and its precursors using artificial neural networks (e.g., Guardani et al., 1999; Gardner and Dorling, 1998-2000; Prybutok et al., 2000) can be achieved in much shorter time than running a 3D photochemical grid model (e.g., UAM) for obtaining similar results. The reason is that the neural network acts as a proxy to the complex physicochemical transfer functions, which are transformed into algebraic expressions that are lighting fast to compute. The target output (i.e., ozone concentration) is obtained by minimizing global error through efficient optimization techniques, e.g., the scaled conjugate gradient algorithm. In general, the statistical methodologies currently applied in the context of atmospheric science can be classified into two broad categories: (1) regression, and (2) spatiotemporal modeling.

2.4.1 Regression

This approach and its derivatives are perhaps the most commonly applied in the field of ozone level prediction (in temporal sense) and/or estimation (in spatial sense). With some twist in the method complexities, many models for the average behavior of ozone, its precursors and relevant meteorological variables exist. The regression-based methods can be further divided into three groups as suggested by Thompson et al. (2001): (1) linear regression, (2) nonlinear regression, and (3) regression tree analysis.

2.4.1.1. Linear Regression

Due to its simplicity, linear modeling is perhaps the most popularly implemented in predicting ozone episodes based on concurrent meteorological conditions. Examples of such works can be found in Feister and Balzer (1991), Korsog and Wolff (1991), Abdul-Wahab et al. (1996), Katsoulis (1996) and Fiore et al. (1998). Some researchers (e.g., Hastie and Pregibon, 1992) argued that many processes have Gaussian (normal) errors and a linear model provides an optimal estimate for the expected value of such problem, at least within a limited interval. Other phenomena exhibit nonlinear behaviors but often can also be modeled linearly by scale-transforming the response variable and predictors. However, such transformations only alter the univariate distribution characterizing the random variable RV $Z(t)$ without affecting the multivariate distributions of a series of RV $\{Z_i(t), i = 1, \ldots, N\}$ taken jointly at all times.

In an example of univariate transformation, Turner (1970) correlated experimental data of the transverse dispersion coefficient σ_y as a function of downwind distance x at various stability categories (A-F). When the variables were logarithmically transformed, he discovered linearity between σ_y and x, which was contrary to the theoretical hypothesis of the form $\sigma_y \sim x^{1/2}$ (see for example, de Nevers, 1995). Hence it should not be a surprise when the EPA adopted Turner's proposal; that is the simple linear relation between the logarithmically-transformed dispersion coefficient σ_y and downwind distance x in the Gaussian plume model should be applied for the air pollution abatement strategies. However, we must proceed with care when dealing with more complicated transformations because physical interpretations in the "unnatural" scales may be misleading.

Consider a classic linear model, which takes the form of:

$$y_i = f(x_i) + \varepsilon, \quad i = 1, \ldots, N \tag{2.18}$$

where y_i are the predicted values of the response Y and f is a function of predictor X with observations x_i. In other words, the error ε is assumed to be Gaussian with zero mean and constant variance. For univariate cases, e.g., forecasting maximum ozone concentrations

(i.e., response Y, possibly in a transformed scale) based on historical ozone data (i.e., predictor X, which may have been transformed prior to analysis), the model can be written as:

$$y_i = a + bx_i, \quad i = 1, \ldots, N \tag{2.19}$$

where a and b are the coefficients of linear best fit intercept and gradient (slope) of the model, respectively.

Unfortunately, univariate analyses often fail to satisfactorily predict future ozone concentrations due to the complexity of the phenomena. This problem can be tackled partly by including the influence of ozone predictors in the model, i.e., by applying multivariate linear analyses. In one study, Chaloulakou et al. (1999) investigated how maximum ozone concentrations relate to meteorological and chemical covariates (predictors). Using a seven-year period (1987-1993) of data sets provided by the Greek Ministry of Environment, City Planning and Public Works (PERPA), the authors applied multivariate linear regression to predict ozone concentrations in 1993. Their approach can be divided into three cases: (1) bivariate analyses of ozone concentrations $[O_3]$ using previous day ozone concentrations $[O_3]_{pd}$ and maximum temperature T_{max}, (2) similar to the former, except that T_{max} is substituted with inverse wind speed WS^{-1}, and (3) multivariate analysis using six cofactors, i.e., $[O_3]_{pd}$, T_{max}, WS^{-1}, dominant wind direction WD, and the concentrations of nitrogen dioxide $[NO_2]$ and carbon monoxide [CO]. For convenience, the results are tabulated in Table 2.2 according to the order of the above cases and the naming conventions used by the authors:

Table 2.2
Models and results of the multivariate linear analysis from the work of Chaloulakou et al. (1999).

	Models	Equations and Results
1.	TEMPER	$[O_3] = a + b*[O_3]_{pd} + c*T_{max}$
		where,
		$a = 43.079$, $b = 0.507$, $c = 1.255$
		$R = 0.54$, $R^2 = 0.29$
2.	WISPER	$[O_3] = a + b*[O_3]_{pd} + c*WS^{-1}$
		where,
		$a = 35.697$, $b = 0.478$, $c = 123.719$
		$R = 0.61$, $R^2 = 0.37$
3.	REGLIN6	$[O_3] = a + b*[O_3]_{pd} + c*T_{max} + d*WS^{-1} + e*WD + f*[NO_2] + g*[CO]$
		where,
		$a = -55.374$, $b = 0.348$, $c = 1.835$, $d = 3.002$, $e = 111.323$, $f = 0.210$, $g = 1.928$
		$R = 0.66$, $R^2 = 0.43$

where a, b, c, d, e, f, and g are the fitted coefficients of the multiple linear regression models. R and R^2 denote the coefficients of correlation and determination, respectively. Among the three regression models, the most reliable ozone-forecasting equation was the REGLIN6, as evidenced from the highest values of R and R^2. However, the authors admitted that the multiple linear regression models could not successfully handle extreme value cases due to their nature; the assumed simple linear and/or additive associations between predictors, i.e., meteorological and chemical data, and response (ozone) were clearly inadequate to capture the nonlinearity in the underlying physical and chemical mechanisms of these phenomena. Bloomfield et al. (1996) also made the same remark and tried the next logical step; that is modeling meteorologically dependent ozone episodes in Chicago area over an eleven-year period (1981-1991) with nonlinear regression techniques.

2.4.1.2. Nonlinear Regression

The next level in model complexity is the nonlinear regression approach. In the study of the meteorological effects on ozone episodes, many workers have acknowledged the nonlinear associations between the response variable (ozone) and predictors (e.g., surface temperature, wind fields). For example, Bloomfield et al. (1993a,b and 1996) correlated ozone concentration [O_3] with a cubic polynomial of temperature T, and a simple function of surface and upper atmospheric (at 700 mbar) wind speeds, respectively, *WSPs* and *WSP*700, as follows

$$[O_3] \sim \frac{poly(T,3)}{1 + \frac{WSPs}{v_s} + \frac{WSP700}{v_{700}}} + \{other\ terms\} \tag{2.20}$$

where v_s and v_{700} are the corresponding fitted wind speed parameters at the surface and upper atmosphere. The seasonal patterns and trend are included in the {*other terms*}, and tailored via a "short" Fourier series (see detailed descriptions in Chapter 7) of the form, e.g., $[a_1 \cos(2\pi t) + b_1 \sin(2\pi t)]$ and $[a_2 \cos(4\pi t) + b_2 \sin(4\pi t)]$ corresponding to the annual and semi-annual frequencies. The a's and b's are the fitted coefficients, and the time (year) variable t was scaled to [Julian year + (Julian day/365) – 1985] for easier analysis. Based on this model, they found that the R^2 value (a measure of nonlinear least square fit) increased to 0.8037 when compared to using the wind speed data alone where R^2 was only 0.7204 or combining the effects of temperature, relative humidity and wind speed where R^2 was 0.7499.

Encouraged by these results, Soja and Soja (1999), and Cobourn and Hubbard (1999) applied similar approaches in the areas of eastern Austria and the Ohio River Valley (U.S.A), respectively. The former employed daily maximum temperature and sunshine duration based on the data sets collected in three-year period (1993-1995). The regression models were given for individual months during the ozone 'season' (May-September) and compared with bivariate linear regression technique ($[O_3] = a + b*T_{max} + c*SUN$) where a, b and c were the fitted coefficients; [O3], T_{max}, SUN are the ozone concentration, maximum temperature and sunshine hours, respectively. For convenience,

the model equations for the case of seven-hour (0900-1600) mean values are tabulated below in Table 2.3:

Table 2.3
Models and results of nonlinear regression analyses from the work of Soja and Soja (1999)

	Months	Model Equations and Coefficients
1.	May	$[O_3] = a + b*\exp(-T_{max}/c)*\exp(-SUN/c)$
		where: $a = 29.2, b = 2.15, c = -13.6$
		$R^2 = 0.716, (R^2)_{blm} = 0.670$
2.	June	$[O_3] = a + b*(T_{max})^{1.5} + c*SUN^{0.5}$
		where: $a = 9.30, b = 0.210, c = 6.00$
		$R^2 = 0.739, (R^2)_{blm} = 0.733$
3.	July	$[O_3] = a + b*(T_{max})^{2.5} + c*SUN^{0.5}$
		where: $a = 15.6, b = 0.00617, c = 4.93$
		$R^2 = 0.641, (R^2)_{blm} = 0.635$
4.	August	$[O_3] = a + b*(T_{max})^{2.5} + c*SUN*\ln(SUN)$
		where: $a = 18.0, b = 0.00720, c = 0.254$
		$R^2 = 0.843, (R^2)_{blm} = 0.824$
5.	September	$[O_3] = a + b*T_{max}*\ln(T_{max}) + c*SUN$
		where: $a = 9.55, b = 0.291, c = 1.527$
		$R^2 = 0.755, (R^2)_{blm} = 0.755$
6.	All months	$[O_3] = a + b*(T_{max})^{2.5} + c*SUN$
		where: $a = 22.0, b = 0.00505, c = 1.491$
		$R^2 = 0.735, (R^2)_{blm} = 0.724$

where the R^2 denotes the coefficient of determination (a measure of prediction variance) and the subscript 'blm' refers to bivariate linear model. Interestingly, the R^2 for the linear model were close to those of the nonlinear models, which means that this approach only serves to predict the best fitted values of ozone concentrations. The complex

relations between ozone and its predictors were difficult to be physically understood, as evidenced from the inconsistency of model representations.

Cobourn and Hubbard (1999) investigated ozone phenomena in the Ohio River Valley using the EPA's Aerometric Information and Retrieval Systems (AIRS) database for a period of five years (1993-1997). The meteorological variables were averaged at several hourly intervals, i.e., cloud cover CC (0900-1400), nighttime "calms" NC (0000-0400), relative humidity RH (0900-1300) and surface wind speed WS (0900-1500). Other variables include maximum temperature TMAX (daily peak value), air mass trajectory corridor TRAJ (upwind, previous 36 hours), day of week DOW (number of Saturdays), rainfall RF (daily total) and length of day LOD (hours from sunrise to sunset). The authors applied a combination of nonlinear and linear regression models in a two-step procedure. First, ozone concentrations were calculated using the nonlinear equation:

$$[O_3]_{nl} = [\theta_1 + (\theta_2 + \theta_3*TMAX + \theta_4*TMAX^2)*\exp(\theta_5*WS)]*\exp(\theta_6*RH)$$

where θ_1-θ_6 ($\theta_1 = 76.5$, $\theta_2 = 181$, $\theta_3 = -9.26$, $\theta_4 = 0.0933$, $\theta_5 = -0.115$, $\theta_6 = -0.0654$) are the ordinary least square parameters of best fit. The resulting nonlinear ozone output $[O_3]_{nl}$ is then used as another predictor variable in the final form of the linear regression model, i.e.,

$$[O_3] = b_o + b_1*[O_3]_{nl} + b_2*CC + b_3*DOW + b_4*LOD + b_5*NC + b_6*RF + b_7*TRAJ$$

where b_o-b_6 ($b_o = -43.7$, $b_1 = 0.800$, $b_2 = -0.732$, $b_3 = 4.14$, $b_4 = 4.16$, $b_5 = 1.55$, $b_6 = -2.29$, $b_7 = 11.3$) are the coefficients of the linear equation, fitted using ordinary least squares and the stepwise method (IMSL, 1992). However, if the data for the linear predictors are scarce or unavailable, the authors suggested the use of only nonlinear ozone model since its R^2 value is 0.724, close to that of the linear model ($R^2 = 0.790$).

However, the implementation of nonlinear models outside of their respective area of development may be inappropriate. For example, Bloomfield et al. (1996) successfully predicted future ozone concentration in Chicago, Illinois but when Davis and Speckman (1999) applied the same model in Houston, Texas, they were disappointed with the

results. The reason for the poor model performance might be caused by dissimilarity in meteorological conditions and/or relations between ozone (response) and the predictors. When this happens, we should resort to tree analysis.

2.4.1.3. Regression Tree Analysis

When dealing with multivariate statistics where complex relations between the response variables and predictors are expected, alternatives to nonlinear regression methods are desired; particularly, the methods based on partitioning the predictor space into mutually orthogonal subspaces depending on some error criteria. Such nonlinear, orthogonal partitioning approaches not only result in more robust predictions but are also useful for identifying the most probable criteria responsible for a given process. Examples include the implementation of the CART (Classification and Regression Tree; Breiman et al., 1984) algorithm and the S-Plus built-in tree functions for analyzing censored survival data (LeBlanc and Crowley, 1993), predicting maximum tropospheric ozone concentrations for the major cities in Canada (Burrows et al., 1995), classifying the meteorological covariates of hurricanes (Elsner et al., 1996) and defining meteorological regimes influencing ozone trends (Huang and Smith, 1999). In another variation of tree analysis, which is based on cluster-specific (average linkage and K-means) generalized additive models (GAM), Davis et al. (1998) were able to identify three covariates, i.e., daily maximum surface temperature, average v component (y-direction) of the surface wind and total global radiation, to be important factors in causing high ozone concentrations in Houston, Texas.

The general idea behind regression tree modeling is analyzing trends within different clusters of the data sets by growing the tree, i.e., forming clusters by recursively partitioning data into two distinct groups (binary tree) until the difference is no longer significant. Consider, for example, a given response variable Y (i.e., ozone) with observations y_i, $i = 1, \ldots, M$, and a set of predictors $\{X_1, \ldots, X_J\}$, which are also termed exploratory variables; then assume Y given the predictor values has a normal (Gaussian) distribution with varying mean μ (depends on predictors, i.e., may be heteroscedastic)

and common variance σ^2. Note that the assumption of normal distribution could be substituted by a different likelihood distribution if the data sets warrant it.

In order to grow a tree, the measurement space of all predictors $X = X_1 \otimes \ldots \otimes X_J$, where $\otimes$ denotes the Cartesian product, are recursively partitioned into two groups X_L (left) and X_R (right) based on certain threshold values x_{jk}, $j = 1, \ldots, J$, $k = 1, \ldots, K$, of some predictor variables X_j. For example, a temperature (X_1) of ~20°C (x_{11}) and low wind speed (X_2) of 7.5 knots (x_{21}) in the springtime were found to be conducive to ozone formation (Gardner and Dorling, 2000); these thresholds x_{11} and x_{21} may determine whether a group of data sets goes left or right after successive splits.

The recursive partitioning regression may be performed via a stepwise procedure; particularly, by using indicator function $I(x_{jm})$ whose value is one if the variable values x_{jm} are less than a threshold value x_{jk}, i.e., $I(x_{jm}) = 1$, $\forall(x_{jm} \leq x_{jk}) \in X_L$, and zero otherwise, i.e., $I(x_{jm}) = 0$, $\forall(x_{jm} \geq x_{jk}) \in X_R$. The primary goal of regression tree is to approximate the response values $\hat{y}(x_j)$ by a linear combination of basis functions $B_m(x_{jm})$ as follows (Friedman, 1990):

$$\hat{y}(x_j) = \sum_{m=1}^{M} a_m B_m(x_{jm}), \quad j = 1, \ldots, J \tag{2.21}$$

where a_m are jointly calculated in such a way that the data fitting error is minimized; the basis functions $B_m(\cdot)$ are modeled as:

$$B_m(x_{jm}) = I[x_{jm} \in R_m], \quad \forall x_{jm} \tag{2.22}$$

where $I(\cdot)$ is the indicator function as explained above and R_m are the subregions of the covariate space representing the entire domain. To obtain a more explicit set of basis functions based on available data, $B_m(\cdot)$ are further defined as a collection of Heaviside step functions $H(\eta)$, whose value is one if the argument η is positive and zero otherwise, i.e.,

$$B_m(x_{jm}) = \prod_{k=1}^{K_m} H[s_{km} \cdot (x_{jm} - x_{jk})], \quad \forall x_{jm}, x_{jk} \tag{2.23}$$

where $[s_{km} \cdot (x_{jm} - x_{jk})]$ denotes the argument η; the parameters s_{km} carry the values of ± 1 in order to always ensure a positive step function; K_m is the total number of splits after the tree stops growing; x_{jm} are the values of the predictor variables X_j and x_{jk} are the corresponding thresholds. To illustrate a partitioning procedure (growing the binary tree), Figure 2.1 shows an example on how to obtain four subregions (R_1-R_4) by splitting the original set of predictor data (x_{jm}) using three threshold values (x_{jk}). Each branch (intermediate node), including the root (first node), of the binary tree is represented by a step function $H(\eta)$, and at the end of the branch(es) lie a leaf (terminal node), which is represented by a specific basis function $B_m(\cdot)$.

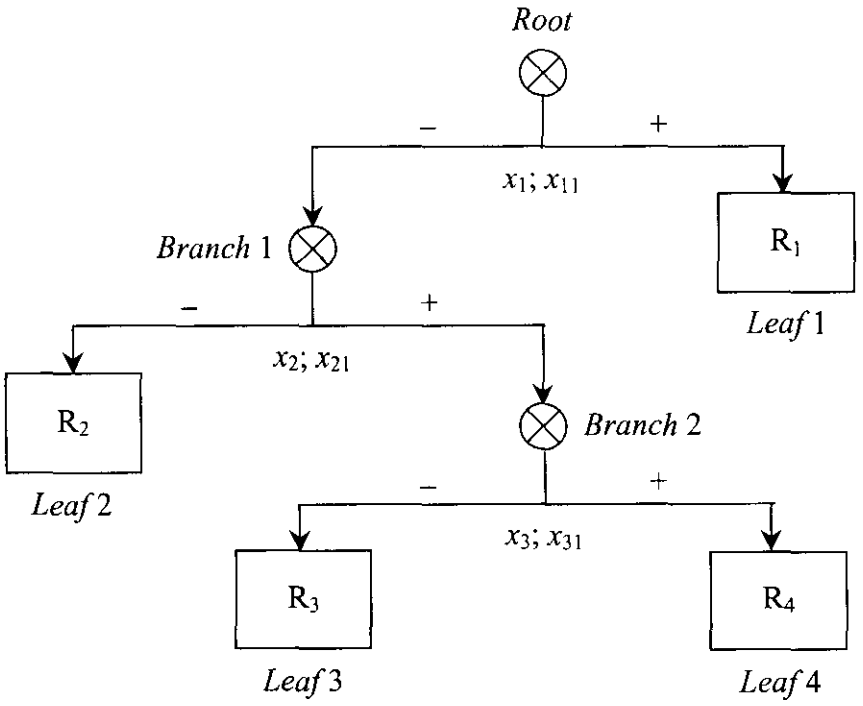


Figure 2.1
Regression tree structure. Four subregions (R_1-R_4) are generated from three thresholds x_{jk} corresponding to each set of predictor data x_j.

From the above figure, the basis functions can be written as follows:

$B_1 = H[+(x_{1m} - x_{11})]$

$B_2 = H[-(x_{1m} - x_{11})]\ H[-(x_{2m} - x_{21})]$

$B_3 = H[-(x_{1m} - x_{11})]\ H[+(x_{2m} - x_{21})]\ H[-(x_{3m} - x_{31})]$

$B_4 = H[-(x_{1m} - x_{11})]\ H[+(x_{2m} - x_{21})]\ H[+(x_{3m} - x_{31})]$

where the minus signs prior to the argument η are utilized to guarantee that all basis functions $B_m(\cdot)$ have nonzero values.

This process must also satisfy maximum deviance D_i so that the difference between the conditional densities $f(y \mid X_L)$ and $f(y \mid X_R)$ is the highest. For a set of observations y_i, $i = 1, \ldots, M$, the deviance D_i is written as (Huang and Smith, 1999):

$$D_i(\mu; y) = \sum_{i=1}^{M} (y_i - \mu)^2 \tag{2.24}$$

where μ is the arithmetic average of y_i, i.e., the minimum deviance estimate. If D_{parent} represents the deviance of the upper node while D_{childL} and D_{childR} are the deviances of the binary split (left L and right R), the maximum value (over j) can be determined from:

$$\Delta D = D_{\text{parent}} - (D_{\text{childL}} + D_{\text{childR}}) \tag{2.25}$$

The tree is usually left to grow up to a certain size (denoted by T_{max}) when it is stopped after the number of observations y_i, $i = 1, \ldots, M$, in a cluster is less than or equal to 5, or when ΔD is less than 1 percent of D_{parent}. However, this large tree must be optimized because excessive splits increase the variance of the estimated means μ on all terminal nodes. To achieve this, the tree can be pruned by omitting the trivial splits according to a certain criterion. In practice, it is common to apply a so-called cost-complexity measure $D_T(\alpha)$ to evaluate the goodness of fit in order to obtain the optimized subtree $T \leq T_{max}$. The deviance D_T of a subtree can be correlated to the tree size $S(T)$, i.e., the number of terminal nodes, and the cost-complexity parameter α by (Clark and Pregibon, 1992):

$$D_T(\alpha) = D_T + \alpha S(T) \tag{2.26}$$

In other words, the cost-complexity measure (total deviance) is a linear combination of the subtree cost (deviance D_T) and its complexity. The tree object $T(\alpha)$ may be readily found by minimizing the cost-complexity measure at constant α. However, to obtain an optimal tree, we must balance the total deviance $D_T(\alpha)$ and tree size $S(T)$ based on the smallest deviance possible as well as a manageable tree size.

The application of regression tree analysis in atmospheric science has gained momentum because of its ability in identifying the underlying physical and chemical mechanisms of complex nonlinear processes. As remarked by Huang and Smith (1999), the regression tree approach has the advantage of estimating "different trends in different clusters." That is the analysis for average and maximum ozone concentrations can be performed simultaneously but separately in a single run, which is useful and time saving. However, if the tree is over- or under-pruned, the end results may be less satisfactory such that its performance (e.g., R^2 statistic) is inferior to, say, a neural network (a form of nonlinear black box model) and only comparable to that of simple linear regression. In addition, if a process is known to be linear *a priori*, the implementation of a regression tree is only as good as, if not worse than, that of simple linear regression.

In essence, regression-based models may be easier to be physically interpreted than other more sophisticated approaches, e.g., artificial neural network. However, the performance of the models may sometime have to be compromised. As Gardner and Dorling (2000a) remarked in the comparative study of surface ozone concentration in the major cities of the U.K., the reliabilities of linear regression, regression tree and neural network modeling approaches could be measured using various goodness-of-fit criteria. In particular, they compared the relative performances of the three approaches based on:

1. Mean bias error (MBE), which indicates the differences between the predicted $\hat{\mu}$ and observed μ mean concentrations as a measurement of under- or over-prediction.

2. Mean absolute error (MAE), which summarizes the absolute differences between the predicted $\hat{Z}$ and observed Z values, i.e., the residual errors.
3. Root mean squared error (RMSE). Similar to MAE, the RMSE also summarizes the residual errors by squaring the differences between the predicted $\hat{Z}$ and observed Z values, adding them and taking the square root of the final result.
4. Coefficient of determination (R^2), which measures the variability in the predicted values as compared with the observed data (Section 4.1). R^2 is useful since it is bounded between 0 and 1.
5. Index of agreement (d_α), which is a more useful measure than the R^2 statistic since it indicates deviation from the observed mean values μ. The index d_α is defined as (Willmott, 1982):

$$d_\alpha = 1 - \frac{\sum_{i=1}^{M} |\hat{Z}_i - Z_i|^\alpha}{\sum_{i=1}^{M} \left(|\hat{Z}_i - \mu_i| + |Z_i - \mu_i| \right)^\alpha}$$

In general, Gardner and Dorling (2000a) found that the regression tree and neural network analyses outperformed those of the linear regression models. However, the reverse was true in the special case occurring at the city of Southampton in which the RMSE, R^2 and d_α statistics of the linear model were superior to those of the regression tree and just short of the neural network performance. Comrie (1997) also compared the neural network with linear regression models and obtained similar results. He found only slight improvements, in term of R^2 statistic, after applying neural network analysis for predicting maximum ozone concentrations in the U.S. urban areas.

On the other hand, nonlinear models may not always outperform the linear regression approaches. Often the relationships between the response and its predictors, like those in the city of Southampton, show linearity and therefore best analyze using linear regression models due to their simplicity. Unfortunately, real-life phenomena are usually complex and nonlinear; to model them linearly is absurd. The use of nonlinear models may be more suitable but even then, the models are only useful at a very limited

area as proven by Davis and Speckman (1999) when they tried to apply the model developed by Bloomfield et al. (1996) for the Houston metropolitan area. In this manner, there is a need for a more versatile approach that can yield good estimations (spatial sense) and prediction (temporal sense), both locally and regionally. This approach known as spatiotemporal modeling is discussed next.

2.4.2 Space-Time Modeling

This approach is historically categorized under the area of geostatistics, which form a "new" branch of statistical science dealing with spatial and temporal phenomena. Since the pioneering works of Matheron (1962) in geology and Gandin (1963) in meteorology, the application of geostatistical methods has been extended into various fields. Examples can be found in the works of: Eynon and Switzer (1983) on the rainfall pH in the Northeastern United States, Laslett (1994) on gilgais, which are geographical phenomena of naturally "gentle depressions in otherwise flat land," Kyriakidis (1999) on sulfate deposition over Europe, and more recently Bechini et al. (2000) on global solar radiation over agricultural area in Italy.

The work of Kyriakidis (1999) instigated significant interest in the environmental science community due to the application of stochastic analyses, or more specifically, batch and recursive-type direct sequential simulation (DSSIM) to estimate (in spatial sense: $\mathbf{u}$) and predict (in temporal sense: t) the monthly average sulfate (SO_4^{2-}) concentrations in various European countries. Here the concepts of space-time phenomena are discussed in detail through the use of random variable (RV) and random function (RF), both of which are also known as random processes. Owing to the decision in stochastic modeling, a random process may be decomposed into two uncorrelated fields, e.g.,

$$Z(\mathbf{u}, t) = M(\mathbf{u}, t) + R(\mathbf{u}, t), \qquad \forall \mathbf{u} \in D, \forall t \in T \tag{2.27}$$

where $M(\mathbf{u}, t)$ is a stochastic trend modeling the "mean" or smooth variability of the spatiotemporal random process (signal) $Z(\mathbf{u}, t)$ for every location in space $\mathbf{u} \in D$ and

instant in time $t \in T$, and $R(\mathbf{u}, t)$ is a stationary space-time residual component around that trend and often modeled independently.

The procedure begins by establishing a deterministic spatiotemporal trend (mean) $\{m(\mathbf{u}_\alpha, t_i), \alpha \in (N)\}$ of the sample time series TS $\{z(\mathbf{u}_\alpha, t_i), i \in T_\alpha\}$, which can be regarded as one realization of a temporal random process $\{Z(\mathbf{u}_\alpha, t_i), i \in T_\alpha\}$ at the monitoring station $\mathbf{u}_\alpha \in D$ and at the discrete time instants t_i. Under stationarity, the trend (mean) is inferable and hence the decomposition (a model) of the random process can be re-defined as:

$$Z(\mathbf{u}_\alpha, t_i) = m(\mathbf{u}_\alpha, t_i) + R(\mathbf{u}_\alpha, t_i), \quad t_i \in T_\alpha \tag{2.28}$$

where $m(\mathbf{u}_\alpha, t_i)$ is a deterministic temporal trend, $R(\mathbf{u}_\alpha, t_i)$ is a stationary, zero-mean stochastic residual process and T_α is the time span of the measurements available at station $\mathbf{u}_\alpha$. Having established the above model, which was adopted by many researchers including Journel and Huijbregts (1972) in the context of mining, we now require the techniques to determine both the trend $m(\mathbf{u}_\alpha, t_i)$ and the residual process $R(\mathbf{u}_\alpha, t_i)$.

2.4.2.1. Modeling the Trend

It is common to have a cyclical variation of the sample data, especially when we deal with the concentrations of air pollutants over a period of time. A good deterministic trend model must be able to reproduce such cyclic variation, and one way to proceed is by using Fourier series of sine and cosine functions. For further simplification, the deterministic trend $m(\mathbf{u}_\alpha, t_i)$ is modeled independently at each station $\mathbf{u}_\alpha$ by applying the following conditional independence assumption:

$$E\{Z(\mathbf{u}, t) \mid [Z(\mathbf{u}, t'), t' \in T], [Z(\mathbf{u}', t'), t' \in T]\} = E\{Z(\mathbf{u}, t) \mid [Z(\mathbf{u}, t'), t' \in T]\} \tag{2.29}$$

which says that the random process $\{Z(\mathbf{u}, t), t \in T\}$ at each station $\mathbf{u}$ is obtained only from sample data $\{Z(\mathbf{u}, t'),\ t' \in T\}$ at the same station $\mathbf{u}$, i.e., the temporal records at location $\mathbf{u}$

screen those obtained from all other locations $\mathbf{u}'$. This particular assumption allows the deterministic trend to be modeled as:

$$m(\mathbf{u}_\alpha, t_i) = \sum_{k=0}^{K} b_k(\mathbf{u}_\alpha) f_k(t_i), \quad i = 1, \ldots, T_\alpha \tag{2.30}$$

where $b_k(\mathbf{u}_\alpha)$ are the coefficients of the Fourier functions $f_k(t_i)$, with $f_o(t_i) = 1$ by convention. In other words, the deterministic trend is modeled as a summation of $(K + 1)$ "known" temporal functions $f_k(t_i)$, i.e., sines and cosines, having certain amplitude A_k and frequency ω_k.

However, the objective is to construct a stochastic space-time trend field $M(\mathbf{u}, t)$, which allows estimation or simulation of a trend at any unmonitored location $\mathbf{u}$. This can be accomplished by regionalizing the trend coefficients $b_k(\mathbf{u})$ at, for example, three monitoring stations $\mathbf{u}_\alpha$, $\mathbf{u}_\beta$ and $\mathbf{u}_\gamma$ at time instant t_i. Therefore, $M(\mathbf{u}, t)$ can be defined as a joint realization of a set of $(K + 1)$ cross-correlated random functions (RFs) $\{B_k(\mathbf{u}), \mathbf{u} \in D\}$, $k = 0, \ldots, K$, over the entire space-time domain $D \times T$:

$$M(\mathbf{u}, t) = \sum_{k=0}^{K} B_k(\mathbf{u}) f_k(t), \quad \forall \mathbf{u} \in D, \forall t \in T \tag{2.31}$$

with the expected value of:

$$E\{M(\mathbf{u}, t)\} = E\{\sum_{k=0}^{K} B_k(\mathbf{u}) f_k(t)\} = \sum_{k=0}^{K} E\{B_k(\mathbf{u})\} f_k(t), \quad \forall \mathbf{u} \in D, \forall t \in T \tag{2.32}$$

or in simple terms, the stochastic spatiotemporal trend model $M(\mathbf{u}, t)$ and its expected value $E\{M(\mathbf{u}, t)\}$ can be obtained from the weighted linear combinations of the $(K + 1)$ RFs $\{B_k(\mathbf{u}), \mathbf{u} \in D\}$ and its expected value $E\{B_k(\mathbf{u})\}$, respectively.

2.4.2.2. Modeling the Residue

The next step is to model the space-time residual field $R(\mathbf{u}, t)$. At the data $z(\mathbf{u}_\alpha, t_i)$ locations, we obtain the residual TS $r(\mathbf{u}_\alpha, t_i)$ as:

$$r(\mathbf{u}_\alpha, t_i) = z(\mathbf{u}_\alpha, t_i) - \sum_{k=0}^{K} b_k(\mathbf{u}_\alpha) f_k(t_i), \quad i = 1, \ldots, T_\alpha \tag{2.33}$$

Note that this residual is algorithm-specific because it depends on the choice of Fourier function $f_k(t_i)$ and the corresponding coefficients $b_k(\mathbf{u}_\alpha)$. In order to obtain the random residual field $R(\mathbf{u}_\alpha, t_i)$, we must regionalize the residual data $r(\mathbf{u}_\alpha, t_i)$ in space by initially standardizing them at every monitoring station $\mathbf{u}_\alpha$ to unit variance. This task can be achieved through division of each residual TS $r(\mathbf{u}_\alpha, t_i)$ by the standard deviation $s_R(\mathbf{u}_\alpha)$ of the residual profile at that particular station, i.e.:

$$\hat{r}(\mathbf{u}_\alpha, t_i) = \frac{r(\mathbf{u}_\alpha, t_i)}{s_R(\mathbf{u}_\alpha)}, \quad i = 1, \ldots, T_\alpha \tag{2.34}$$

where $s_R(\mathbf{u}_\alpha)$ is the square root of the variance $s_R^2(\mathbf{u}_\alpha)$, which is defined below:

$$s_R^2(\mathbf{u}_\alpha) = \frac{1}{T_\alpha} \sum_{i=1}^{T_\alpha} r^2(\mathbf{u}_\alpha, t_i) \tag{2.35}$$

Furthermore, to obtain the standardized residual field $\{\hat{R}(\mathbf{u}, t), t \in T\}$ at any unmonitored location $\mathbf{u}$, the standardized residual TS $\{\hat{R}(\mathbf{u}_\alpha, t), t \in T\}$ at location $\mathbf{u}_\alpha$ may be decomposed into $(L + 1)$ linear summation of component TS $\{\hat{R}_l(\mathbf{u}_\alpha, t), t \in T\}$ weighted with w_l as follows:

$$\hat{R}(\mathbf{u}_\alpha, t) = \sum_{l=0}^{L} w_l(\mathbf{u}_\alpha) \hat{R}_l(\mathbf{u}_\alpha, t), \quad \alpha \in (n) \tag{2.36}$$

with the conditions that the expected value $E\{\hat{R}_l(\mathbf{u}_\alpha, t)\} = 0, \forall l$ and the covariance $Cov\{\hat{R}_l(\mathbf{u}_\alpha, t), \hat{R}_{l'}(\mathbf{u}_\alpha, t+\tau)\} = \delta_{ll'} C_{\hat{R}}[\tau; q_l(\mathbf{u}_\alpha)]$; $\delta_{ll'}$ is the Kronecker delta, whose value is one if $l = l'$ and zero otherwise, to ensure that the correlogram $C_{\hat{R}}[\tau; q_l(\mathbf{u}_\alpha)]$ of the l^{th} component TS $\hat{R}_l(\mathbf{u}_\alpha, t)$ is a diagonal matrix. τ is the temporal lag and $q_l(\mathbf{u}_\alpha)$ is the range of the l^{th} basic correlogram model and by default $q_o(\cdot) = \varepsilon$, i.e., approaching zero. Note that the above is only a modeling decision and carries the implication that the residual field $R(\mathbf{u}, t)$ can be decomposed into $(L + 1)$ mutually independent temporal structures.

Another modeling decision is to constrain all $(L + 1)$ basic correlogram models so that they share the same characteristics, e.g., Gaussian, spherical or exponential and/or the combination of two or more basic structures. The previous decomposition leads to the correlogram function $C_{\hat{R}}[\tau; \mathbf{q}(\mathbf{u}_\alpha)]$, which is expressed simply as a linear summation of components $C_{\hat{R}}[\tau; q_l(\mathbf{u}_\alpha)]$ weighted with positive sill $a_l(\mathbf{u}_\alpha) = [w_l(\mathbf{u}_\alpha)]^2$.

$$C_{\hat{R}}[\tau; \mathbf{q}(\mathbf{u}_\alpha)] = \sum_{l=0}^{L} a_l(\mathbf{u}_\alpha) C_{\hat{R}}[\tau; q_l(\mathbf{u}_\alpha)] \tag{2.37}$$

Since the coefficients $w_l(\mathbf{u}_\alpha)$ can only be determined at the monitoring station $\mathbf{u}_\alpha$, they need to be regionalized in such a way that a set of $(L + 1)$ coefficients $\{w_l(\mathbf{u}_\alpha), \alpha \in (n), l \in L\}$ and range $q_l(\mathbf{u}_\alpha)$ can be modeled as a joint realization of $2 \cdot (L+1)$ cross-correlated RFs $\{[W_l(\mathbf{u}), Q_l(\mathbf{u})], \mathbf{u} \in D\}$. Therefore, the standardized spatiotemporal residual profile $R(\mathbf{u}, t)$ at any location $\mathbf{u}$ and time instant t is redefined as:

$$\hat{R}(\mathbf{u}, t) = \sum_{l=0}^{L} W_l(\mathbf{u}) \hat{R}_l(\mathbf{u}, t) \tag{2.38}$$

and the covariance function $C_{\hat{R}}[\tau; q_l(\mathbf{u})]$ as:

$$C_{\hat{R}}[\tau;\mathbf{q}(\mathbf{u})] = \sum_{l=0}^{L} a_l(\mathbf{u}) C_{\hat{R}}[\tau; q_l(\mathbf{u})] \tag{2.39}$$

where $a_l(\mathbf{u}) = [w_l(\mathbf{u})]^2$ and $q_l(\mathbf{u})$, defined at the same location $\mathbf{u}$, are now the realizations of the RVs $A_l(\mathbf{u})$ and $Q_l(\mathbf{u})$, respectively.

Finally, the space-time residual field $R(\mathbf{u},t)$ can be simulated so as to ensure the reproduction of the covariance model $C_{\hat{R}}[\tau; q_l(\mathbf{u})]$, or in other words, the residual field can be calculated by combining expressions (2.35) and (2.38):

$$R(\mathbf{u},t) = S_R(\mathbf{u})\hat{R}(\mathbf{u},t) \tag{2.40}$$

with the prior assumption that $Cov\{S_R(\mathbf{u})\hat{R}(\mathbf{u},t)\} = 0$, $\forall \mathbf{u}$, $\mathbf{u}'$, t, and also that the following two conditions apply:

1. Expected (mean) value:

$$\begin{aligned} E\{R(\mathbf{u},t)\} &= E\{S_R(\mathbf{u})\hat{R}(\mathbf{u},t)\} \\ &= E\{S_R(\mathbf{u})\}E\{\hat{R}(\mathbf{u},t)\} \\ &= 0 \end{aligned} \tag{2.41}$$

which means that $S_R(\mathbf{u})$ and $\hat{R}(\mathbf{u},t)$ are assumed mutually orthogonal in order to obtain the unbiased residual field $R(\mathbf{u},t)$, and

2. Covariance:

$$\begin{aligned} E\{R(\mathbf{u},t)R(\mathbf{u}',t')\} &= E\{[S_R(\mathbf{u})\hat{R}(\mathbf{u},t)][S_R(\mathbf{u}')\hat{R}(\mathbf{u}',t')]\} \\ &= E\{S_R(\mathbf{u})S_R(\mathbf{u}')\}E\{\hat{R}(\mathbf{u},t)\hat{R}(\mathbf{u}',t')\} \\ &= C_{S_R}(\mathbf{u}-\mathbf{u}')Cov\{\hat{R}(\mathbf{u},t)\hat{R}(\mathbf{u}',t')\} \end{aligned} \tag{2.42}$$

or in words, the covariance of S_R is only a function of space $\mathbf{u}$ whereas that of $\hat{R}$ depends on both space $\mathbf{u}$ and time $\mathbf{t}$.

The space-time phenomena are complex in nature and usually modeled by making certain decisions on the stochastic trend $M(\mathbf{u}, t)$ and residual $R(\mathbf{u}, t)$ components. The former component is usually modeled deterministically at the station locations whereas the latter stochastically. Before going a step further, it is perhaps more useful to gain a prior insight of the temporal profiles and annual trends of relevant variables in order to preliminarily assess the influence of covariates on ozone. If the first two moments (mean and variance) of the temporal phenomena can be correctly reproduced by such methods as stochastic simulation, the prediction of random fluctuations in ozone profiles may be performed with less uncertainty. The implementation of regression-based methods can also be practical due to the more direct approach. However, extreme care must be exercised because ozone formation is highly nonlinear. The application of a simple linear model in predictive modes may require numerous studies at different spatial locations. Otherwise, the statistical analyses investigated at one location may not be useful at the other due to dissimilarities in meteorological events. For this reason and to verify the accuracy of temporal models before regionalization, various statistical methodologies are utilized for investigating ozone phenomena in Calgary, Alberta.

CHAPTER 3

EXPLORATORY DATA ANALYSIS

The majority of data used in this thesis are provided by the Clean Air Strategic Alliance (CASA). These continuous (hourly average) data are stored in the Alberta Ambient Air Data Management System (AAADMS), a publicly accessible repository[1] whose current stakeholders include Alberta Environment (AE), Environment Canada (EC), the Canadian Forest Service (CFS), the West Central Airshed Society (WCAS) and the Strathcona Industrial Association (SIA). In Calgary, Alberta there are three major environmental monitoring stations, respectively located on the Northwest, Central and East sides of the city area. The types of chemical and meteorological variables recorded at each station vary depending on their necessity. For example, the concentration of sulfur dioxide (SO_2) is only recorded at the East station to monitor emissions from the industrial zones and wastewater treatment plants such as those in Ogden-Foothills and Bonnybrook. This way, the massive data collection campaign is cost effective.

The availability of relevant environmental data at the monitoring stations is sufficient for studying ozone phenomena. 'Comma-delimited' (in ASCII/ CSV format) data sets can be downloaded from the CASA's website. However, care should be taken when dealing with enormous amount of data because they may induce confusion due to incorrect formatting and/or different units of measurement. With the exception of the wind speed, and the amount of dust and smoke, which are measured in the respective units of kilometers per hour (km/h) and coefficient of haze (COH), the chemical species, e.g., carbon monoxide, nitric oxide, nitrogen dioxide, total hydrocarbon and ozone, are recorded in parts per million by volume (ppmv).

[1] See CASA's homepage www.casadata.org

Other meteorological variables, i.e., temperature, relative humidity and bright sunshine hours, were obtained from the EC's monitoring station at Calgary's International Airport (in the northeast quadrant of the city). Temperature and relative humidity data, measured in the corresponding units of degree Celsius (^{o}C) and percentage of water in the atmosphere (%), are averaged (by arithmetic means) from the minimum and maximum daily average values. The daily duration (in hours, i.e., minutes are converted to hourly decimal) of which the sunlight "burns" a standard paper is termed the bright sunshine hours, a good surrogate predictor to the intensity of solar radiation (e.g. ultraviolet radiation).

3.1 Selections of Covariates and Temporal Scale

To a certain degree of confidence, previous studies have identified the covariates (predictors) responsible for instigating ozone episodes in major metropolitan areas. For example, Derwent and Davies (1994) showed that ozone is produced from the complex photochemical reactions between nitrogen oxides (NO_x, mainly consisting of NO and NO_2), volatile organic compounds (VOCs) and their radical derivatives. Kajii et al. (1998) observed seasonal variations of tropospheric ozone (O_3) and carbon monoxide (CO) in Happo, Japan and found these two variables to be positively correlated, especially in the months of April through July.

In a study of meteorological-dependency of ozone episodes, Cox and Chu (1993, 1996) studied the effects of about one hundred variables in the major urban areas of the U.S. and determined that only some of them, i.e., maximum surface temperature, relative humidity, mixing (ceiling) height, opaque cloud cover, as well as average wind speed and direction, are significant to ozone prediction. The importance of cloud effects on ozone was emphasized by Matthijsen et al. (1996) who investigated aqueous-phase chemistry and wet depositions in Europe using the long-term ozone simulation (LOTUS) package. These finding were concurred by other researchers, e.g., Bloomfield et al. (1996) who extended the list by including barometric pressure, dewpoint temperature, specific humidity and visibility. However, the exact determination of the dominant variables is difficult due to variations in regional meteorology and pollutant emission patterns. Davis

et al. (1998) verified this fact when they obtained unsatisfactory results for Houston, Texas after trying to apply the nonlinear model developed by Bloomfield et al. (1996) for Chicago region.

Based on various suggestions and conclusive evidence from the literature, nine covariates are selected in an attempt to predict daily average ozone concentrations in Calgary, Alberta. Seven of the variables (1-7; Table 3.1) used in this work are down-loaded from CASA's website. These data are those recorded at the East monitoring station, chosen due to availability of wind field (speed and direction) data and proximity to the meteorological monitoring location (Northeast quadrant). The other three variables (8-10; Table 3.1) are provided by the Environment Canada (EC), available on paper format at the University of Calgary's main library or from the EC's regional headquarters in Calgary. For convenience, the abbreviations of all variables used in this thesis are listed below:

Table 3.1
Name abbreviations of the chemical and meteorological variables.

No.	Variables	Abbreviations
1.	Dust and smoke	COH
2.	Carbon Monoxide	CO
3.	Nitric oxide	NO
4.	Nitrogen Dioxide	NO2
5.	Total hydrocarbon	THC
6.	Ozone	O3
7.	Wind speed	WSPD
8.	Average temperature	Tavg
9.	Average relative humidity	RHavg
10.	Bright sunshine hours	bSUN

The "quality" of data collection is generally excellent but on rare occasions, a few missing data values were observed for several hours in a month. For example, ozone concentrations were not recorded for thirty-seven hours, i.e., from 0400 on April 14 to

1700 on April 15, 1997. To tackle this problem, twenty-four hourly values are averaged to obtain a single daily average "datum;" this approach is also favored by Feister and Balzer (1991). In this case, the hourly average data are considered to be time series TS $z_\alpha(t_i)$, $\alpha = 1, \ldots, 7$, $i \in N \leq 24$ and the daily average values as random variables RVs $Z_\alpha(t_j)$, $\alpha = 1, \ldots, 7$, $j \in J = 365$. Keeping this in mind, the averaging procedure is carried out as follows:

$$Z_\alpha(t_j) = \frac{1}{N}\sum_{i=1}^{N} z_\alpha(t_i) \tag{3.1}$$

$$\forall\, \alpha = 1, \ldots, 7, j = 1, \ldots, J = 365$$

N denotes the total number of hours where the measurement of each time series TS $z_\alpha(t_i)$ is taken continuously on a particular day $j \in J = 365$ (number of days in a year); note that N may be less than 24 due to missing values. The index α refers to the predictor variables (1-7 in Table 3.1) obtained from CASA. However, it should be stressed that the TS $z_\alpha(t_i)$ are converted to RV $Z_\alpha(t_j)$ without prior knowledge of their respective distributions. This decision is valid only if we analyze the average behaviors of certain pollutants in a relatively small region. If we are interested in the study of extreme phenomena, e.g., hourly maximum ozone episodes, the better approach is to model the TS $z_\alpha(t_i)$ using Weibull distributions (Cox and Chu, 1993 and 1996)

Next, the simple averaging process above is repeated for the entire year. When missing daily values are identified, they are estimated from the previous seven-daily average data, i.e.,

$$\hat{Z}_\alpha(t_j) = \frac{1}{K}\sum_{j=0}^{K} Z_\alpha(t_{j-K}) \tag{3.2}$$

$$\forall\, \alpha = 1, \ldots, 7; K = 7$$

where $\hat{Z}_\alpha(t_j)$ are the inferred missing values. Of course there are more explicit but complicated methods, e.g., smoothing splines, available in the literature but these methods are perhaps more appropriate for estimating large number of missing data unlike those in this work. The missing values for a particular RV $Z_\alpha(t_j)$ were less than one percent (< 72 hours or 3 days) of the total number of daily average values, which are always taken to be 365 for the entire period of the case studies. Therefore the decision to simply average the previous seven-daily values in order to "fill-in" for the missing data will not cause large error in the final results. However, if the temporal phenomena are analyzed using hourly data, the more complicated approach should be utilized due to the large number of missing data.

Furthermore, the zero values from the observed TS $z_\alpha(t_i)$ are considered as valid data because they indicate the absence of certain pollutants during the measurement periods. This approach is contrary to that of Caroll et al. (1997) who treated zero values as missing data. Finally, to avoid complications, the chemical and meteorological data values on the last day of the leap year (i.e., December 31, 2000) are assumed negligible and therefore omitted.

3.2 Time Series Plot

Now that the variables have been selected, it is useful, as a preliminary analysis, to understand the seasonal and yearly patterns of ozone and its predictor variables. The argument is that if we are able to graphically identify the covariates responsible for inducing ozone episodes, the statistical analyses will be meaningful because they are more physically interpretable. This task is achieved by plotting time series of: (1) ozone, and (2) its individual predictors for the entire four years, i.e., 1997-2000. Here time is represented in Julian day (JDay) format where Jan 1 and December 31 of the year are denoted as 1 and 365, respectively. Note that the last day of the leap year (2000) is taken as December 30 to simplify the analysis. Keeping this in mind, the thirty-day moving average 30dMA values $W_\alpha(t_j)$ of the random variables RV $Z_\alpha(t_i)$ are calculated as follows:

$$W_\alpha(t_j) = \frac{1}{2k} \sum_{i=j-k+1}^{j+k} Z_\alpha(t_i), \tag{3.3}$$

$$\alpha = 1, \ldots, 10; j = 15, \ldots, J = 365; k = 15$$

where $2k$ denotes the length of the moving window. In words, thirty values of the RV $Z_\alpha(t_i)$ are averaged in a forward process to produce a single value $W_\alpha(t_j)$, which is placed at the center of the window; this process is repeated for all values of the $Z_\alpha(t_i)$ over the entire year. The application of fifteen-day temporal lag (k = 15) is to ensure the more accurate representations of the yearly trends.

From Figure 3.1, the seasonal and yearly trends of ozone can be visualized. Notice the similarity in the overall patterns, especially between 1997 and 1999 data; there is an increasing trend of ozone concentrations, which peak in the spring (April-May), slowly decreasing until late fall (around November), and then increasing again in the winter season (December-March). On the contrary, the 1998 and 2000 yearly trends illustrate an "anomaly" due to the presence of second peaks in the mid-summer season (June-July); otherwise, the trends would have been similar to those of 1997 and 1999. In general, the 30dMA ozone values $W_\alpha(t_j)$ increase from a maximum level of around 0.025 ppmv in 1997 to 0.030 ppmv in 2000. This significant increase of about 0.005 ppmv may be caused by ozone episodes (i.e., sudden increase in ozone levels) occurring in late April or early May, as depicted by the time series plots of 1999 and 2000.

The predictor variables for the year 1997 are plotted in Figure 3.2. From the visual inspection of the trends, it is difficult to speculate which of the chemical variables (COH, CO, NO, NO2, THC) are directly responsible for ozone formation. This is expected since this process is complex and highly nonlinear. However, if the trends of the meteorological variables (WSPD, Tavg, RHavg, bSUN) are inspected, we can immediately notice their influence on ozone levels. Lower values of RHavg and WSPD cause higher ozone concentrations, which can be seen during ozone episodes in May.

Conversely both bSUN and Tavg are positively related to ozone, as they are initially expected based on results from the literature.

3.3 Data Standardization

As briefly mentioned in Chapter 2 of this thesis, both the response (ozone) and predictor (chemical and meteorological) variables are often pre-processed to simplify further statistical analysis. In the work of Rao, Zurbenko and colleagues (1994-1998), the data are initially transformed into logarithmic values and then filtered using the Kolmogorov-Zurbenko [KZ(m,p)] algorithm, where m and p are the length of moving window and the number of iterations, respectively. This way, the seasonal variation can be separated from the short-term component and the analysis can be performed using simple linear regression. In another approach, Carroll et al. (1997) applied a square root transformation of only ozone data. They followed the suggestion by Hasslett and Raftery (1989), who studied the long-term wind speed variations at twelve monitoring stations in Ireland, due to similarity in the data distributions. The predictor variables, in this case sunlight and temperature, are left in their original forms because there is too much fluctuation in the daily values.

A similar approach is applied in this work because the data are not only noisy but also vary in magnitude. For example, ozone is recorded in the unit of part per million by volume (ppmv), in the order of hundredth (0.01) but temperature is in degree Celsius ranging from –30 to 30°C; hence the difference is in the order of one thousand. To account for this "mismatch," all variables used in this thesis are standardized to zero mean and unit variance, which is easily achieved by first subtracting the RV $Z_\alpha(t_j)$ from each variable-specific stationary mean $\overline{Z}_\alpha = \frac{1}{J}\sum_{j=1}^{J} Z_\alpha(t_j)$, α = 1, …, 10, and J is the total number of Julian days in a year (365 for all case studies). The results are then divided by each variable-specific standard deviation s_{Z_α} of the RV $Z_\alpha(t_j), j \in J$, i.e.,

$$X_\alpha(t_j) = \frac{Z_\alpha(t_j) - \overline{Z}_\alpha}{s_{Z_\alpha}}, \quad \forall \alpha = 1, \ldots, 10;\ j \in J \tag{3.4}$$

The expected value of the standardized random variables RV $X_\alpha(t_j)$, i.e., $E\{X_\alpha(t_j)\}$, which is essentially its mean $\overline{X}_\alpha$, is zero from the following:

$$E\{X_\alpha(t_j)\} = \frac{E\{Z_\alpha(t_j)\} - \overline{Z}_\alpha}{s_{Z_\alpha}} = \frac{\overline{Z}_\alpha - \overline{Z}_\alpha}{s_{Z_\alpha}} = 0 = \overline{X}_\alpha$$

and the variance of the RV $X_\alpha(t_j)$, i.e., $s^2_{X_\alpha}$ is one, which can also be easily verified:

$$s^2_{X_\alpha} = E\{[X_\alpha(t_j) - \overline{X}_\alpha]^2\} = E\{[X_\alpha(t_j)]^2\}$$

$$= E\left\{\left[\frac{Z_\alpha(t_j) - \overline{Z}_\alpha}{s_{Z_\alpha}}\right]^2\right\}$$

$$= \frac{1}{(s_{Z_\alpha})^2} E\{[Z_\alpha(t_j) - \overline{Z}_\alpha]^2\}$$

$$= 1$$

because $\overline{X}_\alpha = 0$ and the numerator $E\{[Z_\alpha(t_j) - \overline{Z}_\alpha]^2\}$ is nothing but the variance of the RV $Z_\alpha(t_j)$, $\alpha = 1, \ldots, 10, j \in J = 365$.

3.4 Box Plot

Another important graphical tool that is helpful for understanding the sample distribution is the box plot. The box is bounded by upper (75^{th} percentile) and lower (25^{th} percentile) quartiles of the samples, with the median (50^{th} percentile) displayed within. In addition, the extreme values are depicted by "whiskers," the two lines extending on both sides of the box. The red pluses (+) symbolize the outliers, which are defined as the observations beyond one and a half times the interquartile ranges. For adequately large samples, this plot may qualitatively represent the first three moments, i.e., center of the distribution,

variability and skewness, of the data. Figure 3.3 illustrates the box plots, drawn using the MATLAB® Statistical Toolbox, of the standardized chemical and meteorological variables $X_\alpha(t_j)$, $\alpha = 1, \ldots, 10, j \in J = 365$. As expected, all variables are centered around the mean (i.e., zero) of the distributions. However, there is wide variability in these variables, as indicated by the large interquartile ranges and long whiskers.

Note that the box plots can also be substituted with histograms in order to depict the distributions of the environmental and meteorological data. However, the analysis using such approach is still performed in the univariate sense. It has been shown in this chapter that the influence of covariates on ozone formation cannot be directly determined from individual time series, or more precisely, annual trends. For example, a rise in the average temperature alone does not immediately increase the ozone concentration. The physical process of tropospheric ozone phenomena involves complex and nonlinear associations of multiple covariates at the same time instant. If an 'ingredient' such as NO2 is missing in that process, ozone may not be formed at all. Furthermore, if the intensity of solar radiation is 'low,' NO2 cannot be decomposed to form oxygen radical, a highly reactive chemical species that will attack an oxygen molecule to produce ozone. Perhaps, it is the combinations of two or more covariates that play the major roles in ozone formation. The analysis based on a linear regression approach will be implemented in the next chapter by employing correlation coefficients between ozone and its covariates.

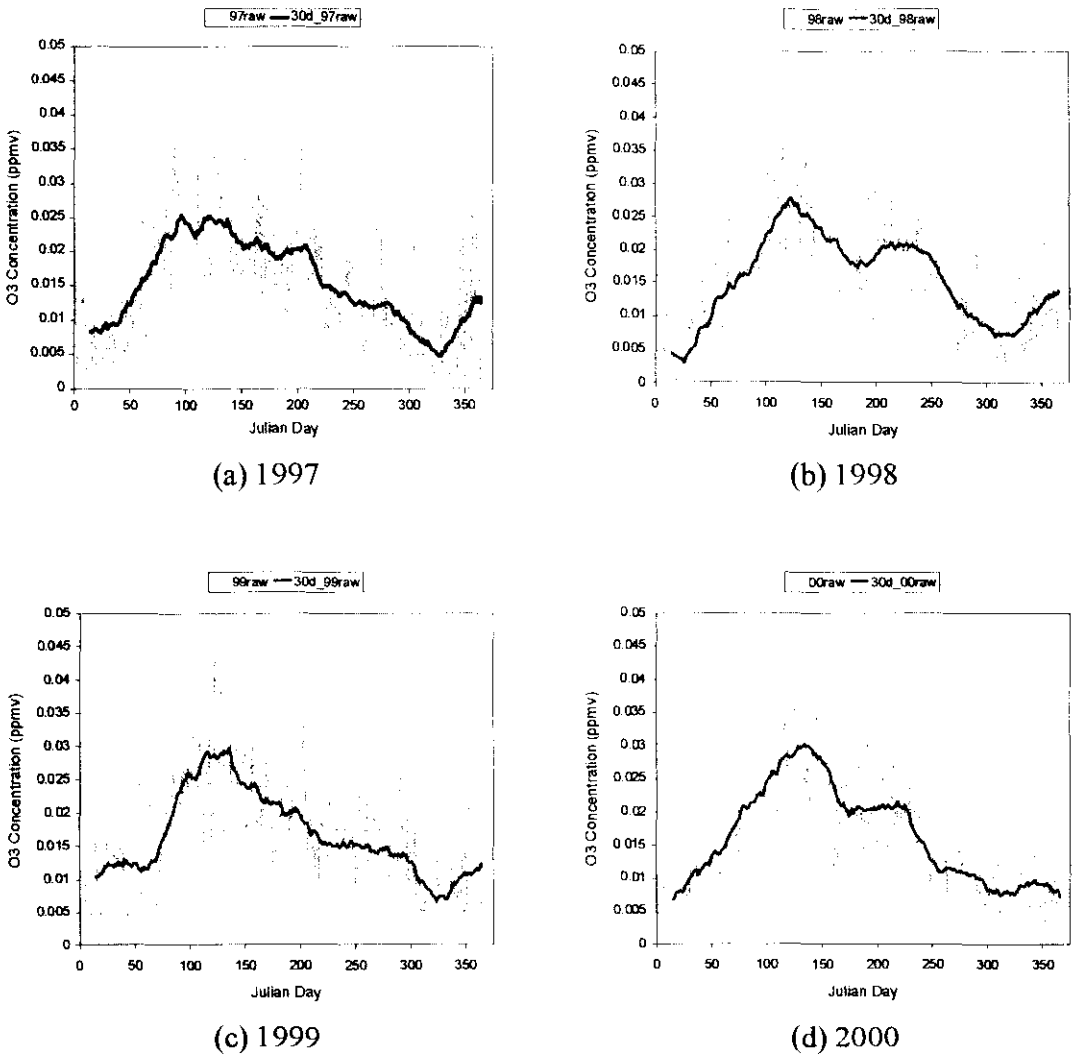


Figure 3.1
Plots of time series and annual trends of ozone. The solid lines are obtained from 30-day moving average values (30d_YYraw; YY is the last two digits of the year and 'raw' denotes the original data) to illustrate the annual trends. Notice the sudden increase in daily average concentration data during the ozone episodes (early May).

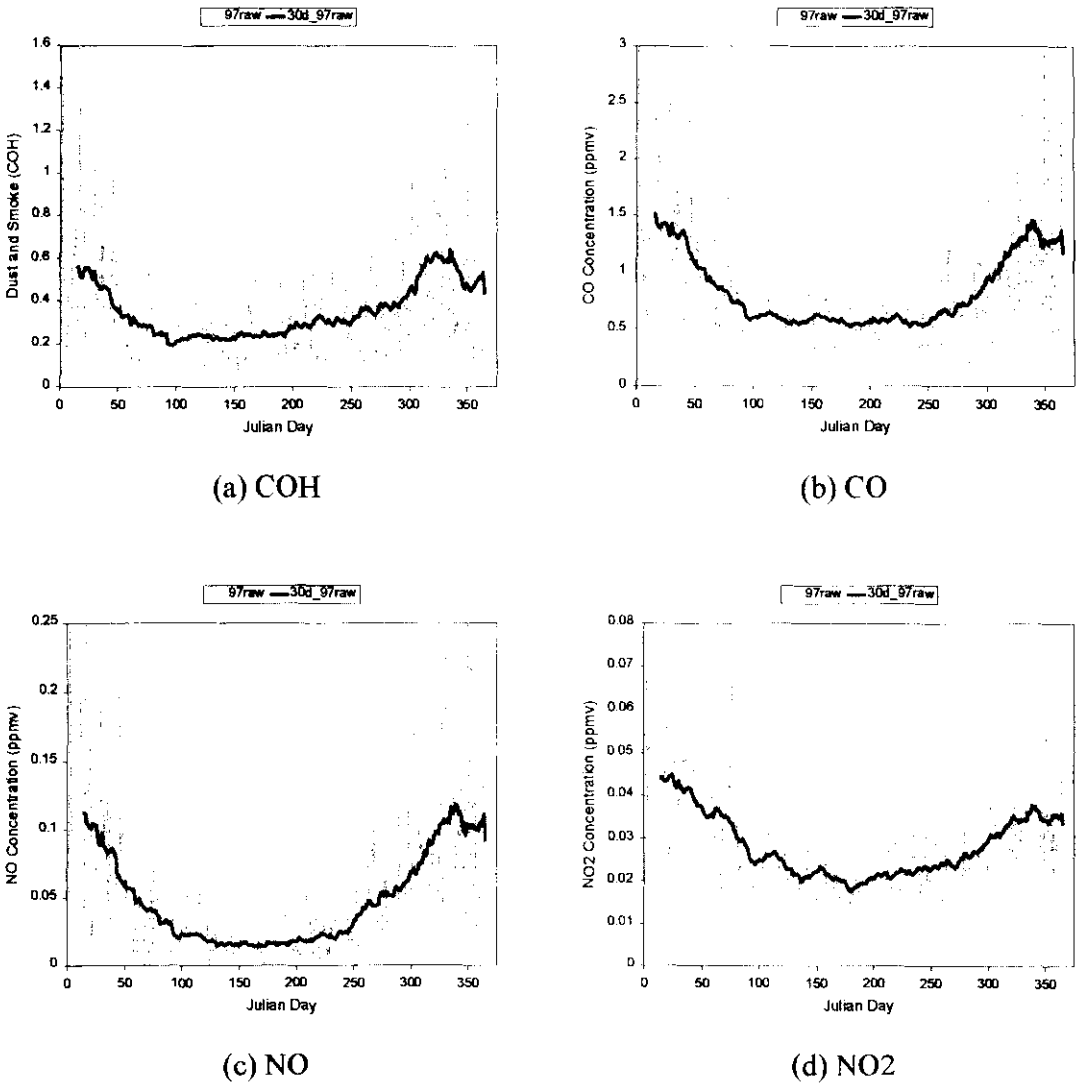
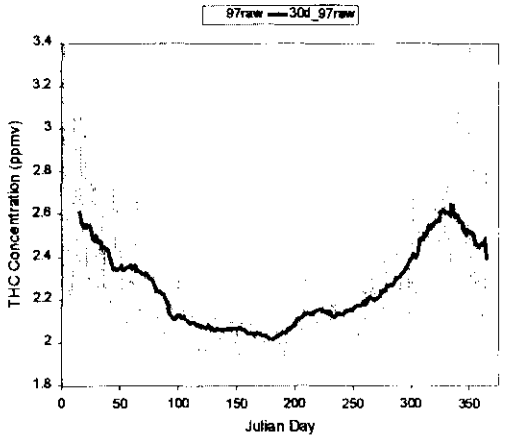
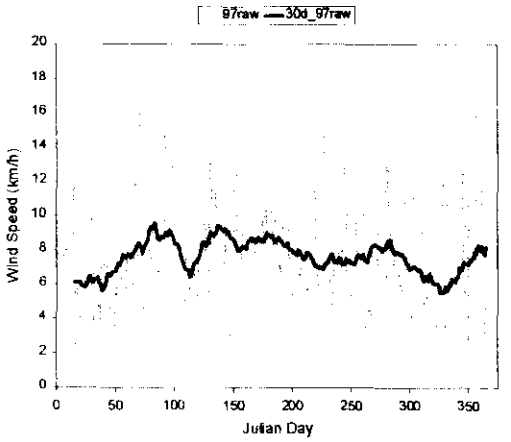


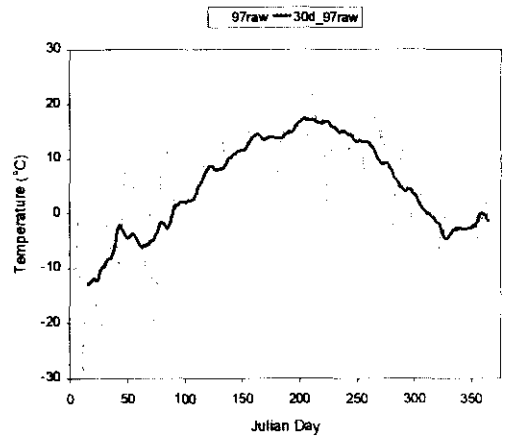
Figure 3.2 (i)
Time series and annual trends of the 1997 chemical and meteorological variables. The thick solid lines (blue) are obtained from 30-day moving average values (30d_YYraw; YY is the last two digits of the year and 'raw' denotes the original data) to illustrate the annual trends.



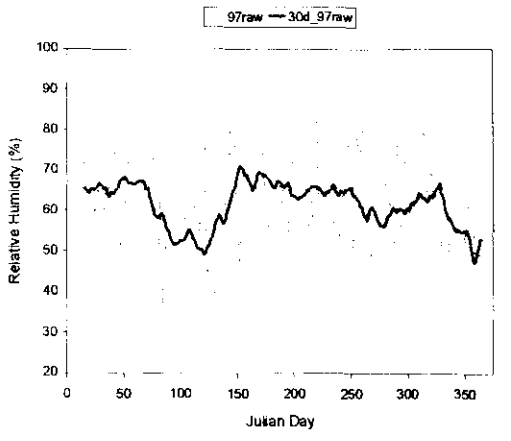
(e) THC



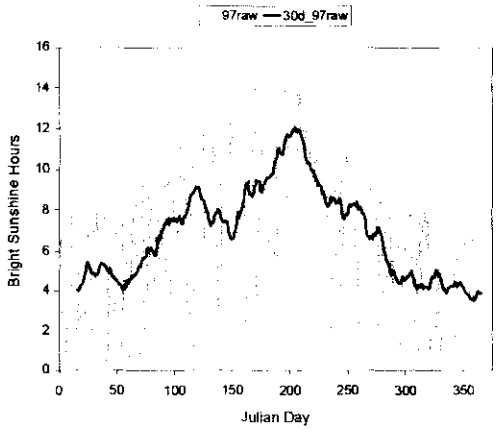
(f) WSPD



(g) Tavg



(h) RHavg



(i) bSUN

Figure 3.2 (ii)
Time series and annual trends of the 1997 chemical and meteorological variables. The thick solid lines (blue) are obtained from 30-day moving average values (30d_YYraw; YY is the last two digits of the year and 'raw' denotes the original data) to illustrate the annual trends.

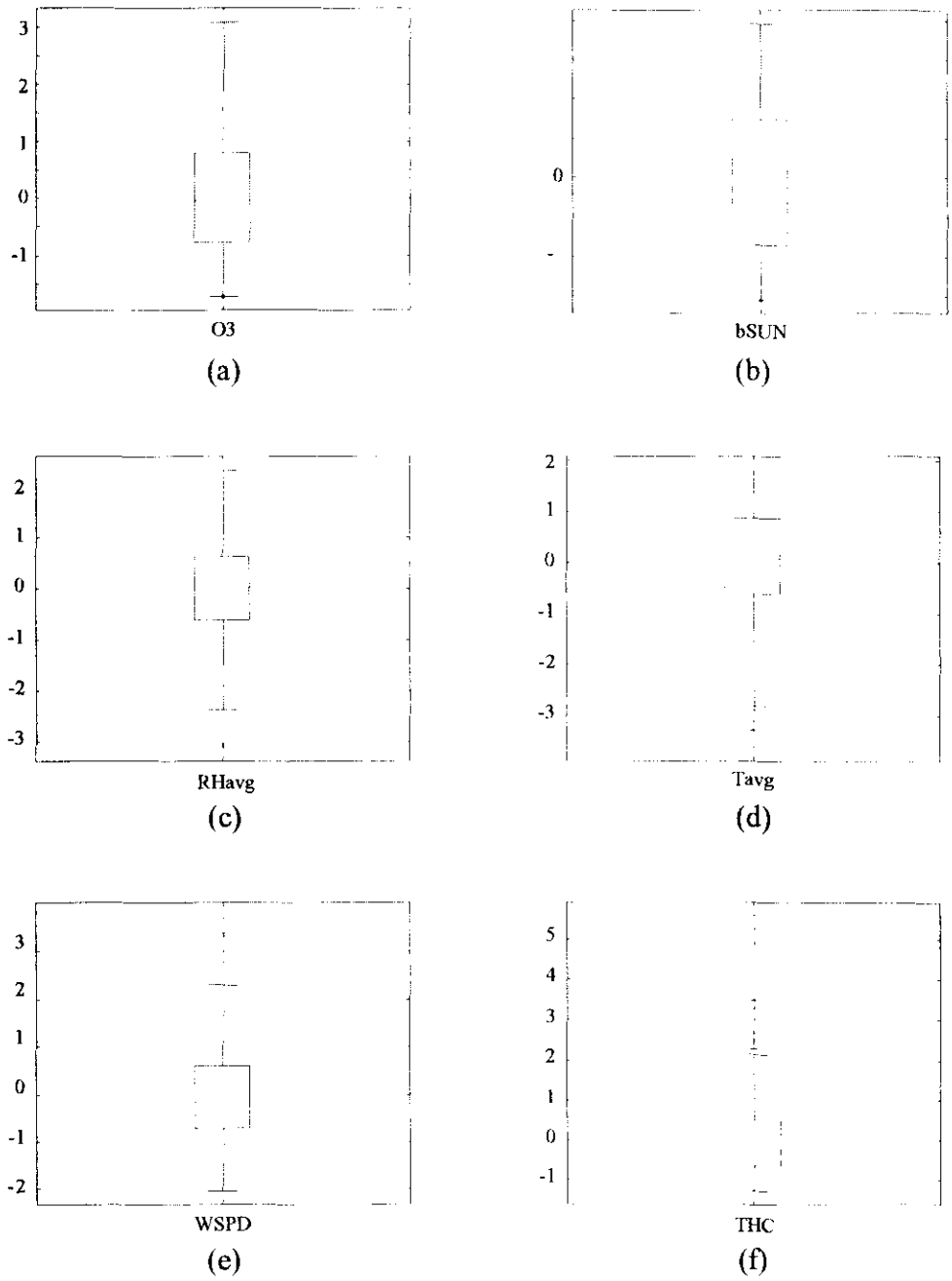


Figure 3.3 (i)
Box plots of meteorological/chemical variables (1997) to show the spread around the mean. All variables are standardized to zero mean and unit variance. The 25th, median and 75th percentiles of the distributions are illustrated by the lower, middle and upper horizontal lines of the box. The outliers are shown by plus (+) symbol.

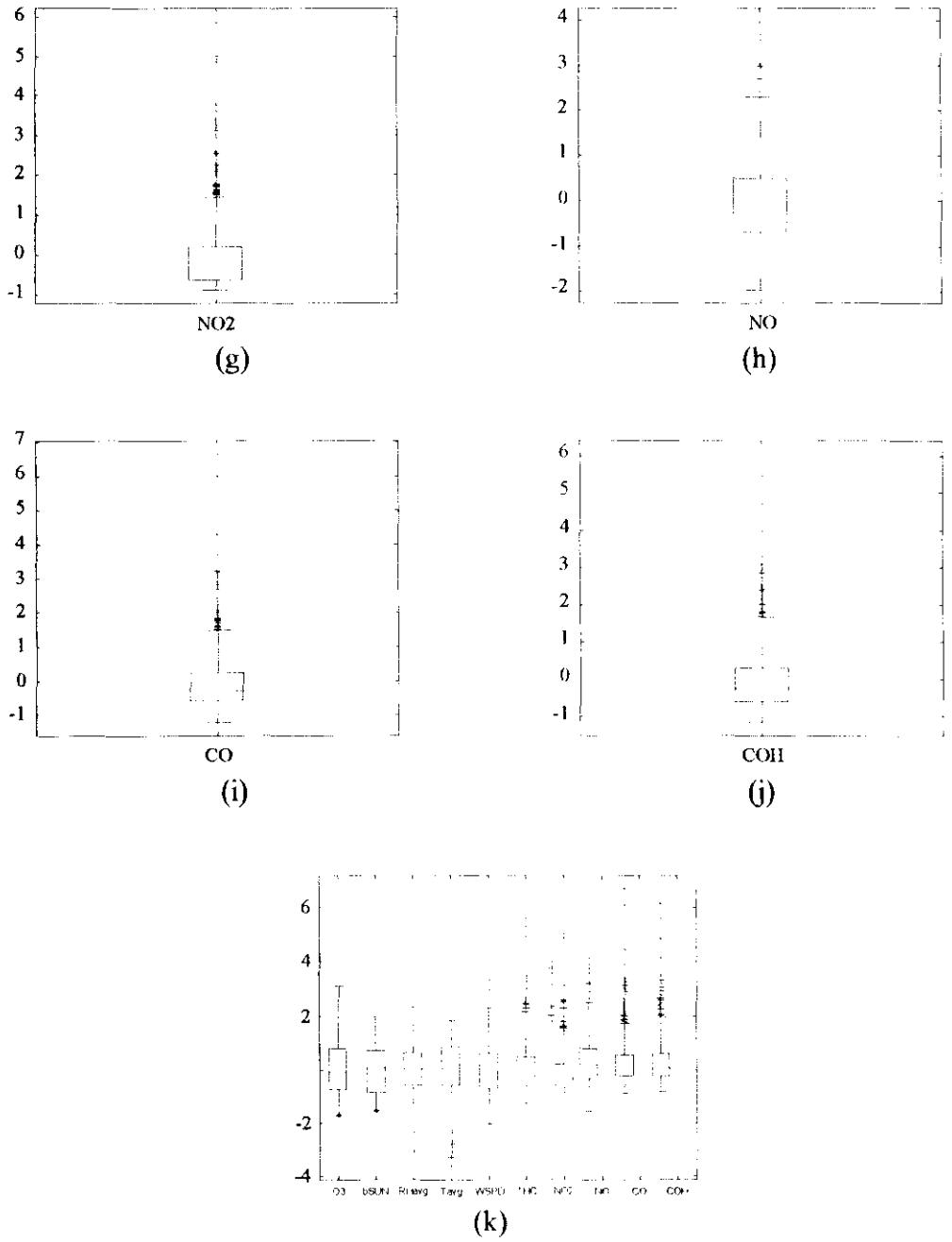


Figure 3.3 (ii)
Box plots of meteorological/chemical variables (1997) to show the spread around the mean. All variables are standardized to zero mean and unit variance. The 25th, median and 75th percentiles of the distributions are illustrated by the lower, middle and upper horizontal lines of the box. The outliers are shown by plus (+) symbol.

CHAPTER 4

REGRESSION APPROACH

Simple linear fitting based on historical or similar ozone records can be first utilized in the statistical approach. The linear fit result obtained from, say, 1997 linear analysis can be applied for predicting ozone trends for 1998-2000. However, the outcomes from the previous chapter illustrate the inadequacy of univariate data (ozone only) in explaining the temporal phenomena. The nonlinear process of ozone formation is better treated in a multivariate sense using correlation coefficients between ozone and the predictor variables, as well as the complete secondary information. Such prediction can be performed via regression accounting for a linear combination of positively correlated variables with ozone.

4.1 Simple Linear Fit of Ozone Data

As a preliminary assessment of the ozone trend, a simple linear fitting is performed. Basically the time series of ozone in a particular year is divided into several windows in which they are "best" represented by linear lines. This decision primarily depends on the annual trends, visualized in the time series plots (see Figure 2.1). Recall that the thirty-day moving averages (30dMA) $W_\alpha(t_j)$ are defined as:

$$W_\alpha(t_j) = \frac{1}{2k} \sum_{i=j-k+1}^{j+k} Z_\alpha(t_i),$$

$$\alpha = 1, \ldots, 4; j = 15, \ldots, J = 365; k = 15$$

where $Z_\alpha(t_j)$ are the raw ozone values and the subscript α refers to a set of ozone data for the year: (1) 1997, (2) 1998, (3) 1999, and (4) 2000; $Z_1(t_j)$ are used as the base case for predicting the 1997 values (validation) and also the 1998-2000 values; the denominator $2k$ represents the length of the moving window.

From the time series plot of ozone in 1997 (Figure 4.1a), there are roughly four windows in which the 30dMA values $W_\alpha(t_j)$ can be expressed linearly. The piecewise linear fit is performed for the following time intervals: (1) 1-100, (2) 101-210, (3) 211-330, and (4) 331-365. For the simple linear equations, intercepts $c_{\alpha\beta}$ are taken as the values on the first day of the respective windows, except for the first window where the intercept is graphically extrapolated to be 0.005 ppmv. The slopes $m_{\alpha\beta}$ are calculated using the first $y_\beta(t_F)$ and the last $y_\beta(t_L)$ values in the respective Julian days (Jdays) t_F and t_L within each corresponding windows, i.e.,

$$m_{\alpha\beta} = \frac{y_{\alpha\beta}(t_F) - y_{\alpha\beta}(t_L)}{t_F - t_L}, \quad \alpha = \beta = 1, \ldots, 4 \tag{4.1}$$

where the index α refers to the ozone data for the year: (1) 1997, (2) 1998, (3) 1999, and (4) 2000; and β denotes the order (left to right) of windows in each particular year. For example, the validation on the same year (1997) and prediction for the years 1998-2000 can be easily performed using this linear fit formula:

$$\hat{y}_{\alpha\beta}(t_j) = m_{\alpha\beta} t_j + c_{\alpha\beta}, \tag{4.2}$$

$$\alpha = \beta = 1, \ldots, 4; j = 1, \ldots, J = 365$$

where $\hat{y}_{\alpha\beta}(t_j)$ are the inferred values on specific days t_j. Note that the above expression should be applied with care because it is not always smooth at the borders of windows. Discontinuities may occur if the data are highly fluctuated. This happens, for example,

when the 30dMA-ozone concentration at Jday 100 is much higher (or lower) than the average value at Jday 101.

The results from the linear analysis are plotted in Figure 4.1. As expected, the piecewise linear fit represents the general trends of the 30dMA ozone values $W_\alpha(t_j)$. Specific features like the second peak in 1998 are not represented correctly by the fitting parameters computed using the 1997 data. The results predicted for 1999 and 2000 do not reflect the increase of the ozone trends due to the difference in average ozone concentrations in those years that cannot be anticipated *a priori*, based on the 1997 ozone data. In addition, the last linear line unsuccessfully predicts the stationary ozone trend at the end of year 2000. It must be emphasized that better linear fits could be obtained by establishing the slope and intercept using appropriate annual data, e.g., using annual summaries of 1999 to predict the trend in the same year; however, that would render the linear fitting method as inadequate to be a prediction tool.

The coefficient of determination (R^2) values for all years are calculated and tabulated below in Table 4.1:

Table 4.1
The R^2 values for simple linear fitting analysis.

Year	R^2
1997	0.94
1998	0.83
1999	0.84
2000	0.90

The R^2 statistic measures the mismatch between the predicted values $\hat{y}_{\alpha\beta}(t_j)$ and the actual 30dMA values $W_\alpha(t_j)$. Typically, R^2 is bounded to within [0, 1] where $R^2 = 1$ refers to the ideal situation in which the prediction exactly identifies the original data. The R^2 values for all simple linear fitting cases are quite high, which is expected since the approach represents the mean fit of over and under-estimations as shown in Figure 4.2. It

should be noted however that ozone phenomena are highly nonlinear and complex; hence representing them with simple linear expressions based on historical ozone data hardly makes physical sense. The simple linear fit approach must be augmented to include the secondary variables (covariates) such as wind speed, bright sunshine hours and average temperature.

4.2 Employing Secondary Variables

Prior to implementing the multivariate regression analysis, it is imperative for us to have some ideas on how the covariates correlate with ozone, and to know whether they share similar distributions. This may be accomplished by respectively analyzing scatter- and quantile-quantile (Q-Q) plots for the bivariate cases. If ozone and its covariates (predictors) are well correlated, all points in the scatterplots will fall on a linear (straight) line. The slope of this line may be positive or negative depending on the effects of these covariates on ozone. Similarly, if all points in the Q-Q plot are linear, ozone and its respective predictors share exactly the same distributions. Figure 4.3 shows both the scatter- and Q-Q plots for all cases in 1997. As evidenced in the scatterplots, ozone is negatively correlated with dust and smoke (COH), carbon monoxide (CO), nitric oxide (NO), nitrogen dioxide (NO2) and total hydrocarbon (THC); the positive association is only found with respect to the wind speed (WSPD). The effects of other variables, i.e., daily average temperature (Tavg), average relative humidity (RHavg) and bright sunshine hours (bSUN), on ozone are less obvious and therefore require further statistical analysis.

Knowing the correlation coefficients between covariates and the response variable (ozone), the predictions for the subsequent years can be made with the help of covariate data in the corresponding years. However, all variables need to be standardized *a priori* due to the large difference in the order of magnitude between the data. The process of standardization of random variables RV $Z_\beta(t_j)$ can be easily accomplished using a simple procedure explained in Chapter 3 of this thesis. Basically, the standardized RV $X_\beta(t_j)$ are obtained by subtracting the respective stationary means $\overline{Z}_\beta$ from the RV $Z_\beta(t_j)$ and dividing by the standard deviations s_{Z_β} as follows:

$$X_{\beta}(t_j) = \frac{Z_{\beta}(t_j) - \overline{Z}_{\beta}}{s_{Z_{\beta}}} \tag{4.3}$$

$$\forall \beta = 0, \ldots, 9;\ j \in J = 365$$

where index $\beta = 0$ refers to the response variable (i.e., ozone concentration); the rests of the predictor variables are denoted by the indices $\beta = 1, \ldots, 9$ as follows: (1) amount of dust and smoke, (2) concentration of carbon monoxide, (3) concentration of nitric oxide, (4) concentration of nitrogen dioxide, (5) concentration of total hydrocarbon, (6) wind speed, (7) average temperature, (8) average relative humidity, and (9) bright sunshine hours. The multivariate regression for the standardized ozone values $X_o^*(t_j)$ can then be written as a linear combination of the standardized meteorological and chemical variables $X_{\beta}(t_j)$ as the following:

$$X_o^*(t_j) = \sum_{\beta=1}^{N} \rho_{\beta o} X_{\beta}(t_j) \tag{4.4}$$

$$N = 9;\ j \in J = 365$$

The variable-specific correlation coefficients $\rho_{\beta o}$ are obtained from the relations between the original (raw) ozone $Z_o(t_j)$ and its predictor variables $Z_{\beta}(t_j)$, i.e.,

$$\rho_{\beta o} = \frac{Cov\{Z_o, Z_{\beta}\}}{\sqrt{Var\{Z_o\} Var\{Z_{\beta}\}}} \in [-1, +1] \tag{4.5}$$

where the numerator is the covariance between the response $Z_o(t_j)$ and corresponding predictors $Z_{\beta}(t_j)$, defined as $Cov\{Z_o, Z_{\beta}\} = \frac{1}{J}\sum_{j=1}^{J}(Z_{oj} - \overline{Z}_o)(Z_{\beta j} - \overline{Z}_{\beta})$, $\beta = 1, \ldots, 9$; J is the number of data or, in this case, observed days. The denominator consists of the square root of the variances (i.e., standard deviations), $Var\{Z_{\beta}\} = \frac{1}{J}\sum_{j=1}^{J}(Z_j - \overline{Z}_{\beta})^2$, $\beta = 0, \ldots, 9$. In other words, the correlation coefficients measure the linear dependency

between the variables where $\rho_{\beta o} = \pm 1$ denotes the perfect linear relationship, either positive or negative, and $\rho_{\beta o} = 0$ signifies no linear dependence between the two variables. The correlation coefficients between ozone and its corresponding predictors (covariates) for the entire four-year period (1997-2000) are summarized in Table 4.2 below:

Table 4.2
Correlation coefficients between ozone and chemical/meteorological variables.

No.	Variables	ρ (1997)	ρ (1998)	ρ (1999)	ρ (2000)
1.	O3 – bSUN	0.35	0.42	0.36	0.40
2.	O3 – RHavg	-0.28	-0.18	-0.16	-0.20
3.	O3 – Tavg	0.38	0.51	0.36	0.45
4.	O3 – WSPD	0.61	0.52	0.55	0.56
5.	O3 – THC	-0.68	-0.67	-0.45	-0.58
6.	O3 – NO2	-0.63	-0.53	-0.61	-0.55
7.	O3 – NO	-0.67	-0.66	-0.64	-0.65
8.	O3 – CO	-0.62	-0.62	-0.63	-0.58
9.	O3 – COH	-0.61	-0.47	-0.61	-0.61

It is important to note that the linear estimator (4.4) has been determined to be unbiased, which can be proven from the following:

$$X_o^* = \sum_{\alpha=1}^{N} \lambda_{\alpha o} X_\alpha \tag{4.6}$$

where the estimator X_o^* is defined as a linear combination of the standardized covariate data X_α weighted with arbitrary weights $\lambda_{\alpha o}$. The residual error with respect to the "true" unknown value X_o can then be written as:

$$X_o - X_o^* = X_o - \sum_{\alpha=1}^{N} \lambda_{\alpha o} X_\alpha$$

$$= \sum_{\alpha=0}^{N} \nu_{\alpha o} X_{\alpha} \tag{4.7}$$

where $\nu_{\alpha o} = 1$ if $\alpha = 0$, and $\nu_{\alpha o} = -\lambda_{\alpha o}$ for all $\alpha = 1, \ldots, N$. The weights $\lambda_{\alpha o}$ can be obtained by minimizing the error variance:

$$Var\{X_o - X_o^*\} = \sum_{\alpha=0}^{N} \sum_{\beta=0}^{N} \nu_{\alpha o} \nu_{\beta o} C_{\alpha\beta} \tag{4.8a}$$

which is the result after employing the linear operator property of the expected value. In the traditional linear regression approach, the covariance between data events α and β ($\neq \alpha$) are ignored; only the covariance between the data 'α' and unknown 'o' is considered. This amounts to setting $C_{\alpha\beta} = 0$ for all $\alpha, \beta \neq 0$ and $\alpha \neq \beta$. As a result, the error variance becomes:

$$Var\{X_o - X_o^*\} = 2\sum_{\alpha=0}^{N} \nu_{\alpha o} C_{\alpha o} + \sum_{\alpha=0}^{N} \nu_{\alpha o}^2 C_{\alpha\alpha} \tag{4.8b}$$

Following a standard procedure, the minimization of the error variance can be achieved by taking the first derivative of expression (4.8b):

$$\frac{\partial}{\partial \nu_{\alpha o}} [Var\{X_o - X_o^*\}] = 2C_{\alpha o} + 2\nu_{\alpha o} C_{\alpha\alpha}$$

Setting the above expression to zero, yielding:

$$\nu_{\alpha o} = -\frac{C_{\alpha o}}{C_{\alpha\alpha}} \tag{4.9}$$

To verify a minimum value, the second derivative of Eq. (4.8b) must be evaluated:

$$\frac{\partial^2}{\partial \nu_{\alpha o}^2} [Var\{X_o - X_o^*\}] = 2C_{\alpha\alpha}$$

which will always be positive since $C_{\alpha\alpha} > 0$, indicating that the weights $\nu_{\alpha o} = -C_{\alpha o}/C_{\alpha\alpha}$ do indeed result in a minimum error variance. Recall that the correlation coefficient is defined as:

$$\rho_{\alpha o} = \frac{C_{\alpha o}}{\sqrt{C_{\alpha\alpha} C_{oo}}} \tag{4.10}$$

Since data standardization is performed at unit variance, i.e., $C_{\alpha\alpha} = Var\{X_\alpha\} = 1$ and $C_{oo} = Var\{X_o\} = 1$, the correlation coefficients $\rho_{\alpha o} = C_{\alpha o}$. Also, by recognizing that the weights $v_{\alpha o} = -\lambda_{\alpha o}$ for all $\alpha \neq 0$ and employing Eq. (4.9), the unbiased linear estimator is obtained exactly as expression (4.4) above.

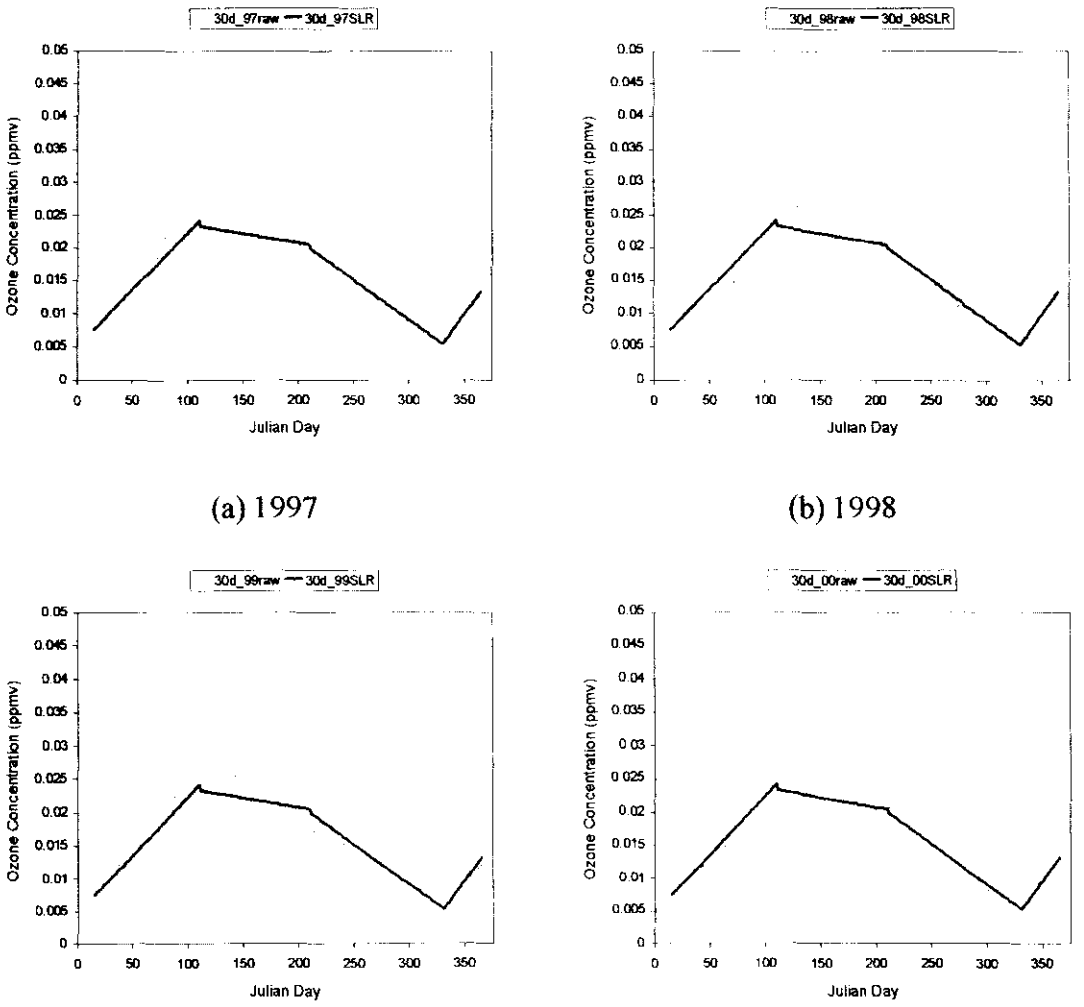
As previously discussed in Chapter 2, all covariates except for COH have been determined by either physicochemical models or experimental results to have direct influence on ozone concentrations. However, the weights associated with the negatively correlated variables will offset the influence of the positively correlated variables. Hence only WSPD, Tavg and bSUN with the respective correlation coefficients of 0.61, 0.38 and 0.35 for the year 1997 are considered in the multivariate regression of ozone. This regression approach performed well in the case of validating the 1997 ozone data and also those in 1998-2000 as shown in Figure 4.4. Here the resultant outputs are highly fluctuated just as the original time series data after converting the resultant standardized values of ozone $X_\alpha(t_j)$ back to those of raw values $Z_\alpha(t_j)$:

$$Z_\alpha(t_j) = s_{z\alpha} X_\alpha(t_j) + \overline{Z}_\alpha \tag{4.11}$$

$$\alpha = 1, \ldots, 4, j \in J = 365$$

where $s_{z\alpha}$ and $\overline{Z}_\alpha$ respectively denote the standard deviation and stationary mean of ozone in a particular year α : (1) 1997, (2) 1998, (3) 1999, and (4) 2000. This approach is promising because the response variable (ozone) can be estimated from its positively correlated predictors. Hence there is a good possibility for accurate prediction at a location where ozone measurement is unavailable but those of secondary variables are, just by employing the correlation coefficients $\rho_{\alpha o}$ obtained from historical bivariate scatterplots. Note that in this multivariate regression analysis, the correlation coefficients

between the response variable (ozone) and its covariates are implemented only at the same time instants t_j. However, it is conceivable that there may be a time lag between a covariate event and a peak in the annual ozone trend. This factor can only be accounted for by modeling a temporal statistic such as a variogram, which is discussed in the next chapter.



(a) 1997

(b) 1998

(c) 1999

(d) 2000

Figure 4.1
The thirty-day averages (30dMA) of the raw data (observation) as compared with the results from simple fit analysis

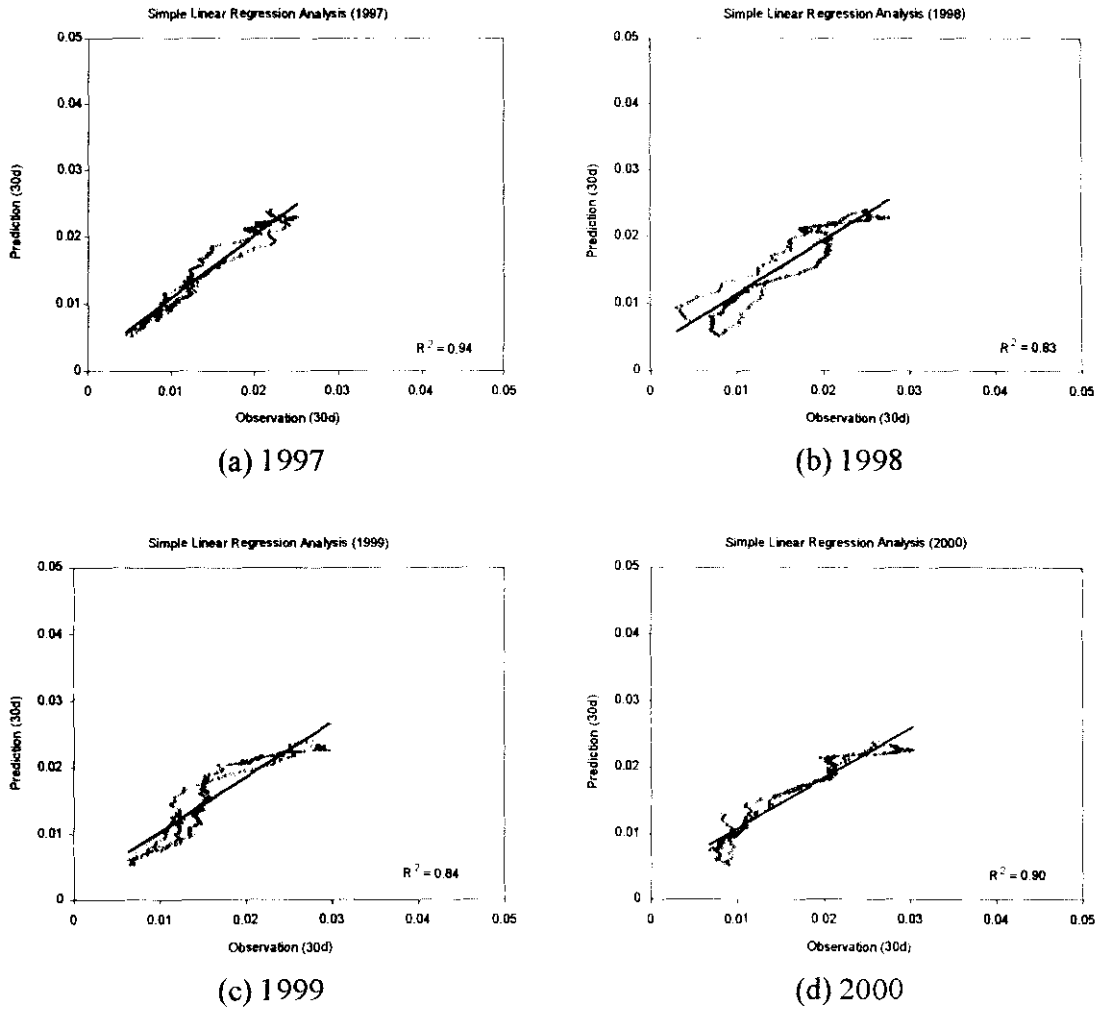
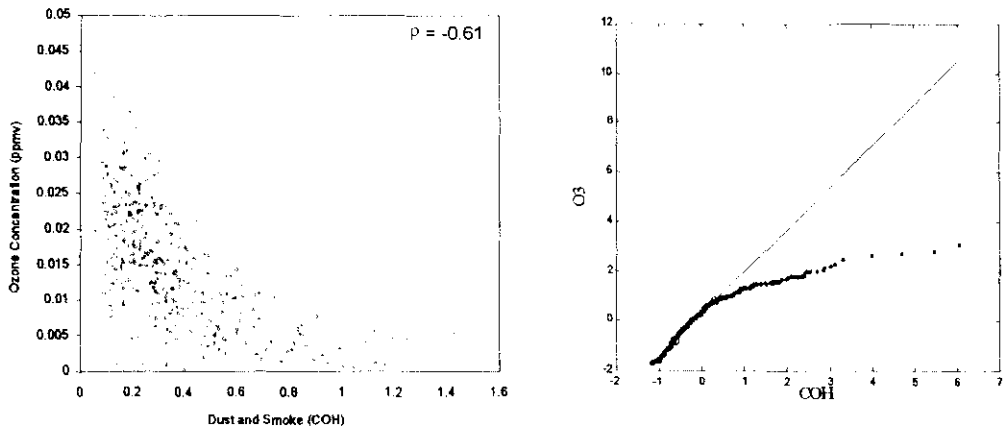
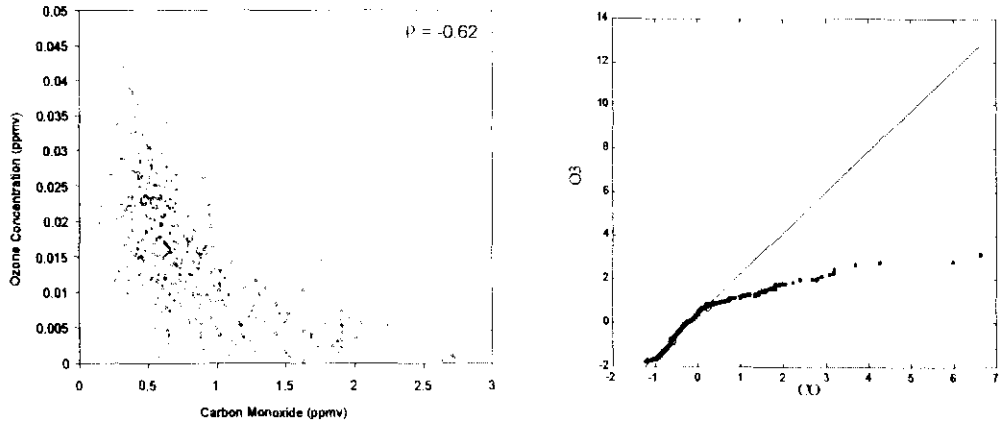


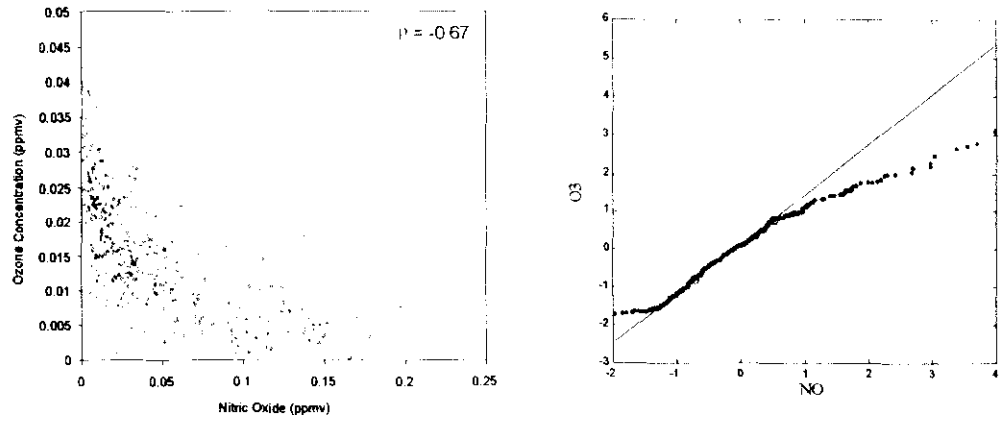
Figure 4.2
Bivariate scatterplots of the thirty-day averages (30dMA) of the raw data (observation) when compared with the results from simple linear fit analysis.



(a) O3-COH



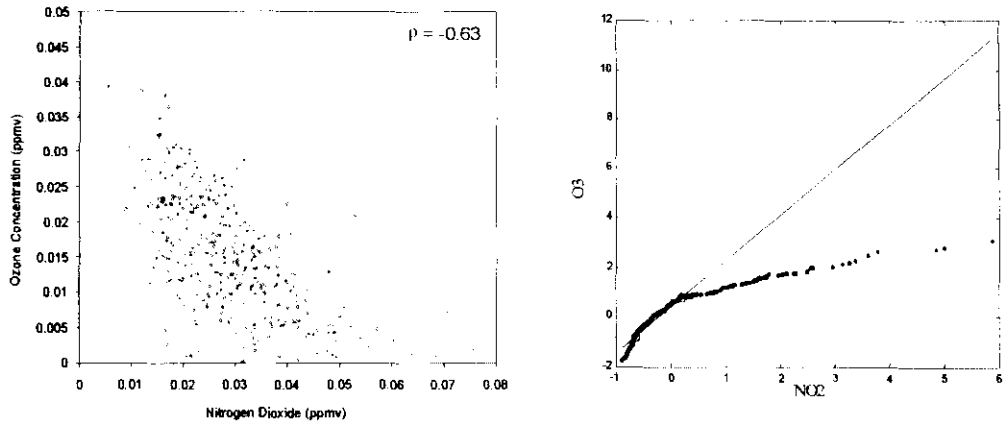
(b) O3-CO



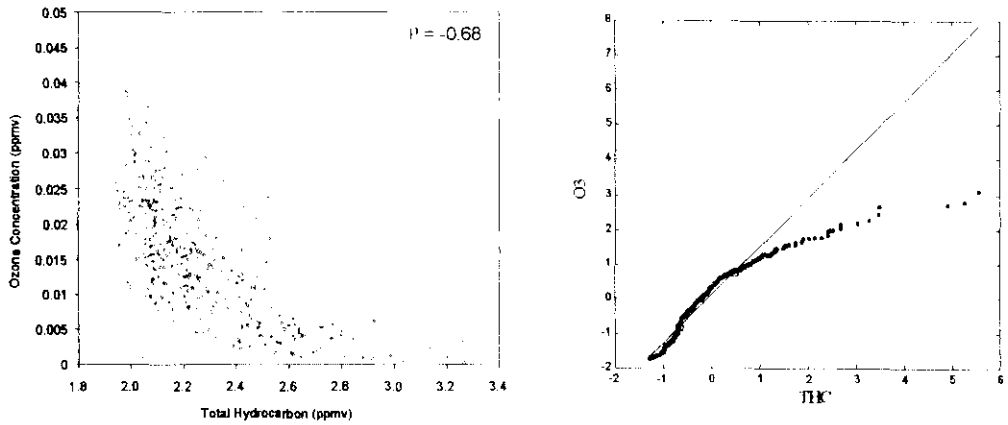
(c) O3-NO

Figure 4.3 (i)

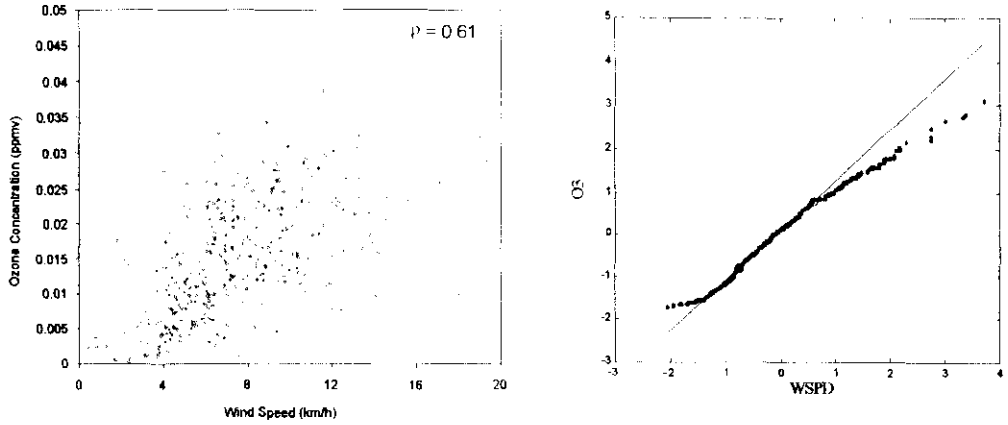
Cross correlation between ozone and meteorological/chemical variables for 1997. The correlation coefficient ρ for individual case is shown on the scatter plot [**LEFT**]. The bivariate distribution is illustrated on the Q-Q plot [**RIGHT**].



(d) O3-NO2



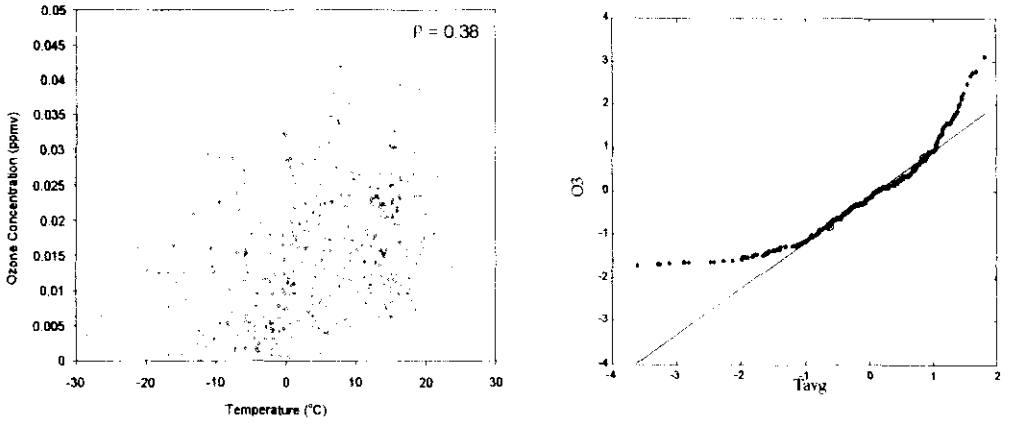
(e) O3-THC



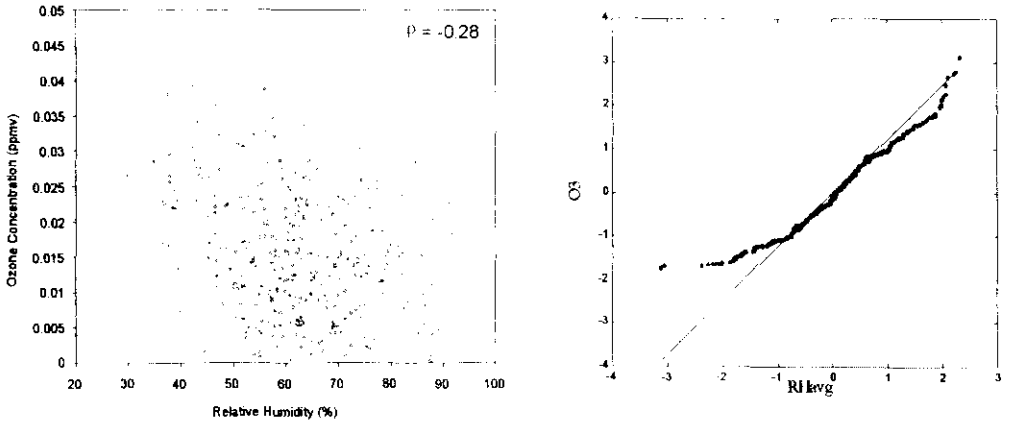
(f) O3-WSPD

Figure 4.3 (ii)

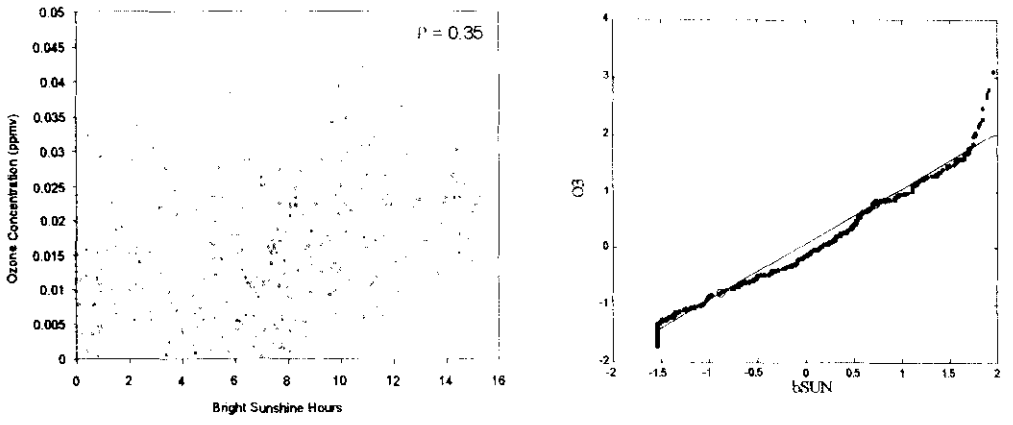
Cross correlation between ozone and meteorological/chemical variables for 1997. The correlation coefficient ρ for individual case is shown on the scatter plot [**LEFT**]. The bivariate distribution is illustrated on the Q-Q plot [**RIGHT**].



(g) O3-Tavg



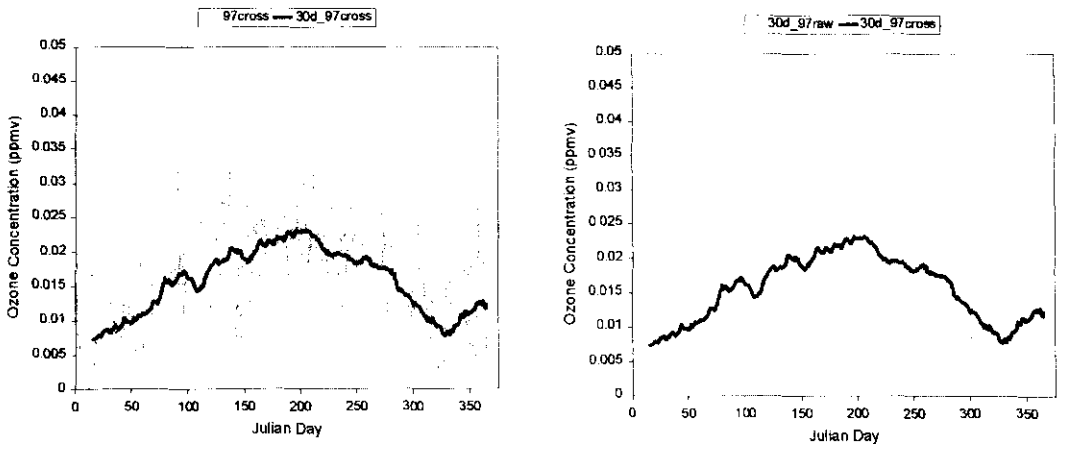
(h) O3-RHavg



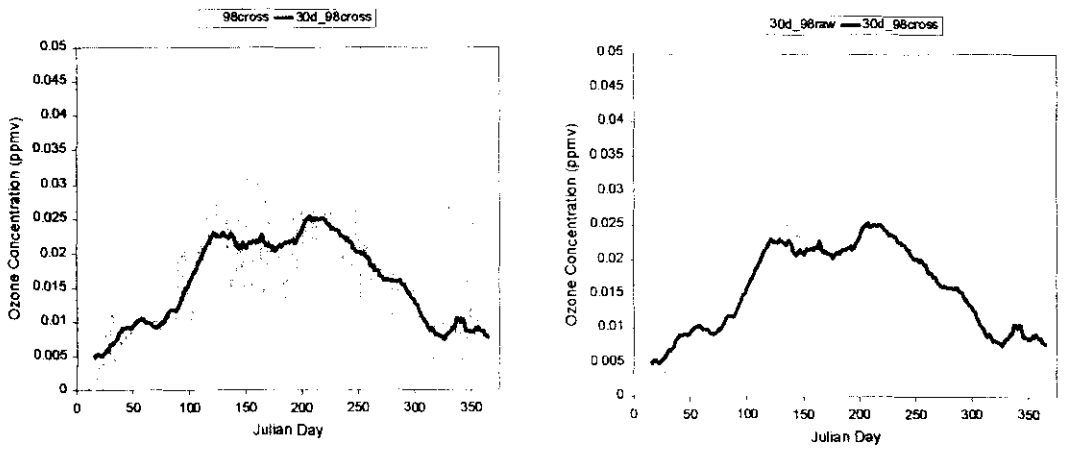
(i) O3-bSUN

Figure 4.3 (iii)

Cross correlation between ozone and meteorological/chemical variables. The correlation coefficient ρ for individual case is shown on the scatter plot [LEFT]. The bivariate distribution is illustrated on the Q-Q plot [RIGHT].

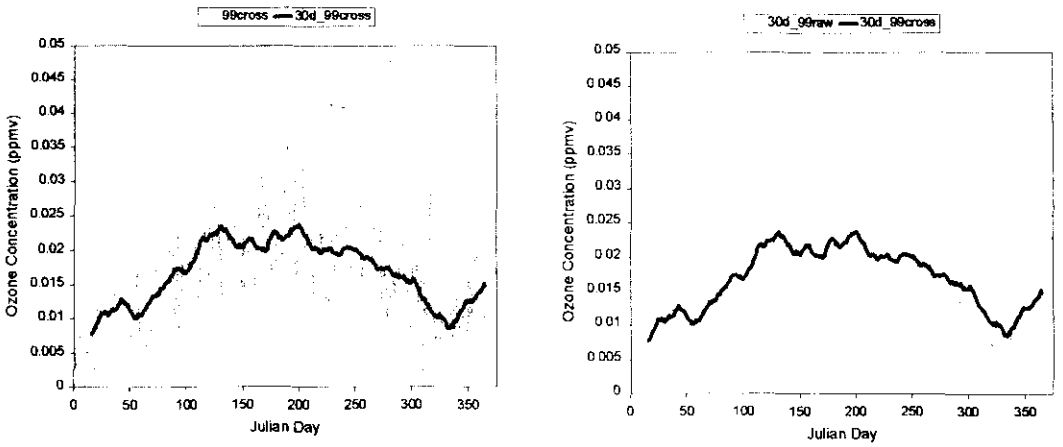


(a) 1997

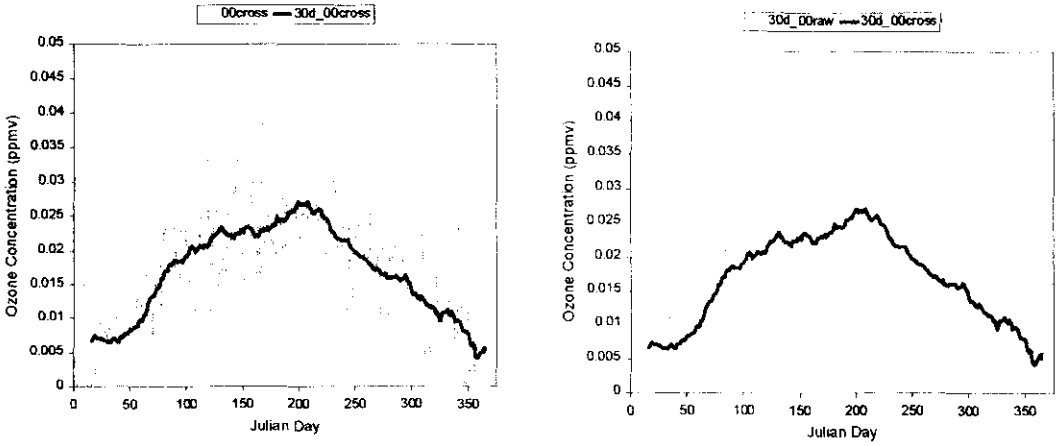


(b) 1998

Figure 4.4 (i)
Employing secondary information in the linear regression analysis. Only the positively correlated variables (WSPD, Tavg and bSUN) are used for inferring ozone concentration.



(c) 1999



(d) 2000

Figure 4.4 (ii)
Employing secondary information in the linear regression analysis. Only the positively correlated variables (WSPD, Tavg and bSUN) are used for inferring ozone concentration.

CHAPTER 5

KRIGING APPROACHES

To gain familiarity with the geostatistical approaches, the basic concepts of stochastic variables are introduced at the beginning of this chapter. The theoretical and modeling aspects of a variogram, a form of "two-point" temporal correlation, is discussed in detail to acknowledge its importance in the kriging procedure. Several kriging algorithms (Deutsch and Journel, 1998), in particular, simple kriging (SK), ordinary kriging (OK), simple cokriging (SCK) and ordinary cokriging (COK), are also discussed. Two of such algorithms, i.e., OK and COK, are implemented in various cases of ozone prediction. In this research study, the variogram, modeled using the 1997 standardized data, is applied for inferring ozone values in 1998-2000. Note that the OK algorithm needs a licit auto-variogram model in order to ensure a positive variance of the estimation points (unknowns). Similarly, the cokriging algorithm requires joint positive-definite auto and cross-variograms, usually modeled via the linear model of coregionalization (LMC).

5.1 The Temporal Framework

A temporal random variable (RV) $Z(t)$ is a variable that can take multiple outcomes (realizations) at any time instant $t \in T$, according to a certain probability density function (pdf). Also this RV $Z(t)$ is fully identified by its cumulative distribution function (cdf), which defines the probability that the variable Z at instant t in time does not exceed a given threshold z:

$$F(t;z) = \text{Prob}\{Z(t) \leq z\}, \qquad \forall z,\ t \in T \tag{5.1}$$

A temporal random function (RF), also termed Z, is defined as a collection of RVs $Z(t)$ that are occurring simultaneously. These RVs $Z(t)$ must be taken jointly resulting in the

regionalization, in a temporal sense, of the RF Z. Therefore, the RF Z is characterized by a multivariate probability distribution:

$$F(t_1,...,t_T;z_1,...z_N)=\text{Prob}\{Z(t_1)\le z_1,...Z(t_T)\le z_N\},\ \forall z,t\in T \tag{5.2}$$

where T is the number of definite stationary time spans, e.g., seasons or years. This set of T-variate cdfs is termed the temporal law of the RF Z.

The inference of the above multivariate distribution requires multiple outcomes of the RF Z at the same instant in time $t \in T$. Since such repeated outcomes are impossible to arrive at, decision on stationarity that amounts to invariance with temporal translation, is employed. Under stationarity, for example, a "two-point" scattergram is inferred by pooling together sets of data points separated by similar temporal lags $\tau = t - t'$, $t \in T$. The moment of inertia of the "two-point" scattergram between $Z(t)$ and $Z(t + \tau)$, $t \in T$, is called a variogram and estimated as:

$$2\gamma(\tau)=\frac{1}{N(\tau)}\sum_{i=1}^{N(\tau)}[z(t_i)-z(t_i+\tau)]^2 \tag{5.3}$$

where $N(\tau)$ is the number of tail-head pairs in the scattergram. The quantity $\gamma(\tau)$, half of the moment inertia above, is called a semivariogram (henceforth, termed a variogram for convenience), and is related to the more conventional covariance function $C(\tau)$ and variance $C(0)$ through:

$$\gamma(\tau) = C(0) - C(\tau) \tag{5.4}$$

Note that expression (5.4) is important because it is the covariance function that must be ensured positive-definite, not the variogram. To clarify, a positive variogram value will always exist but the same is not true for the covariance as in the case of the power law variogram model (Section 5.2), especially when the exponent (power) $\omega \ge 1$.

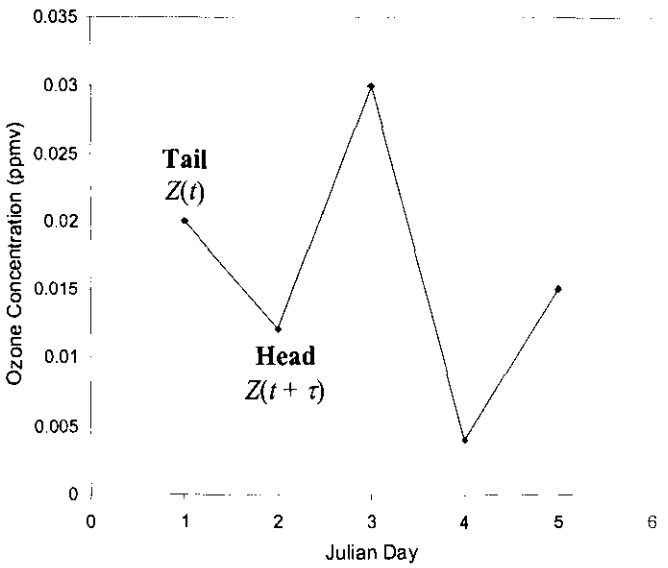


Figure 5.1
A tail $Z(t)$ and a head $Z(t + \tau)$ value separated by a temporal lag $\tau = 1$ day, used in the variogram naming convention.

The RF $Z(t)$ is termed the tail value of a pair with expected value $E\{Z(t)\} = m_{-\tau}$ while the RF $Z(t + \tau)$ is called the head value of a pair with mean $E\{Z(t + \tau)\} = m_{+\tau}$. The covariance measure $C(\tau)$ can be written as:

$$C(\tau) = \frac{1}{N(\tau)} \sum_{i=1}^{N(\tau)} Z(t_i) \cdot Z(t_i + \tau) - m_{-\tau} m_{+\tau} \tag{5.5a}$$

where:

$$m_{-\tau} = \frac{1}{N(\tau)} \sum_{i=1}^{N(\tau)} Z(t_i) \tag{b}$$

$$m_{+\tau} = \frac{1}{N(\tau)} \sum_{i=1}^{N(\tau)} Z(t_i + \tau) \tag{c}$$

and $N(\tau)$ is the number of tail-head pairs in a set of measurements. The covariance $C(\tau)$ can be standardized using the variance of the tail $\sigma_{-\tau}^2$ and head $\sigma_{+\tau}^2$ values. The resultant correlogram $\rho(\tau)$ is written as:

$$\rho(\tau) = \frac{C(\tau)}{\sqrt{\sigma_{-\tau}^2 \sigma_{+\tau}^2}} \in [-1, +1] \tag{5.6a}$$

where:

$$\sigma_{-\tau}^2 = \frac{1}{N(\tau)} \sum_{i=1}^{N(\tau)} \left[Z(t_i) - m_{-\tau} \right]^2 \tag{b}$$

$$\sigma_{+\tau}^2 = \frac{1}{N(\tau)} \sum_{i=1}^{N(\tau)} \left[Z(t_i + \tau) - m_{+\tau} \right]^2 \tag{c}$$

Under second order stationarity, the first two moments of the RF $Z(t)$, i.e., the mean m and covariance $C(\tau)$, remain invariant under translation. Thus:

$$E\{Z(t)\} = E\{Z(t+\tau)\} = m, \quad \forall (t, \tau) \in T, \text{ and} \tag{5.7a}$$

$$Cov\{Z(t), Z(t')\} = Cov\{Z(t+\tau), Z(t'+\tau)\} = C(\tau), \quad \forall (t, t', \tau) \in T \tag{b}$$

or in words, the expected value (mean) m is independent of the lag $\tau \in T$, whereas the covariance $C(\tau)$ is strictly a function of the lag τ and is independent of the time t. As a result, the correlogram becomes:

$$\rho(\tau) = \frac{C(\tau)}{C(0)} \in [-1, +1] \tag{5.8}$$

The above correlogram $\rho(\tau)$ is an important feature because if the historical data for a temporal process, e.g., ozone concentration, is known in a particular year, and if ozone data in subsequent years are assumed to be a realization of an underlying stationary process, then the correlogram inferred based on the historical data can be applied to predict the temporal patterns (at least, in "two-point" sense) of the future profiles.

5.2 Modeling Variogram

The experimental variogram (Eq. 5.3) lacks statistical mass and is affected by outliers especially at larger temporal lag spacing τ, resulting in considerable fluctuations in the

variogram values. A more practical approach is to use a smooth mathematical function (model) that can best fit the experimental variogram. In order to ensure physical plausibility of the RF $Z(t)$ and uniqueness of the estimated values, the variogram $\gamma(\tau)$ or more precisely covariance $C(\tau)$ function must be positive definite. For this reason, experimental variograms are modeled using positive definite functions, as summarized in Table 5.1 below (Deutsch and Journel, 1998):

Table 5.1
Positive definite variogram models.

No.	Abbreviation	Formula
1.	$Sph\left(\frac{\tau}{a}\right)$	$\gamma(\tau)=C(0)\cdot\left[1.5\left(\frac{\tau}{a}\right)-0.5\left(\frac{\tau}{a}\right)^3\right]$ if $\tau \le a$
		$\gamma(\tau)=C(0)$, otherwise.
2.	$Exp\left(\frac{\tau}{a}\right)$	$\gamma(\tau)=C(0)\cdot\left[1-\exp\left(-\frac{3\tau}{a}\right)\right]$
3.	$Gauss\left(\frac{\tau}{a}\right)$	$\gamma(\tau)=C(0)\cdot\left\{1-\exp\left[-\left(\frac{3\tau}{a}\right)^2\right]\right\}$
4.	Power	$\gamma(\tau)=C\cdot\lvert\tau\rvert^{\omega}, \quad 0<\omega<2$
5.	Hole Effect	$\gamma(\tau)=C(0)\cdot\left[1-\cos\left(\frac{\tau}{a}\pi\right)\right]$

where τ is the temporal lag distance. Recall that the variogram (Eq. 5.3) is a measure of temporal variability, or more precisely, its value is higher when the RF $Z(t_i)$ is more dissimilar than RF $Z(t_i + \tau)$. Hence the variogram value at $\tau = 0$ is zero since there is no variability between data points when compared with themselves. However, at a short distance from the origin, the variogram may show discontinuity due to inherent variability of the spatiotemporal phenomena (e.g., the natural occurrence of gold nuggets)

or measurement procedures (e.g., finite sampling interval). This near-origin discontinuity is customarily termed nugget effect C_o. As the temporal lag distance τ increases, the variogram values become larger, signifying less correlation between data points. Theoretically, the variogram reaches a maximum value known as the sill $C(0)$, which is also the variance of the distribution. The lag distance (abscissa value) over which the sill $C(0)$ is reached is called the range of the variogram and denoted as 'a'. This range a represents the greatest distance, over which a datum is related to another within the domain of interest.

The *Exp* and *Gauss* models (Table 5.1) reach the sill $C(0)$ exponentially and their practical (effective) range is defined as the distance over which the variogram reaches 95% of the sill value. Among the five models, the power model is unique in the sense that it lacks a sill value and therefore has no covariance counterpart, indicating a fractal-typed process or lack of stationarity; consequently, the power model is defined only by a positive coefficient (slope) C and an exponent (power) $\omega \in (0, 2)$. For the hole-effect model, the range a is defined as the length from origin to the first peak of the oscillated curve. In a special hole-effect case, i.e., sinusoidal variation, the term in the parentheses may be redefined to comprise a complete period of 2π with a range $a = 365$ days, as implemented in this thesis work.

The behavior of the variogram near the origin, i.e., as $\tau \rightarrow 0^+$, must be known prior to selecting a particular model. The *Sph* model exhibits a linear behavior due to the vanishing of cubic term at small τ. If the *Exp* model is expanded as Taylor series [$e^{-x} = 1 - x + x^2/2! - x^3/3! + \ldots$], the variogram model reduces to x as τ approaches zero; therefore the *Exp* model also behaves linearly near origin. On the other hand, the *Gauss* model when expanded as Taylor series [$\exp(-x^2) = 1 - x^2 + x^4/2! - x^6/3! + \ldots$] will reduce to x^2 as $x \rightarrow 0^+$, i.e., this model approaches the origin parabolically.

The power model, $\gamma(\tau) = C \cdot |\tau|^{\omega}$, with a positive slope C and exponent (power) ω, $0 < \omega < 2$, may show any of the following three behaviors near origin; it can behave linearly ($\omega = 1$), hyperbolically ($0 < \omega < 1$) as well as semi-parabolically ($1 < \omega < 2$)

depending on the value of its exponent ω. As a special case of the power law model, a nugget effect model, $\gamma(\tau)=0$ if $\tau = 0$ and $\gamma(\tau)=C_o$ otherwise, can also be obtained by setting ω to 0^+. The (pure) nugget effect case is an indication of a temporally uncorrelated RF $Z(t)$

Often the experimental variogram exhibits fluctuations (hole effect) near the sill $C(0)$. If the oscillations decay as τ approaches infinity, a general positive definite model, termed 'sinc' model, can be utilized:

$$\gamma(\tau) = C(0) \cdot \left[1 - A\frac{\sin(\tau)}{\tau}\right] \tag{5.9}$$

where A is the first 'overshoot' amplitude and the temporal lag τ (in Julian days) is first converted to radian by employing the complete period of the oscillation ($2\pi = 365$ days), i.e., change τ to $2\pi\tau/365$. For a strongly oscillating variogram, another model, called a half-range Fourier series, may be more useful:

$$\gamma(\tau) = C_o + C(0) \cdot [1 - B\cos(\tau)] \tag{5.10}$$

where the presence of the nugget effect C_o in the variogram model is to ensure the positive definiteness of covariance function $C(\tau)$; B is the amplitude of the fluctuations, and the value of τ is measured in radians. It is interesting to note that both models, (5.9) and (5.10), show parabolic behavior near the origin. However, a major difference is that the former model rises slower than the latter does, which is easily verified from the Taylor series expansion:

(i) $\sin(\tau) = \tau - \tau^3/3! + \tau^5/5! - \tau^7/7! + \ldots$, and

(ii) $\cos(\tau) = 1 - \tau^2/2! + \tau^4/4! - \tau^6/6! + \ldots$

As $\tau \to 0^+$, the variograms models of (5.9) and (5.10) reduce to $\tau^2/6$ and $\tau^2/2$, respectively.

The temporal correlation in terms of variogram (covariance) can be incorporated into a more generalized linear regression algorithm, known as kriging.

5.3 Kriging

Kriging is a geostatistical tool for interpolation, or more specifically, the value of an attribute in unknown areas (or time) is obtained by considering the conditioning data, and the correlation between the data and unknowns (estimates) within the domain of interest. Although traditionally formulated for spatial estimations in the contexts of mining and reservoir characterization, kriging has also gained acceptance in the environmental sciences, e.g., for interpolating the temporal trends of air pollutants and spatial variations of terrestrial contaminants. In this thesis, the applications of kriging algorithms for estimating ozone temporal trends are explained in the following sections.

5.3.1 Simple Kriging

The simple kriging (SK) estimator Z^*_{SK} at each time instant t_j is the best linear unbiased estimator/predictor (BLUE/P) and is usually written as:

$$Z^*_{SK}(t_j) - m(t_j) = \sum_{i=1}^{N} \lambda_i(t_j)[Z(t_i) - m(t_i)] \tag{5.11}$$

where $m(t_j) = E\{Z(t_j)\}$, $j = 1, \ldots, J = 365$, is the "known" stationary expected value (mean) of the RF $Z(t_j)$ defined at time instant t_j; in some cases, it may simply be assumed equal to $m(t_i)$ (a model decision *not* a hypothesis) if "enough" data are available, or it can be estimated *a priori* from historical records. The locally averaged value $m(t_i) = E\{Z(t_i)\}$, $i = 1, \ldots, N$, of the random function RF $Z(t_i)$ defined at time instant t_i, may be inferred from N data points available for the purpose of estimation. However, prior to solving for the SK estimator Z^*_{SK}, kriging weights $\lambda_i(t_j)$, $\forall i = 1, .., N$, must be calculated from the following system of normal equations:

$$\sum_{k=1}^{N} \lambda_k(t_j) C(t_i - t_k) = C(t_i - t_j), \qquad \forall i = 1, \ldots, N \tag{5.12}$$

where t_j is the time instant corresponding to the estimation point (unknown). The kriging system is obtained by minimizing the estimation (error) variance σ^2_{SK}:

$$\sigma_{SK}^2(t_j) = C(0) - \sum_{i=1}^{N} \lambda_i(t_j) C(t_i - t_j) \tag{5.13}$$

where $C(0) = Var\{Z(t_i)\}$ is the sill or positive variance of the RF $Z(t)$, and is generally inferred from N sample values or readily known from the prior knowledge of the temporal process. Note that the estimation variance σ_{SK}^2 is independent of data values, i.e., homoscedastic, and therefore is not a measure of local accuracy of kriging. Other measures such as the cross-validation error must be subsequently applied to the final results in order to evaluate the goodness of data fit. Simple kriging (Eqs. 5.11-5.13) requires complete prior knowledge of:

- the stationary mean $m(t_j) = m(t_i) = m$,
- (N x N) square matrix of covariances $[C(t_i - t_k)]$: $i, k = 1, \ldots, N$, between the sample data, and
- (N x 1) vector of data-to-unknown covariance $[C(t_i - t_j)]^T$: $i = 1, \ldots, N$, where superscript T denotes the transpose operation,

The SK estimator can also be applied in the case of estimating a RF $Z(t)$ by integrating the N "hard" data with the secondary ("soft") information. As an example, consider bright sunshine hours (bSUN) and surface wind speed (WSPD) that may be well correlated with ozone but not to each other. This happens, especially, when the secondary variables, RFs Z_α and Z_β, $\alpha, \beta = 1, \ldots, N$, $\forall \alpha \neq \beta$, are weakly associated with respect to the physicochemical processes and yet exhibit almost no linear statistical dependence. In such cases, the correlation coefficient $\rho_{\alpha\beta}$ between RFs Z_α and Z_β is zero, which implies that $Cov\{Z_\alpha, Z_\beta\} = 0$ from:

$$\rho_{\alpha\beta} = \frac{Cov\{Z_\alpha, Z_\beta\}}{\sqrt{Var\{Z_\alpha\} Var\{Z_\beta\}}} \tag{5.14}$$

However, the correlation coefficients $\rho_{\alpha j}$ and $\rho_{\beta j}$ are not necessarily zero since each datum may be independently correlated with the estimation node at time instant t_j. Hence

the SK system, a variation of Eq. (5.12), may be simplified to $\lambda_\alpha \cdot Cov\{Z_\alpha, Z_\alpha\} = Cov\{Z_\alpha, Z(t_j)\}$, or after rearrangement:

$$\lambda_\alpha = \frac{Cov\{Z_\alpha, Z(t_j)\}}{Cov\{Z_\alpha, Z_\alpha\}} = \frac{Cov\{Z_\alpha, Z(t_j)\}}{Var\{Z_\alpha\}} \tag{5.15}$$

since the covariance of the same variable Z_α is nothing but its variance. Consequently, Eq. (5.14) can be combined with (5.15) to reformulate the SK estimator (5.11), or rather its variation, as:

$$\left(\frac{Z(t_j) - m(t_j)}{\sigma(t_j)}\right)^* = \sum_{\alpha=1}^{M}\sum_{k=1}^{N_\alpha} \rho_{\alpha j}(t_k - t_j)\left(\frac{Z_\alpha(t_k) - m_\alpha}{\sigma_\alpha}\right) \tag{5.16}$$

where

$$\rho_{\alpha j}(t_k - t_j) = \frac{Cov\{Z_\alpha(t_k), Z(t_j)\}}{\sigma_\alpha^2} \tag{5.17}$$

in which σ_α^2 = Var$\{Z_\alpha(t_k)\}$, the stationary variance of RF $Z_\alpha(t_k)$. Prior to prediction (future years) or validation (same year), the stationary mean $m(t_j)$ and standard deviation $\sigma(t_j)$ of the target variable $Z(t_j)$, as well as the correlation coefficients $\rho_{\alpha j}(t_k - t_j)$ must be inferred based on the information from historical or similar records. The variable-specific mean m_α and standard deviation σ_α are obtained from the available data $z_\alpha(t_k)$, $\forall \alpha = 1, \ldots, M, k = 1, \ldots, N_\alpha$.

Expression (5.16) may also be interpreted as kriging of the standardized RF $X(t)$ by assuming independence between the secondary RFs X_α and X_β,, $\alpha, \beta = 1, \ldots, M$, $\forall \alpha \neq \beta$. The coefficients $\rho_{\alpha j}(t_k - t_j)$ are merely correlograms expressed as a continuous function at temporal lag $\tau = | t_k - t_j |$, separating the datum and estimation node at time instants t_k and t_j, respectively. If regression is performed at lag $\tau = 0$, this will result in ordinary linear regression (Eq. 4.4) of the standardized RFs X_α, weighted with regression coefficients $\rho_{\alpha j}$. Note that the variable standardization is merely a process of transforming

RFs Z_α, $\alpha = 0, \ldots, M$, into standardized RFs X_α by subtracting from their respective stationary means m_α and dividing by the stationary standard deviations σ_α. The implementation of this approach for validating ozone concentrations in Calgary, Alberta throughout 1997-2000 has been successful as previously shown in Figure 4.4 of this thesis.

If the attribute α is different from that of the estimated values $Z(t_j)$, then a positive definite measure of the cross-covariance between RFs $Z_\alpha(t_k)$ and $Z(t_j)$ is required; this discussion will be deferred to the latter section, i.e., under cokriging. For the rest of this section, $Z_\alpha(t_k)$ is assumed to be the data pertaining to the same RF $Z(t_j)$. In a predictive mode, the above model (Eq. 5.16) requires the implementation of a positive definite variogram model $\gamma(\tau)$, or alternatively a covariance model, i.e.,

$$C(\tau) = C(0) - \gamma(\tau) \tag{5.18}$$

The variogram model $\gamma(\tau)$ may be selected from Table 5.1 based on its suitability to the sample (experimental) variogram plot. Since the RF $Z(t_i)$ is standardized to zero mean and unit variance, expression (5.14) may be re-written as:

$$\rho_{ij} = Cov\{Z(t_i), Z(t_j)\} = C(\tau), \qquad \forall i = 1, \ldots, N \tag{5.19}$$

where $\tau = | t_i - t_j |$ is the temporal lag distance between time instants t_i and t_j. As a result, the simple kriging estimator assuming independence between data (5.16) may be reformulated as:

$$X(t_j) = \sum_{i=1}^{N} C(\tau) \cdot X(t_i), \qquad \forall j = 1, \ldots, J = 365 \tag{5.20}$$

where $X(\cdot)$ is the standardized RF, and $C(\tau)$ is obtained from Eq. (5.18). In other words, standardized unknown points $X(t_j)$ can be predicted by taking a linear combination of independent sample data $X(t_i)$, weighted with a covariance model $C(\tau)$.

In general, when data collection is well distributed and are dependent on each other, the RF $Z(t_i)$ can be assumed stationary with constant mean m and covariance function $C(\tau) = C(t, t + \tau)$, $\forall t, \tau \in T$. After rearranging Eq. (5.11), the SK estimator Z^*_{SK} becomes:

$$Z^*_{SK}(t_j) = \sum_{i=1}^{N} \lambda_i(t_j) Z(t_i) + \left[1 - \sum_{i=1}^{N} \lambda_i(t_j)\right] m \tag{5.21}$$

and the kriging weights $\lambda_i(t_j)$ can be obtained by the same SK system as expressed by Eq. (5.12). Therefore, the geostatistical concept of kriging is a generalized form of linear regression, obtained by considering the redundancy between sample data.

However, it is difficult and often premature to make any decision on stationarity, especially when lacking enough "hard" data. This drawback of SK method is overcome by ordinary kriging (OK), which is discussed next.

5.3.2 Ordinary Kriging

The robustness of simple kriging (SK) algorithm is enhanced by eliminating the prior knowledge requirement for the stationary mean m. This is especially true when the mean is deemed unreliable and therefore can be ignored in the formulation (5.21) by setting the second term to zero. This amounts to imposing the quantity $[1 - \sum_{i=1}^{N} \lambda_i(t_j)] = 0$, or equivalently, $\sum_{i=1}^{N} \lambda_i(t_j) = 1$, which is an additional constraint to the kriging algorithm. This improvement gives rise to another approach, termed ordinary kriging (OK) in which the mean m is assumed stationary but unknown. The OK estimator Z^*_{OK} at each time instant t_j is simply:

$$Z^*_{OK}(t_j) = \sum_{i=1}^{N} \lambda_i(t_j) Z(t_i) \tag{5.22}$$

which minimizes the error (estimation) variance σ^2_{OK} :

$$\sigma_{OK}^2(t_j) = C(0) - \sum_{i=1}^{N} \lambda_i(t_j)\, C(t_i - t_j) - \mu(t_j) \tag{5.23}$$

where $C(0) = \mathrm{Var}\{Z(t_i)\}$ denotes the variance of the RF $Z(t_i)$, inferred from N sample data points $z(t_i)$, $i = 1, \ldots, N$. The presence of Lagrange multiplier μ is crucial to satisfy the minimization of σ_{OK}^2 subject to the additional constraint $\sum_{i=1}^{N} \lambda_i(t_j) = 1$. The associated OK system is:

$$\begin{cases} \sum_{k=1}^{N} \lambda_k(t_j) C(t_i - t_k) + \mu(t_j) = C(t_i - t_j), \quad i = 1, \ldots, N & \text{(5.24a)} \\ \sum_{k=1}^{N} \lambda_k(t_j) = 1 & \text{(b)} \end{cases}$$

which is a system of $(N+1)$ equations to solve for $(N+1)$ unknowns $\lambda_k(t_j)$ and $\mu(t_j)$. The OK estimator Z_{OK}^* can be readily calculated as soon as the weights $\lambda_k(t_j)$ are found from the OK system (5.24). In essence, the pre-requisites of the OK algorithm are:

- $(N \times N)$ square matrix of covariances $[C(t_i - t_k)]$: $i, k = 1, \ldots, N$, of the sample data, and
- $(N \times 1)$ vector of data-to-unknown covariance $[C(t_i - t_j)]^{\mathrm{T}}$: $i = 1, \ldots, N$, where the superscript T denotes the transpose operation,

One important feature of ordinary kriging is that it implicitly re-estimates the locally varying mean m_{OK}^* at each new time instant t_j. This condition can be easily proven by taking the expected value of OK estimator Z_{OK}^*, i.e.,

$$E\{Z_{OK}^*(t_j)\} = E\{\sum_{i=1}^{N} \lambda_i(t_j)\, Z(t_i)\}$$

$$= \sum_{i=1}^{N} \lambda_i(t_j)\, E\{Z(t_i)\}$$

$$= m^*_{OK} \sum_{i=1}^{N} \lambda_i(t_j) = m^*_{OK}$$

since the constraint $\sum_{i=1}^{N} \lambda_i(t_j) = 1$ ensures the unbiasedness of the OK estimator Z^*_{OK}. In this sense, the OK algorithm is similar to that of the traditional SK (5.11) where the stationary $m(t_j)$ and locally averaged $m(t_i)$ means are substituted by m^*_{OK}, i.e.,

$$Z^*_{OK}(t_j) - m^*_{OK}(t_j) = \sum_{i=1}^{N} \lambda_{i,SK}(t_j)\left[Z(t_i) - m^*_{OK}(t_j)\right] \tag{5.25a}$$

After rearrangement, the OK estimator is simply re-written as:

$$Z^*_{OK}(t_j) = \sum_{i=1}^{N} \lambda_{i,SK}(t_j) Z(t_i) + \left[1 - \sum_{i=1}^{N} \lambda_{i,SK}(t_j)\right] m^*_{OK}(t_j) \tag{5.25b}$$

where the SK weights $\lambda_{i,\,SK}(t)$ are determined from system (5.12) and differ from those obtained in terms of the unbiased constraint $\sum_{i=1}^{N} \lambda_{OK}(t_j) = 1$ in the OK system (5.24). Therefore, ordinary kriging can be thought of as an algorithm corresponding to a nonstationary RF $Z(t_i)$ with locally varying mean m^*_{OK} but constant covariance $C(\tau)$, especially in the direction orthogonal to the moving trend. This robust feature of ordinary kriging makes it suitable for predicting average concentrations of ozone from historical data correlation.

However, just like simple kriging (SK), the ordinary kriging (OK) algorithm can also be improved by including secondary information, i.e., "soft" data, in the formulation. This addition, in turn, gives rise to another algorithm known as cokriging, which is discussed next.

5.3.3 Cokriging

In many cases, a variety of environmental data are measured at monitoring stations $\mathbf{u}_\delta$, $\delta = 1, \ldots, D$. Thus it is advantageous to include some of the auxiliary information ("soft data") in the algorithm in order to improve prediction (or estimation). For example, nitric oxide (NO) and volatile organic compounds (VOCs) are two of the many well-known chemicals that induce tropospheric ozone formation. These precursors $Z_\alpha(t_i)$, $\alpha = 1, \ldots, M$, $i = 1, \ldots, N$, to ozone can "add value" to prediction because of their direct influence in the photochemical reaction mechanisms (Section 2.2). The simple kriging (SK) paradigm can be extended to incorporate the relevant secondary information. The resulting simple cokriging (SCK) estimator Z^*_{SCK} is written as:

$$Z^*_{SCK}(t_j) - m_o(t_j) = \sum_{\alpha=0}^{M} \sum_{i=1}^{N_\alpha} \lambda_{\alpha i}(t_j)\,[Z_\alpha(t_i) - m_\alpha(t_i)] \tag{5.26}$$

where subscript $\alpha = 0$ refers to the primary variable (i.e., ozone); the stationary mean $m_o(t_j)$ corresponds to the RF $Z_o(t_j)$ defined at time instant t_j and must be either postulated *a priori* from historical records or estimated from the knowledge of similar phenomena; the variable-(α)-specific means $m_\alpha(t_i)$ are stationary averages of the respective RFs $Z_\alpha(t_i)$, $\alpha = 0, \ldots, M$. The kriging weights $\lambda_{\alpha i}(t_j)$ are solved via the following system of normal equations:

$$\sum_{\beta=0}^{M} \sum_{k=1}^{N_\beta} \lambda_{\beta k}(t_j)\, C_{\alpha\beta}(t_{\alpha i} - t_{\beta k}) = C_{\alpha o}(t_{\alpha i} - t_j), \tag{5.27}$$

$$\forall \alpha = 0, \ldots, M,\ i = 1, \ldots, N_\alpha$$

where:

- $[C_{\alpha\beta}(t_{\alpha i} - t_{\beta k})]$ is the $[(M + 1)\cdot(N)$ by $(M + 1)\cdot(N)]$ square matrix of auto and cross-covariances,

- $[C_{\alpha o}(t_{\alpha i} - t_{\alpha o})]^{\mathrm{T}}$ is the $[(M + 1)\cdot(N)$ by 1] vector of data-to-unknown auto and cross-covariances,

which can be used to solve for $(M + 1)\cdot(N)$ weights $\lambda_{\alpha i}(t_j)$, $\alpha = 0, \ldots, M$; $i = 1, \ldots, N_\alpha$, and thus the simple cokriging estimator Z^*_{SCK} at each time instant t_j.

However, the SCK algorithm suffers the same problem as SK does in term of the stationarity decision. It is difficult to accurately infer the stationary means $m_\alpha(t)$ because: (1) for estimation, the availability of "hard" data is often scarce due to missing values at the monitoring stations $\mathbf{u}_\delta$, and (2) the future values are, of course, unavailable for predictive purposes in which case, we have to stochastically determine a few sample values from the historical records. To alleviate this predicament, the stationary means $m_\alpha(t)$ are filtered out from cokriging expression (5.26). The resulting ordinary cokriging (COK) estimator Z^*_{COK} is given by:

$$Z^*_{COK}(t_j) = \sum_{\alpha=0}^{M}\sum_{i=1}^{N_\alpha} \lambda_{\alpha i}(t_j) Z_\alpha(t_i) \tag{5.28}$$

which should satisfy the following constraints in order to maintain unbiasedness:

- the weights related to the primary variables (denoted by subscript $\alpha = 0$) must add up to one, i.e., $\sum_{i=1}^{N_o} \lambda_{oi} = 1$, and
- for the secondary variables ($\alpha > 0$), the sum of the variable-specific weights must vanish, i.e. $\sum_{i=1}^{N_\alpha} \lambda_{\alpha i} = 0$, $\alpha = 1, \ldots, M$.

The kriging weights can be solved from the following COK system:

$$\sum_{\beta=0}^{M}\sum_{k=1}^{N_\beta} \lambda_{\beta k}(t_j) C_{\alpha\beta}(t_{\alpha i} - t_{\beta k}) + \mu_\alpha(t_j) = C_{\alpha o}(t_{\alpha i} - t_j) \tag{5.29}$$

$$\forall \alpha = 0, \ldots, M,\ i = 1, \ldots, N_\alpha$$

where $\mu_\alpha(t_j)$ are the Lagrange multipliers corresponding to specific secondary RFs $Z_\alpha(t)$, $\alpha = 0, \ldots, M$. Just like in the SK and OK algorithms, the covariance matrix needs to be positive definite in order to ensure a positive estimation variance σ_K^2. However, in the cokriging paradigm, the large $[(M + 1)\cdot(N)$ by $(M + 1)\cdot(N)]$ covariance matrix has to be positive definite. Such a condition for covariance model legitimacy is usually performed by a technique called the linear model of coregionalization (LMC) (Journel and Huijbregts, 1978) and is discussed next.

5.3.4 Linear Model of Coregionalization

In the case of simple or ordinary kriging, the positive-definiteness of the covariance matrix can be ensured through the use of a licit (legitimate) variogram model, thus ensuring positive variance of the predicted random function RF $Z(t_j)$. However, ensuring positive-definiteness of the covariance matrix in cokriging is more complicated since the combination of auto and cross-covariances (variograms) has to be modeled jointly.

The linear model of coregionalization (LMC) is one avenue to ensure positive-definiteness of the covariance matrix. It states that any positive linear combination of simpler, licit variogram structures will result in a positive-definite variogram model. As an example of a bivariate case, the auto-variogram $\gamma_Z(\tau)$ and $\gamma_Y(\tau)$, as well as the cross-variogram $\gamma_{ZY}(\tau)$ models must share common basic structures $\gamma_k(\tau/a_k)$, $k = 1, \ldots K$, and ranges a_k as the following:

$$\gamma_Z(\tau) = b_{00}^{(1)}\gamma_1\left(\frac{\tau}{a_1}\right) + b_{00}^{(2)}\gamma_2\left(\frac{\tau}{a_2}\right)$$

$$\gamma_Y(\tau) = b_{11}^{(1)}\gamma_1\left(\frac{\tau}{a_1}\right) + b_{11}^{(2)}\gamma_2\left(\frac{\tau}{a_2}\right)$$

$$\gamma_{ZY}(\tau) = b_{01}^{(1)}\gamma_1\left(\frac{\tau}{a_1}\right) + b_{01}^{(2)}\gamma_2\left(\frac{\tau}{a_2}\right) = \gamma_{YZ}(\tau)$$

where $b_{\alpha\beta}^{(k)}$: $\alpha, \beta = 0, 1$, are the sill contributions for specific variogram structures. Note that the basic structures are limited to only two for an illustrative purpose; they may assume more than two structures depending on the shapes of experimental variograms. The sill components $b_{\alpha\beta}^{(k)}$ act as positive weights for the linear combination and can be combined into $(M + 1)$ by $(M + 1)$ matrices, called the coregionalization matrices $\mathbf{B}^{(k)}$, and written as:

- For the first structure, i.e., $k = 1$: $\mathbf{B}^{(1)} = \begin{bmatrix} b_{00}^{(1)} & b_{01}^{(1)} \\ b_{10}^{(1)} & b_{11}^{(1)} \end{bmatrix}$, and

- For the second structure, i.e., $k = 2$: $\mathbf{B}^{(2)} = \begin{bmatrix} b_{00}^{(2)} & b_{01}^{(2)} \\ b_{10}^{(2)} & b_{11}^{(2)} \end{bmatrix}$

In order to ensure positive-definiteness of the variogram models, the determinants $| \mathbf{B}^{(k)} |$ must be greater than zero, i.e.,

- $b_{00}^{(1)} b_{11}^{(1)} - [b_{01}^{(1)}]^2 > 0$, and

- $b_{00}^{(2)} b_{11}^{(2)} - [b_{01}^{(2)}]^2 > 0$

This task is quite difficult, especially when modeling with more than two secondary variables $Z_\alpha(t_i)$, $\alpha = 3, \ldots, N$.

The following procedure is followed in the LMC modeling:

1. Model the auto-variograms utilizing the same structures (e.g., *Gauss*, *Exp*) with the largest possible identical ranges. Any structure not shared by all the auto-variogram models are neglected in the cross-variograms;

2. Calculate the determinants of all coregionalization matrices $\mathbf{B}^{(k)}$, $k = 1, \ldots, K$. If any of these determinants is negative, adjust the contributions $b_{\alpha\beta}^{(k)}$: $\alpha, \beta = 0, \ldots, M$, of the k basic variogram $\gamma_k(\tau)$ structures in order to satisfy the positive-definite conditions; and

3. Check the overall quality of auto and cross-variogram (or covariance) models by visually examining the plots. Modify the range and/or model types (Section 5.2) to obtain the best possible licit models. It is important to remember that all models must share the same number of structures, ranges and types. The fitted variograms may not exactly match the experimental variograms because the result from the LMC procedure is the best compromise for ensuring the positive-definiteness of the coregionalization matrices $\mathbf{B}^{(k)}$.

5.4 Results and Discussion

For determining the temporal ozone correlations, first the experimental variograms of four-year ozone data (standardized to zero mean and unit variance) were plotted. Here, GSLIB programs (Deutsch and Journel, 1998) called **gam** and **vmodel** (Appendices A.1 and A.2) are utilized for calculating the experimental and model values, respectively. The experimental variograms in Figure 5.2 exhibit a high nugget effect, as evidenced by the discontinuity near the origin. Hence prior to modeling the variograms, the source of the nugget effect must be investigated. Figure 5.3 shows the variogram computed using the hourly average ozone values, calculated at lags up to 7 days (168 hours). There is little or no nugget effect displayed by these higher resolution data. It can therefore be concluded that the high nugget effect seen in Figure 5.2 is due to the coarser resolution (daily averages) of the data used for variogram computation compared to the actual resolution of the ozone data in the form of hourly averages. For this reason, a variogram model (Eq. 5.30) based on the 1997 ozone data is developed assuming no nugget effect. This variogram model is also applied for the predictive mode, i.e., as a representation of the temporal variations in all subsequent years (1998-2000).

$$\gamma(\tau) = 0.50 \cdot Exp\left(\frac{\tau}{5}\right) + 0.50 \cdot Gauss\left(\frac{\tau}{100}\right), \quad \forall \tau \in [1, 365] \tag{5.30}$$

In addition, a variogram model reflecting the periodicity of the observed long-term ozone behavior is also developed. This hole-effect model shown at the bottom of Figure 5.2 utilizes a cosine function:

$$\gamma(\tau) = 1 - 0.5 \cdot \cos\left(\frac{2\pi\tau}{365}\right), \quad \forall \tau \in [0, 365] \tag{5.31}$$

and will be implemented later in the stochastic simulation (Chapter 6). Note that a nugget effect of 0.5 should be added to the covariance function (5.18) in order to ensure positive-definiteness.

As a preliminary study, "kriging assuming independence between data" (Eq. 5.20) is implemented using rigid data sampling. Here twelve evenly spaced data, sampled from the daily average values are selected at every 30th Julian day, i.e., on the 30th, 60th, ..., and 360th, and correlated via temporal correlation (5.30). Each set of the resulting outputs is obtained by taking a linear combination of the twelve rigid data and then superimposed on the 30-day moving averages (30dMA) of the actual ozone values for comparison (Figure 5.4). Several small spikes are clearly visible at the data locations but the overall outputs reflect smooth trends at all other locations. In general, the resulting outputs mimic the actual 30dMA ozone behavior for all four years, signifying that this type of kriging can only capture the average annual trends of the ozone phenomena without honoring the sample data at their respective locations.

Further analysis is performed using ten sets of twelve randomly selected data at the interval of [25, 30] Julian day of the month. Figure 5.5 shows that the 30dMA of actual ozone values are bounded within the minimum and maximum (in a least-square sense) 30dMA of the results, implying that sampling uncertainties may greatly influence the accuracy of the ozone prediction. To verify this, the correlation coefficients $\rho_{\alpha o}$, $\alpha =$ 1 ,..., 10, between the 30dMA of the actual ozone values and the resulting outputs are plotted at the bottom graphs. The large variations in the $\rho_{\alpha o}$ values [0.45, 0.89] and [0.57, 0.96] for the 1997 and 1998 cases, respectively, confirm the above statement with regards to sampling uncertainties. Conversely, high $\rho_{\alpha o}$ values [0.85, 0.94] and [0.94, 0.97] for the corres-ponding 1999 and 2000 cases signify that accurate prediction may be obtained if the randomly sampled data are representative of the annual ozone trend when they are properly placed at particular locations.

The above procedure is updated via ordinary kriging (OK) to account for data interactions. It should be noted that simple kriging (SK) can also be implemented but the stationary mean must be known previously. The OK algorithm removes this requirement and proceeds without prior knowledge of the stationarity mean by constraining the kriging weights to one (Section 5.3.2). To determine the performance of the OK algorithm (GSLIB `kt3d` parameter file: Appendix A.3), the variogram inferred based on 1997 data was used to estimate the 1997 temporal phenomena. This process serves as a validation to the suitability of the variogram model, which is subsequently used for predicting the ozone trends in 1998-2000. The kriging procedure is carried out for two different cases using the following data sampling procedure:

1. twelve evenly spaced data, selected at every 30th Julian day throughout the year, and
2. ten sets of twelve data, randomly sampled between the 25th and 30th day of the month.

In the first case, the twelve data used for inferring the highly fluctuating daily values are sampled at higher resolution (daily average) and therefore may consist of extreme high and/or low values. Consequently, this will result in the presence of about twelve spikes as depicted in Figure 5.6. These spikes are caused by the extreme short-range structure in the variogram model (5.30) and the exactitude of kriging at the data locations. Another important observation is the smooth variability of the kriged outputs at other locations. This phenomenon is due to the second-order stationarity assumption of kriging, in which only the first two moments are reproduced. For this reason, it is more reasonable to compare the kriged profiles against the 30dMA of the ozone values as a measure for evaluating the annual trends. The general profiles show that the temporal highs and lows are not accurately modeled, especially in the period between the 90th and 150th Julian day of 1999. Since the datum at the 120th Julian day is low, kriging will honor this datum at its location and thus "bring down" the neighboring values.

To circumvent the above data-sampling problem, ten sets of twelve data points are randomly selected between the 25th and 30th day of the month for all four years. As

mentioned previously, this procedure will permit for evaluating the influence of high variability in the daily ozone values. The results shown in Figure 5.7 exhibit a marked improvement in the accuracy of prediction. Here, the 30-day moving average values of the raw data are enclosed within those of the minimum and maximum (in a least-square sense) kriged outputs corresponding to different sets of the conditioning data (top graphs). The correlation coefficients $\rho_{\alpha o}$, α = 1, ..., 10, between the 30dMA of kriged profiles and those of the actual values are calculated for the ten sets of random data. The resultant distributions of the $\rho_{\alpha o}$ values shown at the bottom of Figure 5.7 are generally greater than 0.8 about 90% of all cases. The correlation coefficients in 2000 are the highest and exhibit the least variability, i.e., $\rho_{\alpha o} \in [0.96, 0.98]$ implying excellent kriging performance. In contrast, the kriging results for 1997 demonstrate the most variability where $\rho_{\alpha o} \in [0.25, 0.94]$.

Next, the ordinary kriging (OK) algorithm is extended to include the influence of secondary information via ordinary cokriging (COK) (GSLIB **cokb3d** parameter file: Appendix A.4). In this case, the temporal correlations between covariates (e.g., total hydrocarbon THC and nitric oxide NO) and ozone are captured using cross-variograms, which are calculated at incremental temporal lag τ and given as:

$$\gamma_{ZY}(\tau) = \frac{1}{2N(\tau)} \sum_{i=1}^{N(\tau)} \{z(t_i) - z(t_i + \tau)\} \cdot \{y(t_i) - y(t_i + \tau)\} \tag{5.32}$$

where $N(\tau)$ is the total number of pairs at lag τ apart. Note that variogram is a two-point statistic and thus only two attributes, i.e., RFs $Z(t_i)$ and $Y(t_i)$, can be correlated at one time. Consequently, the multipoint dependencies between ozone, THC and NO variables taken jointly (or simultaneously) cannot be modeled using a single cross-variogram. The influences of THC as well as NO on ozone, however, can be modeled using three sets of joint cross-variograms, i.e., between (1) ozone and THC, (2) ozone and NO, and finally (3) THC and NO in order to account for redundancies between these two covariates. Another important point to note is that in the case of a systematic trend for RF $Z(t)$, the values of the tail may sometimes be greater than the head, i.e., $z(t_i) > z(t_i + \tau)$, and at the

same time, the reverse is true for RF $Y(t)$, i.e., $y(t_i) < y(t_i + \tau)$. This will result in a negative cross-variogram between these two RFs. To circumvent such cases, the cross-variogram is calculated by initially taking a negative transformation of the covariate values.

Based on the sample cross-variogram plots (Figure 5.8), two covariates, i.e., total hydrocarbon (THC) and nitric oxide (NO) are selected because they show distinct trends similar to that exhibited by the ozone auto-variogram (Figures 5.2 and 5.9). Moreover, these two covariates have relatively high correlation coefficients (Table 4.2) with ozone over the entire four-year study period and are known to be the chemical precursors to ozone formation (Chapter 2). The other chemical covariates such as nitrogen dioxide (NO2), carbon monoxide (CO), and dusk and smoke (COH) also exhibit the above cross-variogram qualities and may be used as substitutes for THC and NO as the secondary information.

The sample auto-variogram of ozone, and those of the negatively transformed THC and NO exhibit discontinuity near origin (Figure 5.9). The cross-variograms plotted in the same figure also demonstrate similar behavior but their nugget effects are slightly lower. However, it was explained previously that this nugget phenomenon is only due to calculating variograms using the coarser resolution of the daily average values as opposed to the hourly values. To rectify this circumstance, all variograms are modeled with short-range structure (5 days) with appropriate sill contribution.

The models for cross-variograms are developed following the LMC procedure as outlined in Section 5.3.4. Note that the upper limit of the sill values $C(0) = Var\{Z(t_i)\}$ of all auto-variogram models are always one; however, for the cross-variogram models, the maximum sills $C(0) = Cov\{Z_\alpha(t_i), Z_\beta(t_i)\}$ (at lag $\tau = 0$) are limited to the correlation coefficients $\rho_{\alpha\beta}$, $\alpha, \beta = 0, \ldots, 2$, $\alpha \neq \beta$, between the respective variables. To ensure positive-definiteness of the coregionalization matrices $\mathbf{B}^{(k)}$, $k = 1, \ldots, K$, the actual sills of the cross-variograms are adjusted until all determinants $|\mathbf{B}^{(k)}|$ are positive. The general isotropic auto and cross-variogram model is given below and the values of the sill contributions ($c_{\alpha\beta}$ and $d_{\alpha\beta}$) are summarized in Table 5.2:

$$\gamma_{\alpha\beta}(\tau) = c_{\alpha\beta} \cdot Exp\left(\frac{\tau}{5}\right) + d_{\alpha\beta} \cdot Gauss\left(\frac{\tau}{120}\right) \tag{5.33a}$$

where α and β are the variable indices for the cross-variogram models. Note that the range for the Gauss model structure is 120 as opposed to 100 days in the auto-variogram model of ozone (1997). This compromise is necessary in order to better fit the rest of the sample auto and cross-variograms.

Table 5.2
Sill contributions for the auto and cross-variogram models to be used in the linear model of coregionalization (LMC).

Variable		**Sill Contributions**	
Indices	**Names**	$c_{\alpha\beta}$	$d_{\alpha\beta}$
0	O3	0.57	0.43
1	THC	0.46	0.54
2	NO	0.50	0.50
01	O3-THC	0.24	0.44
12	NO-THC	0.39	0.50
02	O3-NO	0.26	0.41

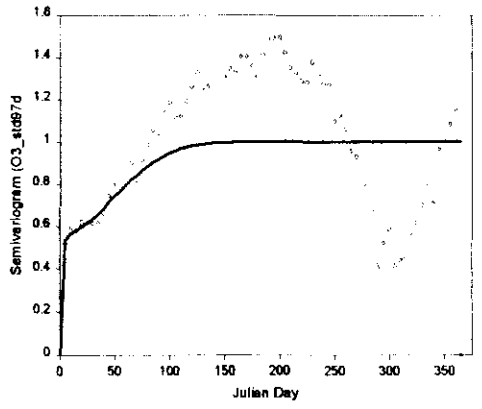
In matrix form, the coregionalized variogram model above is written as:

$$\gamma(\tau) = \begin{bmatrix} 0.57 & 0.24 & 0.26 \\ 0.24 & 0.46 & 0.39 \\ 0.26 & 0.39 & 0.50 \end{bmatrix} \cdot Exp\left(\frac{\tau}{5}\right) + \begin{bmatrix} 0.43 & 0.44 & 0.41 \\ 0.44 & 0.54 & 0.50 \\ 0.41 & 0.50 & 0.50 \end{bmatrix} \cdot Gauss\left(\frac{\tau}{120}\right) \tag{5.33b}$$

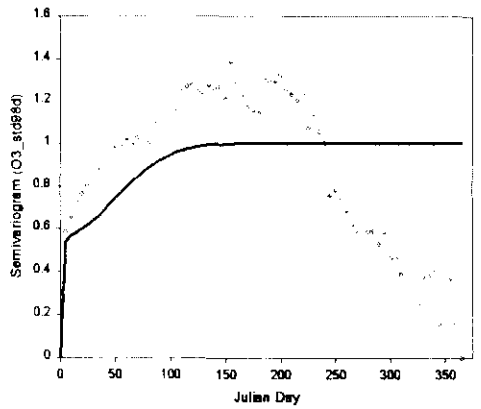
It can be easily verified that the determinants $|\ \mathbf{B}^{(k)}\ |$ corresponding to the above matrix systems are indeed positive, thus ensuring that the estimation procedure gives rise to positive variance. The resultant variogram models for the auto and cross-variograms are shown in Figure 5.9 to assess the goodness of fit.

Before discussing the kriging results, it is useful to look at the temporal variations of the covariates. Figure 5.10 shows that during the four-year study period, the THC and NO do not necessarily peak at the same time the ozone does. For example, in 1999, NO and ozone peak at about the same months but THC reaches maximum at a later time of the year. Hence depending on the weights attributed to THC during cokriging procedure, the predicted ozone profile may not emulate the actual trend due to the influence of THC. This observation is confirmed by ordinary cokriging (COK) results for 1999 as shown in Figure 5.11. In general, the integration of covariates in kriging paradigm improves the accuracy of ozone prediction, at least for the years 1997, 1998 and 2000. The mean temporal variation of the predicted profiles is brought up, resulting in the better emulation of the peaks in the respective years. The results also point to the adequacy of the auto and cross-variogram models inferred from the 1997 data for predicting the ozone trends in the subsequent years.

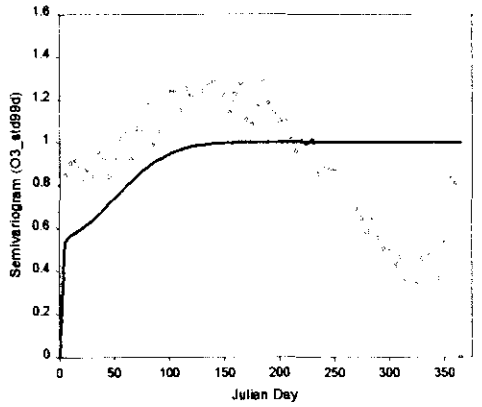
However, similar to SK and OK or any other regression approaches, cokriging can only yield smooth estimates. Recall that in order to solve for the weights, kriging system only ensures the reproduction of the covariance model between data and the estimation points (unknowns). However, covariance between the unknowns themselves does not identify the covariance model, resulting in the poor emulation of the highly fluctuating values observed in the daily average ozone data. Furthermore, it will be proven in the next chapter that the kriged profiles exhibit less variance and thus generating smooth temporal trends. This is readily evidenced from the results presented in this chapter. Apart from these shortcomings, the exactitude of kriging at the data locations is an asset towards stochastic simulation, which is discussed next.



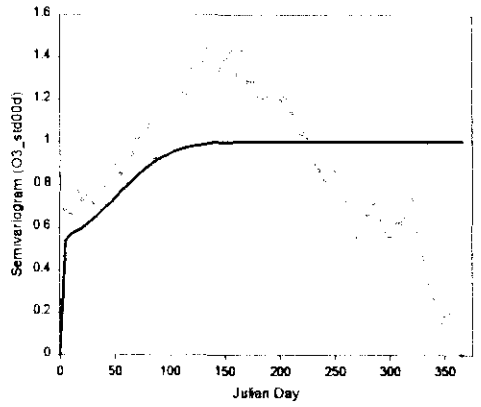
(a) 1997



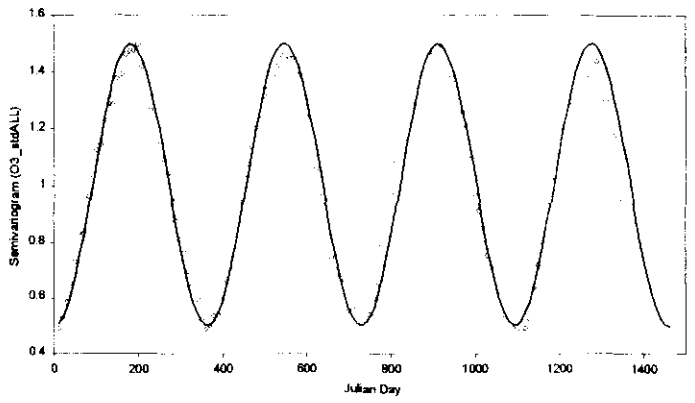
(b) 1998



(c) 1999

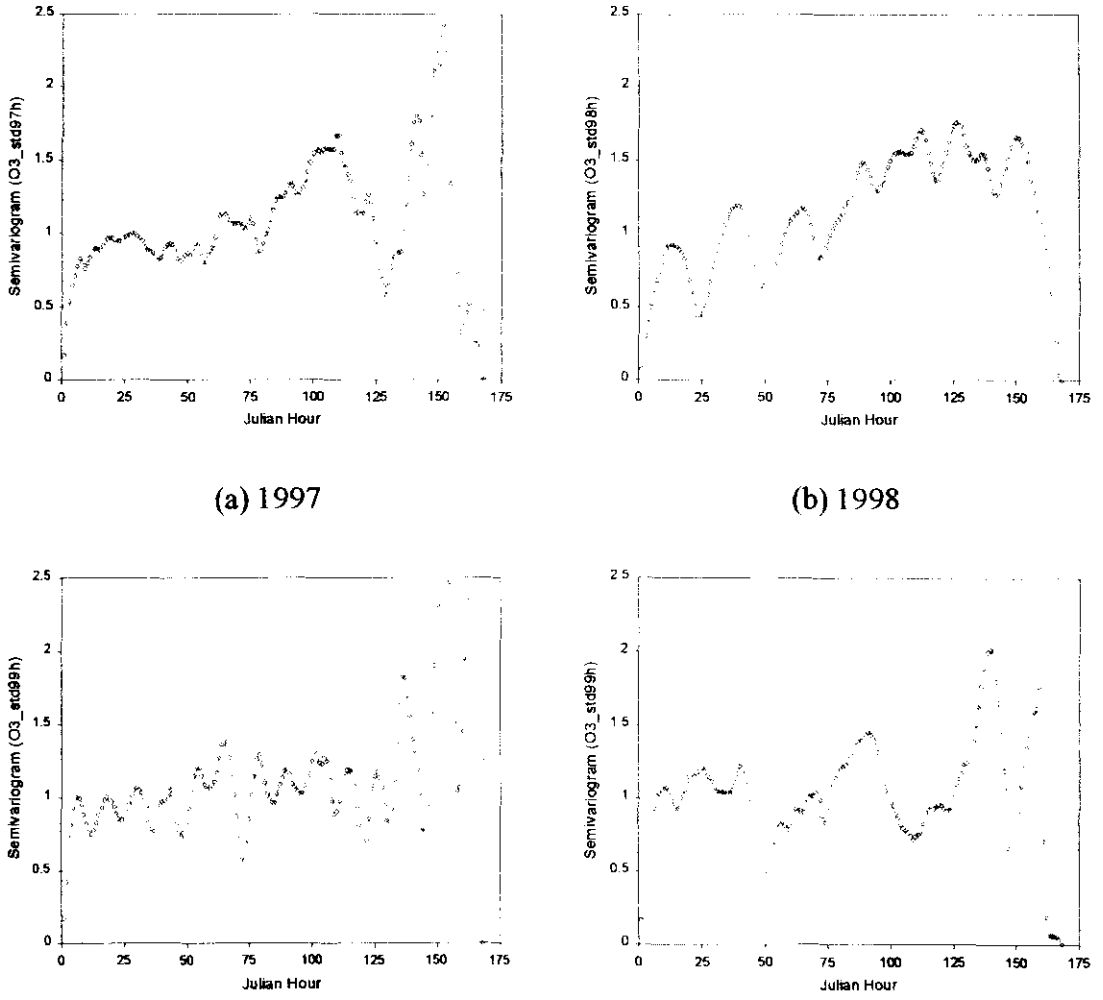


(d) 2000



(e) All four years (1997-2000)

Figure 5.2
Semivariograms of the daily average standardized ozone data (open diamonds with thin gray lines). The variogram models (thick blue lines) based on the sample variogram for 1997 are shown superimposed on the sample variograms for 1997-2000 in the case of (a) the two-structure model $\gamma(\tau) = 0.50 \cdot Exp(\tau/5) + 0.50 \cdot Gauss(\tau/100)$. The periodicity of variogram behavior over four-year period is better captured via (e) the hole-effect model $\gamma(\tau) = 1 - 0.50 \cdot \cos(2\pi\tau/365)$.



(a) 1997 (b) 1998

(c) 1999 (d) 2000

Figure 5.3
Semivariograms of hourly average standardized ozone data, calculated up to a maximum lag of seven Julian days (168 hours) of the year. The higher resolution of the ozone data results in the elimination of the nugget effect observed previously for the daily average standardized data.

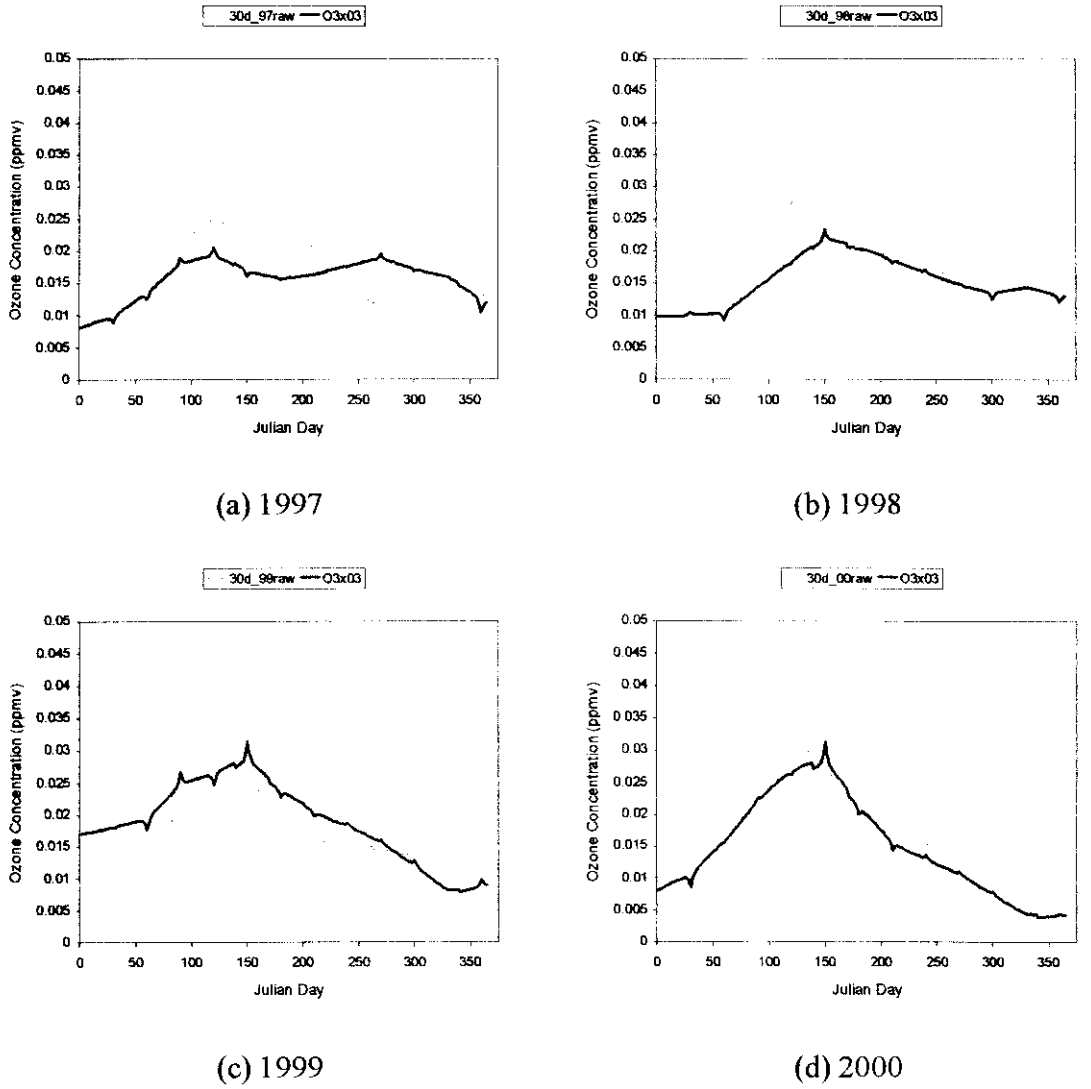


Figure 5.4
Linear regression results (thick blue line) based on one of the ten sets of twelve randomly selected data between the 25th and 30th Julian day of the month are superimposed on the actual 30-day moving average values of the raw data (thin gray line) in the respective year. Sample data and the unknowns are correlated using a variogram model $\gamma(\tau) = 0.50 \cdot Exp(\tau/5) + 0.50 \cdot Gauss(\tau/100)$. This type of linear regression is also known as "kriging assuming independence between data."

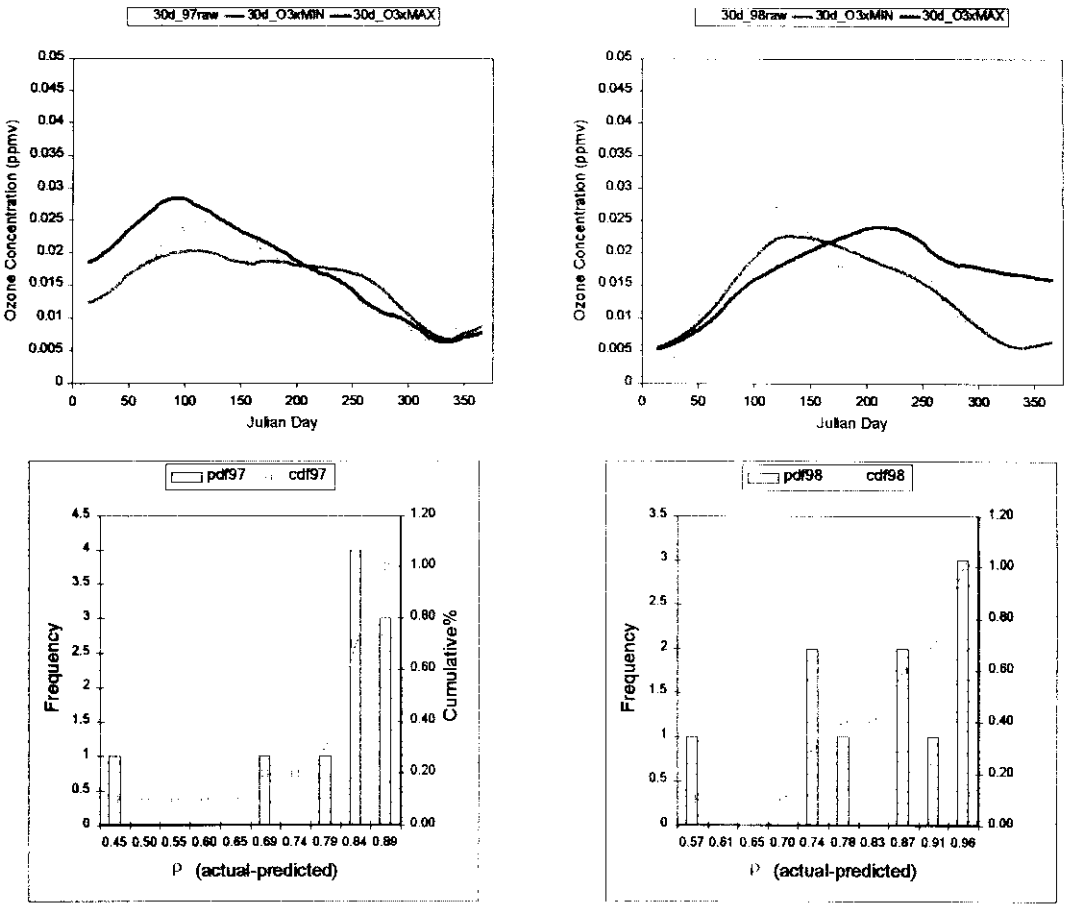


Figure 5.5 (i)

Linear regression results using ten sets of twelve randomly selected data between the 25th and 30th Julian day of the month for 1997 [**LEFT**] and 1998 [**RIGHT**]. In a least-square sense, the predicted minimum (green) and maximum (red) annual trends, i.e., the 30-day moving averages (30dMA) of the regression profiles, are superimposed on those of raw data (gray) [**top**]. The correlation coefficients between the regression and actual 0dMA were calculated for all ten cases [**bottom**].

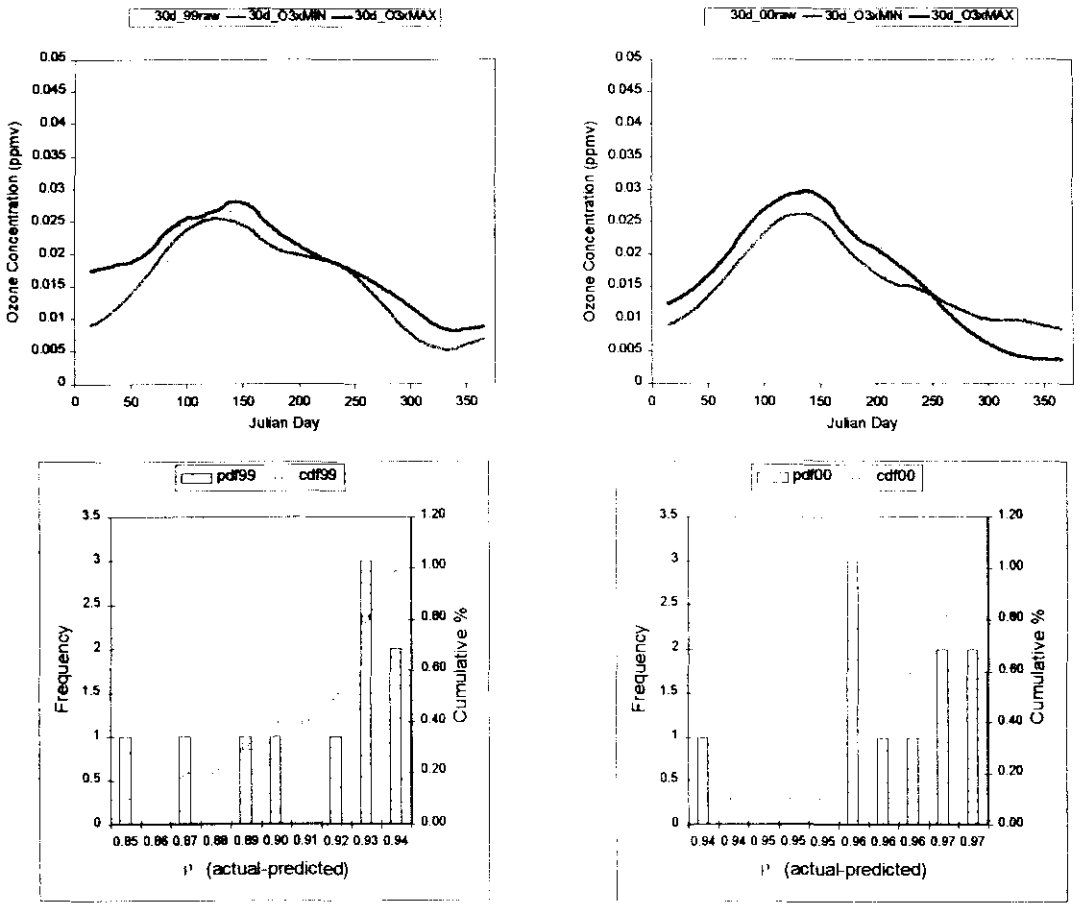


Figure 5.5 (ii)

Linear regression results using ten sets of twelve randomly selected data between the 25th and 30th Julian day of the month for 1999 [**LEFT**] and 2000 [**RIGHT**]. In a least-square sense, the predicted minimum (green) and maximum (red) annual trends, i.e., the 30-day moving averages (30dMA) of the regression profiles, are superimposed on those of raw data (gray) [**top**]. The correlation coefficients between the regression and actual 30dMA were calculated for all ten cases [**bottom**].

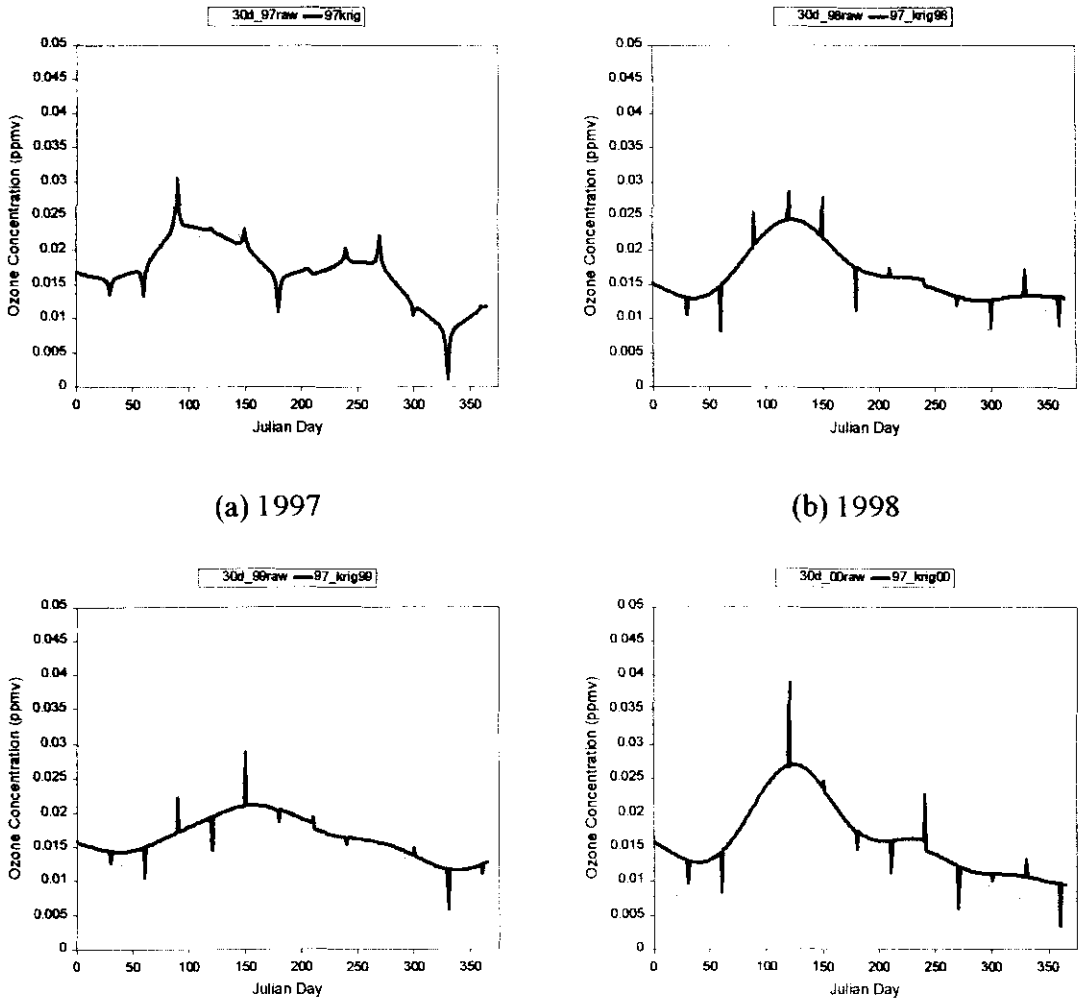


Figure 5.6
Ordinary kriging results (thick blue line) based on twelve evenly spaced data points selected at every 30th Julian day are superimposed on the measured 30-day moving averages of the raw data (thin gray line) for the respective year.

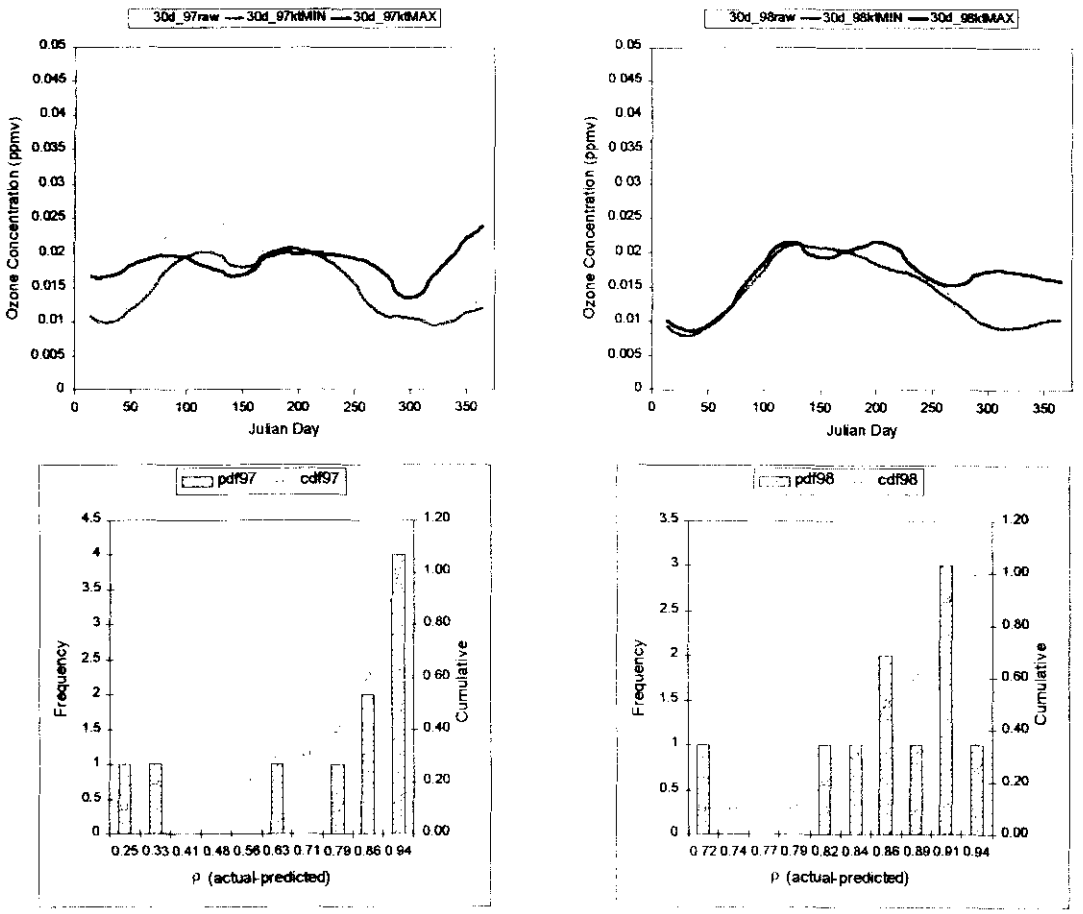


Figure 5.7 (i)

Ordinary kriging results using ten sets of twelve randomly selected data between the 25th and 30th Julian day of the month for 1997 [LEFT] and 1998 [RIGHT]. In a least-square sense, the predicted minimum (green) and maximum (red) annual trends, i.e., the 30-day moving averages (30dMA) of the kriged profiles, are superimposed on those of raw data (gray) [**top**]. The correlation coefficients between the kriged and actual 30dMA were calculated for all ten cases [**bottom**].

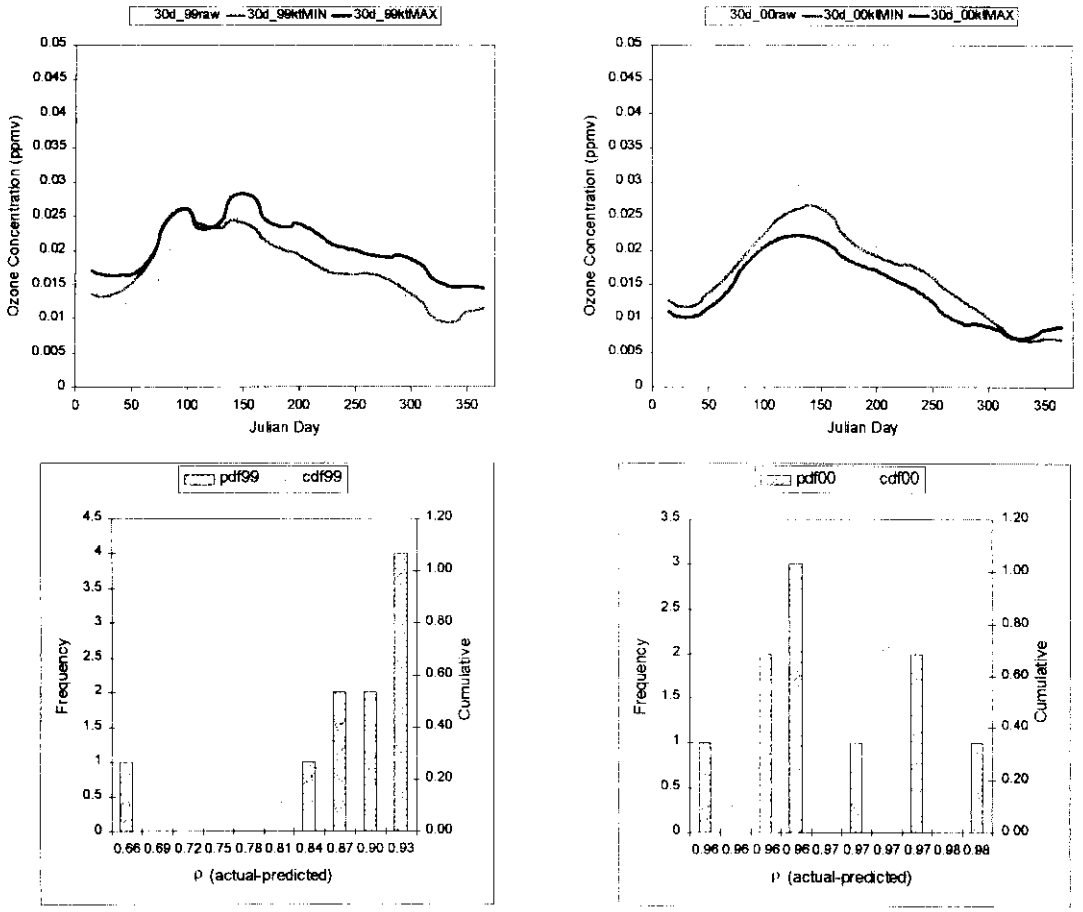


Figure 5.7 (ii)

Ordinary kriging results using ten sets of twelve randomly selected data between the 25th and 30th Julian day of the month for 1999 [**LEFT**] and 2000 [**RIGHT**]. In a least-square sense, the predicted minimum (green) and maximum (red) annual trends, i.e., the 30-day moving averages (30dMA) of the kriged profiles, are superimposed on those of raw data (gray) [**top**]. The correlation coefficients between the kriged and actual 30dMA were calculated for all ten cases [**bottom**].

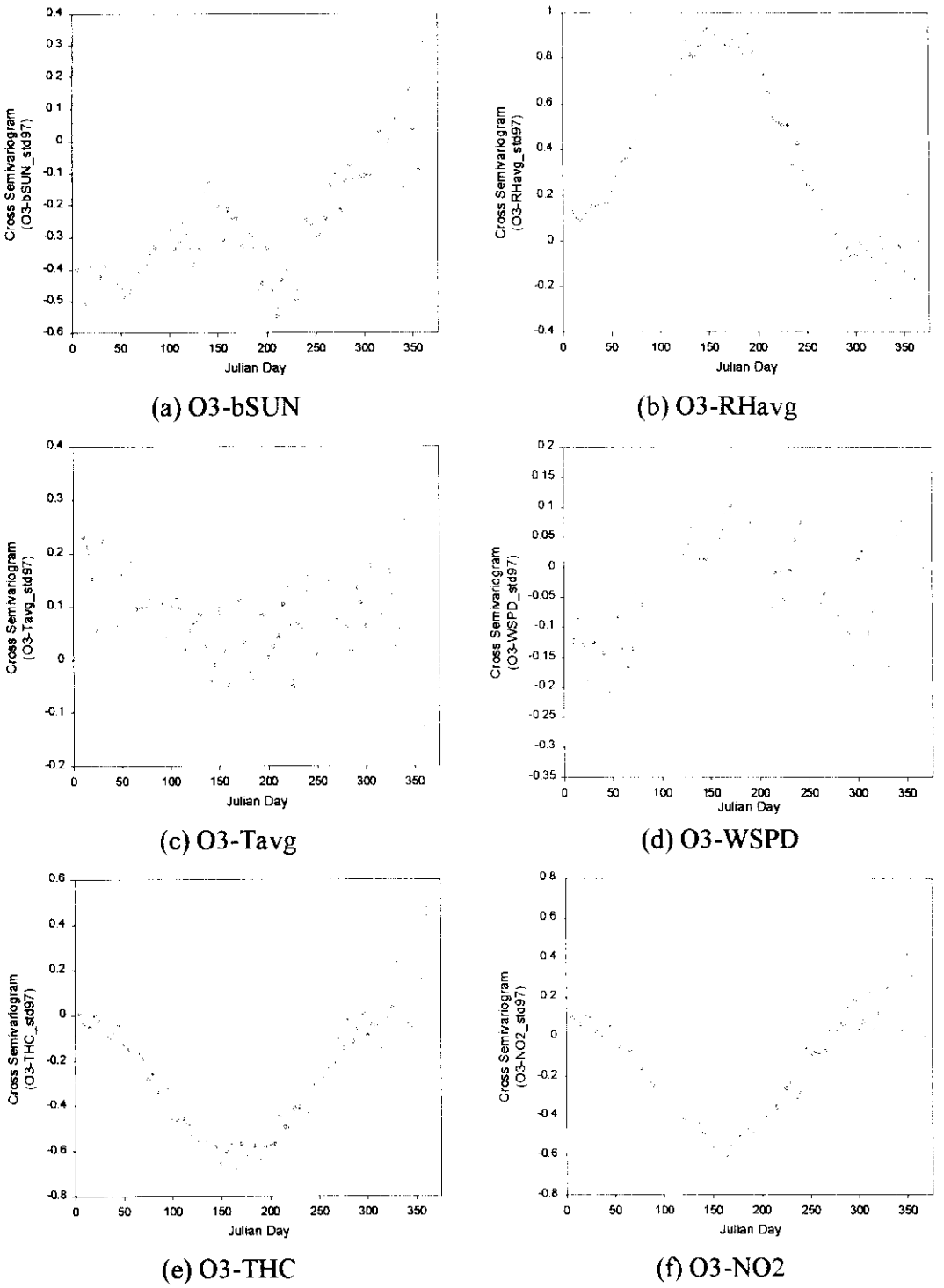
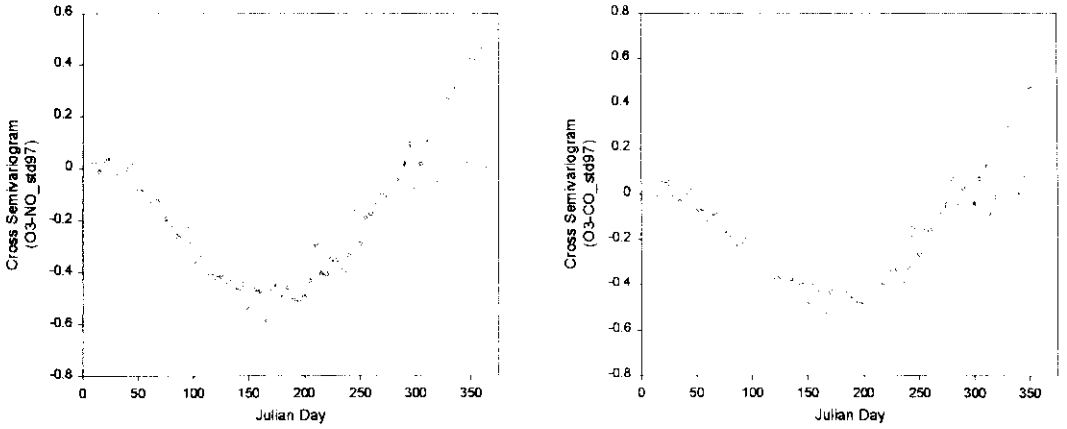
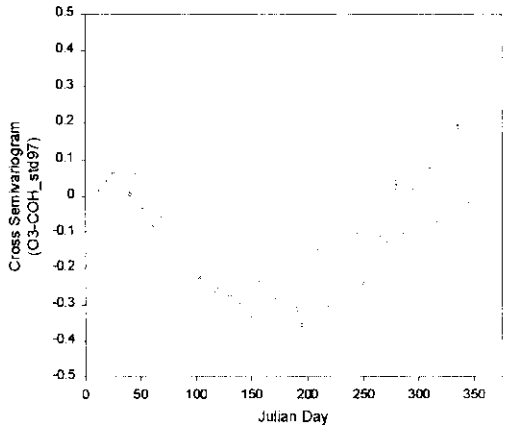


Figure 5.8 (i)
Cross semivariograms of ozone and the meteorological/chemical variables based on the standardized (zero mean and unit variance) 1997 data.



(g) O3-NO (h) O3-CO



(i) O3-COH

Figure 5.8 (ii)
Cross semivariograms between ozone and the meteorological/chemical variables of the standardized (at zero mean and unit variance) 1997 data.

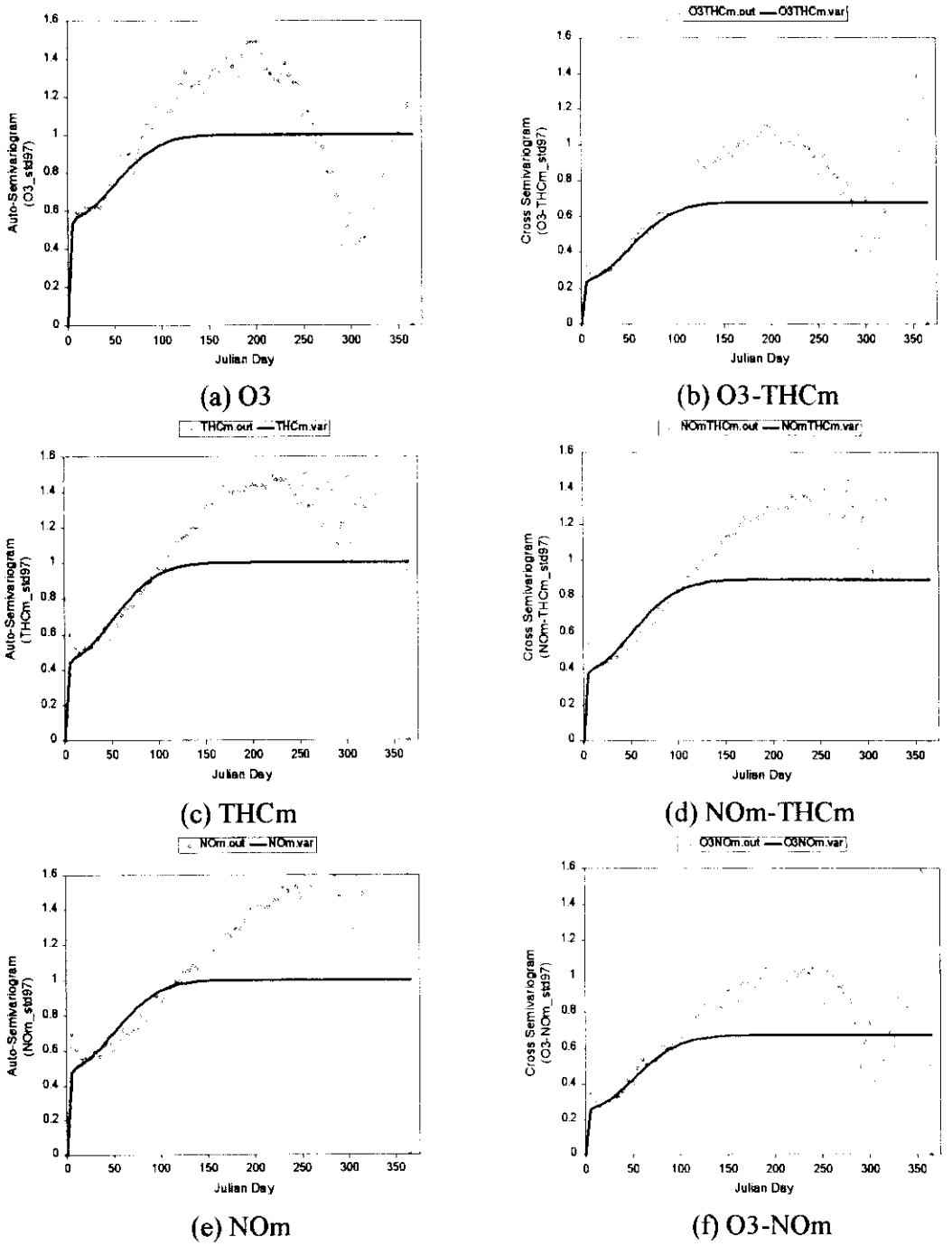


Figure 5.9
Experimental auto and cross-semivariograms (open diamonds with a solid gray line) and the model fit (thick blue line). The model sills for auto-variograms [**LEFT**] are always one but those for cross-variograms [**RIGHT**] are adjusted to ensure positive-definiteness of the coregionalization (LMC) matrices, and subject to a maximum dictated by the correlation coefficients between the primary and secondary variables.

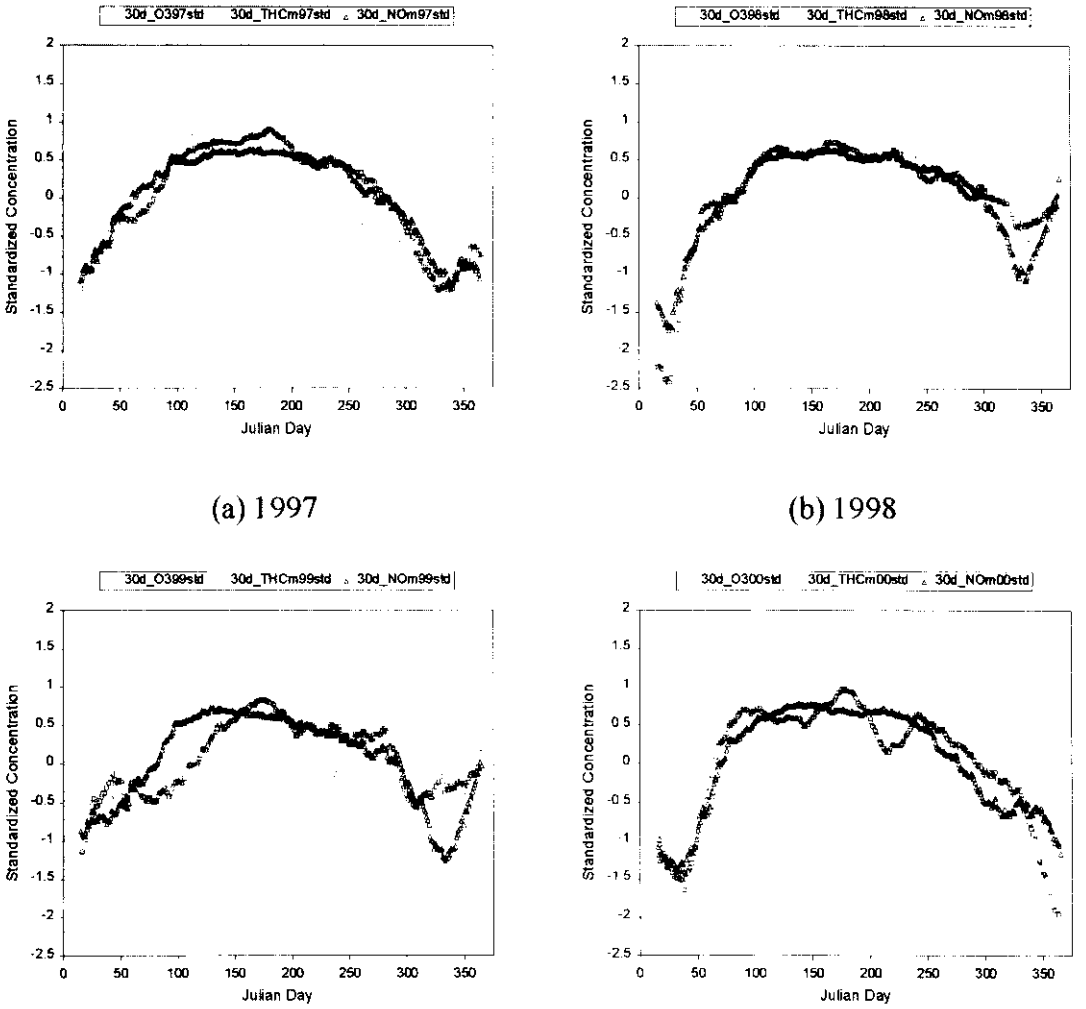


Figure 5.10
The standardized (at zero mean and unit variance) 30-day moving average values of ozone (thin gray line) and its covariates, i.e., total hydrocarbons THC (open pink box) and nitric oxide NO (open blue triangle) are plotted on the same graph for comparing the annual trends of all variables.

CHAPTER 6

STOCHASTIC SIMULATION

Stochastic simulation differs from process simulation in the sense that the latter incorporates physicochemical models, which are adapted from the previously published studies. Process simulation packages such as the Urban Airshed Model (UAM: SAI, 1999, Appendix B), have received wide acceptance for regulatory purposes due to the better understanding of physicochemical mechanisms underlying the process. However, tropospheric ozone phenomena involve many factors, some of which are still not fully understood and therefore ignored in the physicochemical models. An alternate approach is to consider the temporal variability in ozone values as realizations of a multivariate probability distribution underlying a stochastic random functions RF $Z(t)$ (e.g., Deutsch and Journel, 1998). The kriged results discussed earlier can be interpreted as the mean of the RF characterizing ozone concentrations at each time instant. In stochastic simulation, one way to proceed is to sample the entire multivariate distribution following a sequential procedure as outlined in the succeeding sections.

6.1 Theoretical Background

As discussed in Chapter 5, kriging is an excellent tool for honoring sample data and estimating the trends of ozone phenomena. However, these are accomplished at the expense of reduced variance, as previously substantiated by the smooth kriged profiles and irreproducibility of the correct patterns of temporal variations. Therefore, stochastic simulation is implemented in order to rectify the "smoothness effect" of kriging. In this sense, kriging serves as the first step towards stochastic simulation, which is a procedure of generating a set of L alternative realizations $Z^{(l)}(t)$, $l = 1, \ldots, L$, $\forall t \in T$ of a temporal

(and/or spatial) process. These realizations are basically sampled from the multivariate probability distribution underlying a random function RF $Z(t)$.

The reduction in variance of the kriged estimator $Z^*(t)$ can be corrected by adding a residual error component $R(t)$ in such a way that:

$$Z_s(t) = Z^*(t) + R(t) \tag{6.1}$$

where subscript s indicates the simulated value of RF $Z(t)$. To verify this, consider the estimation of a standardized RF $Z(t)$ (zero mean and unit variance). The kriged estimator $Z^*(t)$ is given by a linear combination of RFs Z_α, $\alpha = 1, \ldots, M$, i.e.,

$$Z^*(t) = \sum_\alpha \lambda_\alpha Z_\alpha$$

Thus its variance can be written as:

$$\begin{aligned} Var\{Z^*(t)\} &= E\{[\sum_\alpha \lambda_\alpha Z_\alpha]^2\} \\ &= E\{\sum_\alpha \sum_\beta \lambda_\alpha \lambda_\beta Z_\alpha Z_\beta\} \\ &= \sum_\alpha \sum_\beta \lambda_\alpha \lambda_\beta \cdot E\{Z_\alpha Z_\beta\} \\ &= \sum_\alpha \sum_\beta \lambda_\alpha \lambda_\beta \cdot C_{\alpha\beta} \\ &= \sum_\alpha \lambda_\alpha \cdot \sum_\beta \lambda_\beta C_{\alpha\beta} \\ &= \sum_\alpha \lambda_\alpha C_{\alpha o} \end{aligned}$$

where subscript 0 refers to the estimation point (unknown), and the last step is obtained by employing the kriging system of normal equations. However, the estimation variance is given as $\sigma_K^2 = C(0) - \sum \lambda_\alpha C_{\alpha o}$, or after rearranging and substituting for the target variance $C(0) = 1$, the result is $\sum \lambda_\alpha C_{\alpha o} = 1 - \sigma_K^2$. This implies that the variance of the

kriged profiles is less than the target variance by a quantity of σ_K^2. To resolve this problem, we add the residual $R(t)$ to the kriged estimator $Z^*(t)$ and calculate the resultant variance, i.e.,

$$Var\{Z(t)\} = Var\{Z^*(t) + R(t)\}$$

$$= Var(Z^*(t)\} + Var\{R(t)\} + 2Cov\{Z^*(t), R(t)\}$$

If the residual $R(t)$ is orthogonal (independent) to $Z^*(t)$, then $Cov\{Z^*(t), R(t)\} = 0$. In turn, the new variance of the RF $Z(t)$ is:

$$Var\{Z(t)\} = \text{Var}(Z^*(t)\} + Var\{R(t)\}$$

$$= 1 - \sigma_K^2 + Var\{R(t)\}$$

entailing that the RF $R(t)$ should have a variance of σ_K^2 in order to restore the correct variance of RF $Z(t)$.

In other words, to uphold unbiasedness, and also to regain the correct variance of the simulated values RF $Z_s(t)$, the residual $R(t)$ must satisfy the following criteria:

1. its expected value (mean) must vanish, i.e., $E\{R(t)\} = 0$,
2. its variance $Var(R(t)\} = \sigma_K^2(t)$, and
3. $R(t)$ must be orthogonal to the kriged estimate $Z^*(t)$ so that $Cov\{Z^*(t), R(t)\} = 0$.

If the expected value $E\{R(t)\} = 0$, expression (6.1) reduces to $E\{Z_s(t)\} = E\{Z^*(t)\}$, which by construction is exact at the data location due to exactitude of kriging estimator $Z^*(t)$. The last two conditions ensure that the RF $Z_s(t)$ has the correct variance to overcome the "smoothness effect" of kriging algorithm. The remaining task is to ensure that the simulated values $Z_s(t_i)$, $\forall t_i \in T$ reproduce the correct temporal variation, i.e., variogram model. Recall that kriging identifies the covariance (or variogram) between data and the estimation points. However, the drawback with kriging is that the covariance between two estimated points does not identify the model covariance, as shown in the following:

$$E\{Z^*(t_i)Z^*(t_j)\} = E\{\sum_\alpha \lambda_\alpha(t_i)Z_\alpha \cdot \sum_\beta \lambda_\beta(t_j)Z_\beta\}$$

$$= E\{\sum_\alpha \sum_\beta \lambda_\alpha(t_i)\lambda_\beta(t_j)Z_\alpha Z_\beta\}$$

$$= \sum_\alpha \lambda_\alpha(t_i) \sum_\beta \lambda_\beta(t_j) E\{Z_\alpha Z_\beta\}$$

Again, assuming RF $Z(t)$ to be standardized at zero mean and unit variance, the expected values $E\{Z_\alpha Z_\beta\} = C_{\alpha\beta}$. Hence, the covariance between two kriged estimators is:

$$Cov\{Z^*(t_i), Z^*(t_j)\} = E\{Z^*(t_i)Z^*(t_j)\}$$

$$= \sum_\alpha \lambda_\alpha(t_i) \sum_\beta \lambda_\beta(t_j) C_{\alpha\beta}$$

$$= \sum_\alpha \lambda_\alpha(t_i) C(t_i - t_j)$$

which is different from the target model covariance $C(t_i - t_j)$.

To rectify this deficiency, consider updating the kriged data set to include the values $Z_s(t)$, which are simulated prior to the current time instant t_j. Then,

$$Cov\{Z_s(t_i), Z_s(t_j)\} = E\{Z_s(t_i)\cdot[Z^*(t_j) + R(t_j)]\}$$

$$= E\{Z_s(t_i)\cdot Z^*(t_j)\} + E\{Z_s(t_i)\cdot R(t_j)\}$$

$$= E\{Z_s(t_i)\cdot Z^*(t_j)\} + E\{[Z^*(t_i) + R(t_i)]\cdot R(t_j)\}$$

$$= E\{Z_s(t_i)\cdot Z^*(t_j)\} + E\{Z^*(t_i)\cdot R(t_j)\} + E\{R(t_i)\cdot R(t_j)\}$$

Recall that the kriged values $Z^*(t)$ are orthogonal to the residual $R(t)$ and the expected value of the residual is zero; hence the above expression reduces to:

$$Cov\{Z_s(t_i), Z_s(t_j)\} = E\{Z_s(t_i)\cdot Z^*(t_j)\}$$

Since the current data set includes the previously simulated value $Z_s(t_i)$, and kriging identifies the data-to-unknown covariance, then by construction the $Cov\{Z_s(t_i), Z_s(t_j)\} = C(t_i - t_j)$ as it should be. Thus the simulation procedure consists of:

1. Perform kriging at the first time instant (node) t_i to obtain $Z^*(t_i)$ and $\sigma_K^2(t_i)$.
2. Assuming $Z^*(t_i)$ and $\sigma_K^2(t_i)$ to be the mean and variance of the probability distribution characterizing the RF $Z(t_i)$, draw a value from the distribution to yield a simulated value $Z_s(t_i)$.
3. Update this simulated value $Z_s(t_i)$ into the conditioning data set.
4. Proceed to the next node t_j, along a random path. Perform kriging at this node with the new conditioning data set.
5. With the kriged estimate $Z^*(t_j)$ and estimation variance $\sigma_K^2(t_j)$, draw a value from the probability distribution to obtain a new simulated value $Z_s(t_j)$ and update the conditioning data set by adding $Z_s(t_j)$.

Repeating Steps 4-5 until all nodes are exhaustively visited, the simulated values $Z_s(t)$ will reflect the target variance and reproduce the pattern of temporal variability in the form of covariance (variogram) identification. It should also be noted that since temporal nodes are visited along a random path and the simulated values $Z_s(t)$ are obtained by random sampling from the kriged distribution, multiple realizations $Z_s^{(l)}(t)$, $l = 1, \ldots, L$, of the temporal process are possible. These multiple realizations are all equiprobable representations of the temporal process.

Stochastic simulation can therefore be thought of as a procedure or process of representing the true underlying physical phenomena through probabilistic modeling of the temporal (and/or spatial) distribution of the RF $Z(t)$. Each outcome (realization), denoted by superscript (l), from the simulation process is equally probable. The higher the number of realizations, the better the simulated results reproduce the desired statistics (mean and variance). Among the many simulation algorithms currently available in the literature, the sequential simulation approach as outlined above is preferred due to its ability to ensure the correct pattern of temporal variability between the unknown samples and its effectiveness in restoring the variance of the true RF $Z(t)$.

The sequential simulation paradigm can also be interpreted as a technique for sampling from the multivariate probability distribution underlying the RF $Z(t)$, characterized by N-variate conditional cumulative density functions (ccdf's) with a set of N data, denoted as (N):

$$\text{Prob}\{Z(t_1) \le z(t_1), \ldots, Z(t_N) \le z(t_N) \mid (N)\} = F_Z\{z(t_1), \ldots, z(t_N) \mid (N)\} \tag{6.2}$$

The above multivariate distribution can be decomposed into multiple univariate ccdf's by applying Bayes' Rule, i.e.,

$$\begin{aligned} &\text{Prob}\{Z^{(l)}(t_1) \le z(t_1), \ldots, Z^{(l)}(t_N) \le z(t_N) \mid (N)\} \\ &\quad = F_Z\{z^{(l)}(t_N) \mid z^{(l)}(t_1), \ldots, z^{(l)}(t_{N-1}), (N)\} \\ &\quad \cdot F_Z\{z^{(l)}(t_1), \ldots, z^{(l)}(t_{N-1}), (N)\} \end{aligned} \tag{6.3a}$$

The second and the subsequent terms can be further decomposed as:

$$\begin{aligned} F_Z\{z^{(l)}(t_1), \ldots, z^{(l)}(t_{N-1}), (N)\} &= F_Z\{z^{(l)}(t_{N-1}) \mid z^{(l)}(t_1), \ldots, z^{(l)}(t_{N-2}), (N)\} \\ &\quad \cdot F_Z\{z^{(l)}(t_1), \ldots, z^{(l)}(t_{N-2}), (N)\} \\ &\quad \vdots \\ &\quad \cdot F_Z\{z^{(l)}(t_2) \mid z^{(l)}(t_1), (N)\} \\ &\quad \cdot F_Z\{z^{(l)}(t_1) \mid (N)\} \end{aligned} \tag{b}$$

where subscript s referring to the simulated values $z_s(t)$ is omitted for convenience. This implies that the multivariate distribution can be obtained by multiplying a series of univariate ccdf's, each increasingly updated by the conditioning data. To elaborate, the first step is to simulate a value $z^{(l)}(t_1)$ from the univariate ccdf of $Z(t_1)$, given a set of the original N data. Then add the simulated value $z^{(l)}(t_1)$ to the conditioning data set, whose size is now increased to $(N + 1)$. Draw another value $z^{(l)}(t_2)$ from the updated univariate ccdf , i.e., $F_Z\{z^{(l)}(t_2) \mid z^{(l)}(t_1), (N)\}$, and add it to the current data set, now with a larger dimension of $(N + 2)$. Repeat the subsequent steps of continually drawing a value from the updated ccdf and adding it to the data set until all nodes are considered. This process

will finally yield simulated values $z^{(l)}(t)$, $\forall t \in T$, that are correctly sampled (drawn) from the multivariate probability distribution characterizing RF $Z(t)$. In essence, this entire procedure is the core paradigm of the sequential simulation, of which a special case, i.e., sequential Gaussian simulation, is implemented in this thesis work.

6.1.1 Sequential Gaussian Simulation

The sequential simulation paradigm requires that the first two moments (mean and variance) established by kriging at any node be identified with the mean and variance of the conditional distribution $F_Z\{z(t_i) \mid z_s(t_j)\}$, $\forall j \neq i$, where $z_s(t_j)$ are the previously simulated values. If the conditional probability distributions are assumed Gaussian, the sequential simulation procedure yields samples from a multivariate Gaussian distribution. Thus the following properties hold:

- The conditional mean is a linear combination of the conditioning data, i.e.,

$$E\{Z(t_i) \mid Z_s(t_j)\} = \sum_{j=1}^{N_i} \lambda_j(t_i) Z_s(t_j), \quad \forall j \neq i$$

- The variance is independent of the data and strictly a function of variance or more precisely covariance, i.e.,

$$Var\{Z(t_i) \mid Z_s(t_j)\} = \sum_{j=1}^{N_i} \lambda_j(t_i) C_{ij}(t_i - t_j), \quad \forall j \neq i$$

It can be surmised by comparing the above expressions to those of kriging (for example, Eqs. 5.11-5.13) that the kriged estimator $Z^*(t)$ will be an ideal choice for estimating the mean and variance of the Gaussian conditional distributions.

The sequential Gaussian simulation approach capitalizes on these properties, and its procedure is discussed below, following Deutsch and Journel (1998):

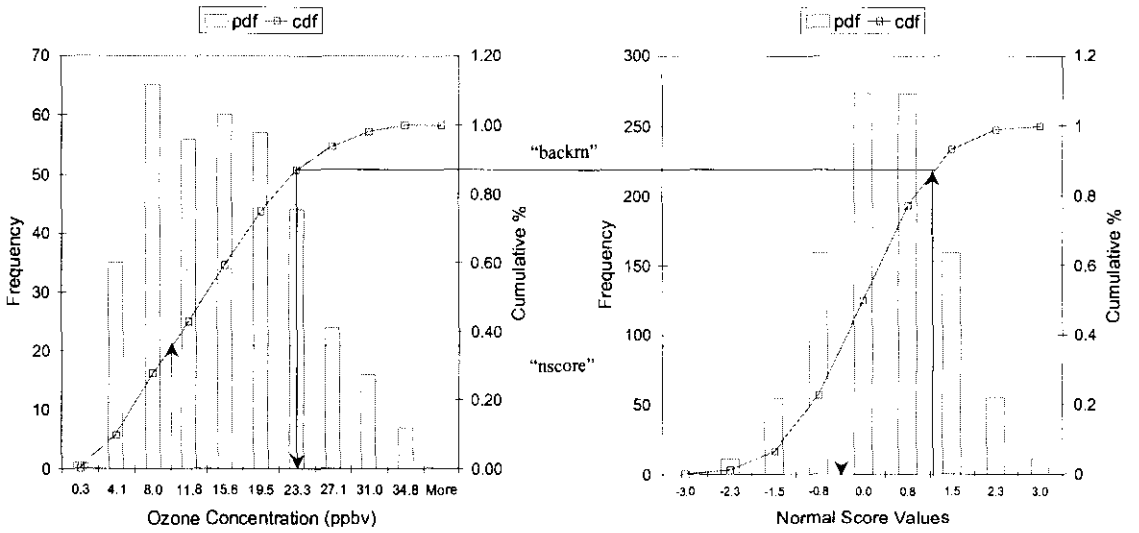
1. For a random function RF $Z(t)$ relevant to the case study, determine its univariate cumulative distribution function (cdf) $F_Z(z)$, which must be representative for the whole stationary domain. This means that the collection of data must be performed

wisely, in the sense that there is no preferential clustering over the temporal framework. If not, data declustering may be required to ensure the unbiasedness of data sampling.

2. Often the univariate distribution of the RF $Z(t)$ is not Gaussian; hence normal score transform of the sample data $z(t_k)$, k =1, …, N, into normally distributed data $y(t_k)$ is required prior to simulation. To elaborate, a set of evenly spaced (every 30 days) raw data $\{z(t_k), k = 1, \ldots, 12\}$ are transformed (Figure 6.1: "nscore") into the normal score values $y(t_k) = G_Y^{-1}\{F_Z[z(t_k)]\}$; G is the standardized (zero mean and unit variance) Gaussian operator defined as:

$$G(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(-\frac{1}{2}\zeta^2\right) d\zeta$$

Or, graphically:



(a) Raw ozone data (b) Normal score values

Figure 6.1
Graphical representation of normal score ("nscore") and back-transformed ("backtrn") procedures, denoted by the dashed and solid lines, respectively. For a better comparison, the corresponding histograms (pdf) and probability distributions (cdf) of the raw data and normal score (Gaussian) values are plotted on the same graph.

3. Before performing sequential Gaussian simulation, it is necessary to at least verify whether the data $y(t_i)$ are bivariate normal (Section V.2.2 in Deutsch and Journel, 1998). This verification is necessary since univariate Gaussianity (normality) does not automatically ensure normality of higher order distributions.

4. The steps for the sequential Gaussian simulation are:

   (a) Visit J simulation grid nodes $\{t_1, \ldots, t_j, \ldots, t_J\}$ sequentially along a pre-defined random path.

   (b) At each grid node t_j:

      i. Calculate the first two moments (mean and variance) of the local ccdf $F_Y[y(t_i) \mid y(t_k), k = 1, \ldots, N_i]$ using the kriging algorithm (Section 5.3.2). Note that on the first node t_1 along the random path, the original conditioning data set only consists of N normally distributed sample data $\{y(t_k), k = 1, \ldots, N\}$. As the simulation progresses, the conditioning data set, hereafter denoted as (N_i), also increases in size to include the N original data and all previously simulated values $y^{(l)}(t_i)$ in the neighborhood of node t_i.

      ii. Then, perform Monte Carlo drawing from the local ccdf $F_Y[y(t_i) \mid (N_i)]$ to obtained a simulated value $y^{(l)}(t_i)$ at time instant t_i. To elaborate, the drawing process starts with the generation of a uniformly distributed random number $\varphi^{(l)}$ within [0, 1] interval. Once $\varphi^{(l)}$ is identified, the simulated value $y^{(l)}(t_j)$ is easily acquired by applying the quantile transform:

$$y^{(l)}(t_j) = F^{-1}_{Y|(N_i)}[\varphi^{(l)}]$$

      where $F^{-1}(\cdot)$ refers to the quantile function of the distribution $F_Y[y(t_i) \mid (N_i)]$.

      iii. This simulated value $y^{(l)}(t_j)$ acts as an additional datum and therefore is included in the conditioning data set (N_i) for the next visit to node t_k along the random path. In essence, the size of the contemporaneous data set is now $N_j = N_i + 1$ due to adding the additional simulated value $y^{(l)}(t_i)$ at time instant t_i.

(c) The simulation is completed after all J grid nodes are visited by repeating step (3.b). This will, in turn, produces a set of simulated values $\{y^{(l)}(t_j), j = 1, \ldots, J\}$, which represents the l^{th} realization of the temporal process $Y^{(l)}(t)$.

5. Finally, the simulated normal values $y^{(l)}(t_j), j \in J = 365$ are back-transformed (Figure 6.1: "backtrn") using the quantile transform $z^{(l)}(t_j) = F_Z^{-1}\{G[y^{(l)}(t_j)]\}$; G is the Gaussian operator defined above.

6. In the case of multiple realizations, i.e., $\{y^{(l)}(t_j), j \in J\}$, $l = 1, \ldots, L$, Steps 4(b-c) are repeated L times. But now the time instants t_j, $\forall j \in J$, are visited along a different random path. The kriged conditional distribution is sampled using Monte Carlo drawing with a different random seed. The resultant realizations $y^{(l)}(t_j)$, $j \in J$, represent uncertainty in the ozone temporal profile due to limited samples.

6.1.2 Sequential Gaussian Co-Simulation

As mentioned above, sequential Gaussian simulation requires the kriging algorithm as its "driving engine." Similarly, the co-simulation procedure needs cokriging as well as the temporal correlations, i.e., auto and cross-variogram models (see Figure 5.10), of the entire process in order to improve ozone prediction. This way, not only the sample values $z_o(t_i)$, $i = 1, \ldots, N$, of the primary RF $Z_o(t_i)$ are used, the "soft data" from other RFs $Z_\alpha(t_i)$, $\alpha = 1, \ldots, M$, can also be capitalized upon for secondary information because, in some cases, they are often more extensively collected than the primary data. As an example, imagine a case when the measurement of ozone data in the year 2001 is interrupted for a month because of equipment failure, but data collections for other pollutants (e.g., total hydrocarbon THC) are still continued. By applying the temporal correlations, i.e., the variogram models, obtained from the previous year (2000) and using a few representative samples of ozone and THC in that particular year (2001), the ozone concentrations for the entire month can be fully predicted.

In other words, the sequential co-simulation procedure is very much alike to the one previously discussed except for one major difference; that is, the ordinary kriging (OK) algorithm applied in Step (4.b.i) is replaced with the ordinary cokriging (COK)

algorithm as the "driving engine" of the simulation process. Note that (co)kriging is retained as the underlying mechanism due to its excellent ability in data identification, i.e., at the data locations, the co-simulated outputs are ensured exact by construction. However, the tedious task is of course performing the linear model of coregionalization (LMC: Section 5.3.4) in order to determine the licit variogram model, and hence to guarantee the existence of a unique cokriging solution.

6.2 Results and Discussion

As mentioned earlier, sequential Gaussian simulation requires kriging as its "driving engine" that in turn utilizes a temporal correlation or variogram model (Figure 5.2). Recall that the variogram was modeled with two structures (Eq. 5.30). For all simulation studies, the GSLIB program **sgsim** (Appendix A.4) is utilized and the RF $Z(t)$, $\forall t \in T$, is assumed multivariate normal. The sequential Gaussian simulation was attempted using the two-structure variogram model (Eq. 5.30) and twelve data points, evenly placed at every 30th Julian day. The sequential simulation was carried out for three different cases:

1. ten realizations (L = 10), generated by the Monte Carlo sampling from the local conditional probability density function (cpdf),
2. ten sets of twelve randomly selected data between the 25th and 30th day of the month to account for the variance due to sampling, and
3. simulation using the hole-effect variogram model: $\gamma(\tau) = 1 - 0.5 \cdot \cos(\tau)$ with nugget effect C_o = 0.50, where τ is in radians, to replace the two-structure variogram model (Eq. 5.30).

The simulated results were averaged using a thirty-day moving window and then compared with the similar average profiles calculated using the raw data. Figure 6.2 shows the minimum and maximum (in a least-square sense) 30-day moving averages (30dMA) computed over the ten realizations. Figure 6.2 also shows the average outputs computed over the ten realizations (middle graphs) and the distributions of the correlation coefficients between the 30dMA of the raw data and the corresponding 30dMA over all realizations (bottom graphs). The temporal fluctuations of ozone concentrations are

reproduced by **sgsim**. While kriging and other forms of linear regressions yield average (smooth) predictions, stochastic simulation preserves the two-point correlation (variogram) and yields the correct fluctuations in ozone concentrations. In general, the correlation coefficients $\rho_{\alpha o}$ are greater than 0.70 except for 1999 where the $\rho_{\alpha o} \in$ [0.50, 0.79]

The influence of the conditioning data on the simulated profiles was explored by randomly selecting data between the 25th and 30th Julian day of each month over the whole year. Similar to the kriging case, ten sets of twelve randomly selected data are utilized. The 30dMA values of the minimum and maximum (in a least-square sense) simulated models (top graphs), and the average simulated outputs (middle graphs) over the ten realizations, as well as the distributions of the correlation coefficients between the 30dMA values of the raw data and the simulated results over the ten realizations are plotted in Figure 6.3.

In general, the results from the first two cases of sequential Gaussian simulation can be interpreted as follows:

- The results presented in Figure 6.2 represent the uncertainty in the temporal trend arising due to the lack of information beyond the two-point statistic (variogram) for predicting the temporal profiles. Thus the results in Figure 6.2 were obtained by ignoring the uncertainty in sampling. On the other hand, the results in Figure 6.3 represent the uncertainty in the temporal profiles due to the variance in sampling. The worst model (ρ = 0.38) therefore represents one realization conditioned to the worst possible configuration of data.
- The results for 1999 indicate that the correct peak between the 100th and 150th day of the year can indeed be correctly reproduced if the conditioning data reflect the underlying temporal phenomena accurately. Any one configuration of the conditioning data can result in further uncertainty in the simulated profiles due to the inadequacy of the two-point statistic to depict the temporal trends. Thus the data configuration that corresponds to ρ = 0.77 in Figure 6.3 can further yield a

range of uncertainty $\rho \in [0.50, 0.79]$ (Figure 6.2) when multiple realizations are simulated using that conditioning data.

Specifically, the discrepancy between the simulated and target ozone profiles can be attributed to:

- Inconsistency between the variogram interpreted on the 1997 data and the actual temporal variation exhibited in 1999.
- Inadequacy of the twelve conditioning data to represent the underlying temporal phenomena. For example, a peak in ozone trend is observed between the 100^{th} and 150^{th} day of 1999. The nearest conditioning datum is on the 120^{th} day and the instantaneous ozone concentration on that day happens to be low. This causes the simulated profile peak at a later time resulting in the offset with respect to the actual trend.

Next, simulation using the hole-effect variogram model was attempted in order to account for the annual periodicity in the observed ozone behavior. The 30dMA values of the hole-effect results are plotted in Figure 6.4, and compared with those of the raw data and the simulated (sequential Gaussian) results using the two-structure variogram model (Eq. 5.30). In 1997, the hole-effect model overestimates the temporal trend whereas in 2000, it grossly underestimates the yearly ozone trend. This is expected since the hole-effect model was fitted using the long-term ozone trend, i.e., it summarizes the temporal correlation over a four-year study period, and thus only represents the ozone trends in the sense of the long-term average. To elaborate, the 30-day average time series plots of ozone data have distinct features from one year to another; those of 1997 and 1999 share some resemblance since they only have one large peak occurring between the 100^{th} and 150^{th} Julian day of the year. On the other hand, the trends in 1998 and 2000 depict one large peak around the same period of the year, and in addition, a smaller peak (flattened profile) in the summer season (around the 170^{th} to 240^{th} Julian day). In essence, the hole-effect model imposes cyclic phenomena of mean amplitude (0.5) and intra-year periodicity (about 365 days) on the simulated results based on the data for 1998-2000.

To account for the influence of secondary information ("soft data"), co-simulation was performed. Since the total hydrocarbon (THC) is slightly better correlated with ozone in 1997 ($\rho_{10} = 0.68$) than nitric oxide (NO) ($\rho_{20} = 0.67$), sequential co-simulation was performed with THC as the covariate. The linear model of coregionalization (LMC) was developed in a manner similar to that for cokriging. A variogram model with two structures was selected. Since there is only one covariate, the size of the coregionalized matrices $\mathbf{B}^{(k)}$, $k = 1, \ldots, K$, is reduced to (2 x 2). Expression (6.6) below summarizes the variogram model in the matrix form:

$$\gamma(\tau) = \begin{bmatrix} 0.56 & 0.24 \\ 0.24 & 0.46 \end{bmatrix} \cdot Exp\left(\frac{\tau}{5}\right) + \begin{bmatrix} 0.44 & 0.44 \\ 0.44 & 0.54 \end{bmatrix} \cdot Gauss\left(\frac{\tau}{120}\right) \tag{6.4}$$

The legitimacy of the above model can be verified by calculating the corresponding determinants of the coregionalized matrices, i.e., $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$. The co-simulation results in the form of one realization are plotted in Figure 6.5. The simulated results indicate that the covariate (THC) does indeed provide additional information regarding the amplitude of the temporal phenomena. The troughs observed in the "true" profiles between the 300th and 350th day of 1997 and 1999 are reproduced quite well as compared to the previous simulation results in Figure 6.2. The peak in ozone trend between the 100th and 150th day of the 1999 profile is again simulated much later in the year. As explained previously, the offset between the simulated and the actual profiles can be explained on the basis of the location and magnitude of the conditioning data. In general, the simulated profiles for all four years (1997-2000) exhibit much less amplitude of variations. This reduction in variance is due to the integration of secondary information. The peaks appear shifted and the extent of the shift is directly related to the temporal profiles of the secondary variable (THC) as previously illustrated in Figure 5.10.

Co-simulation was also attempted using NO as the covariate. The coregionalized matrices $\mathbf{B}^{(k)}$, $k = 1, \ldots, K$, for the variogram model was recalculated to uphold the positive-definite constraints, and hence the existence of unique solutions. Since the auto-variogram of NO is very similar to that of THC, the LMC model for NO is likely to

resemble that of THC. The corresponding isotropic variogram model, written in matrix notation, is:

$$\gamma(\tau) = \begin{bmatrix} 0.56 & 0.26 \\ 0.26 & 0.50 \end{bmatrix} \cdot Exp\left(\frac{\tau}{5}\right) + \begin{bmatrix} 0.44 & 0.41 \\ 0.41 & 0.50 \end{bmatrix} \cdot Gauss\left(\frac{\tau}{120}\right) \tag{6.7}$$

and the co-simulation results obtained using the above LMC model are plotted in Figure 6.6. Except for the slight improvement in the simulated profile for 2000 (between the 250^{th} and 300^{th} day), the overall trend is strikingly similar, which is anticipated since the LMC models corresponding to NO and THC (Figure 5.9) bear a very close resemblance to each other. To investigate further, the sample variograms of the co-simulation results using THC and NO as covariates are plotted in Figure 6.7 and 6.8. Evidently, the variograms are very much alike suggesting that the co-simulation results using these two covariates are similar. Thus the presence of NO as another secondary information is redundant, i.e., it makes no extra contribution towards predicting temporal ozone trends.

The above co-simulation results are obtained over one realization only. However, stochastic simulation can be fully capitalized if more realizations are performed. Figure 6.9 shows the co-simulation results corresponding to ten realizations and the "true" profiles. Except in 1999, the minimum and maximum (in a least-square sense) 30dMA values of the co-simulated outputs (top graphs) enclose the raw trends throughout the whole year. The average trends computed over the ten realizations (middle graphs) also emulate the "true" trends quite well. The correlation coefficients calculated over the ten realizations indicate a general increase. Thus the addition of covariate data enhances the accuracy of sequential simulation. The simulated results for 1999 do indicate little (if any) improvement. The relatively poor performance for 1999 is mainly due to the non-representative conditioning data used for the simulation. The integration of the auxiliary data in the form of THC or NO does relatively little to mask the effect of the conditioning ozone data.

By and large, the implementation of sequential simulation algorithms for modeling the temporal trends of ozone concentrations does result in improved

reproduction of the "true" temporal fluctuations. The pattern of temporal variation captured in the form of a temporal variogram is reproduced in an ergodic sense (average over multiple realizations) by the simulated realizations. Improved accuracy of the simulated profiles is achieved by integrating auxiliary information in the form of THC or NO covariate data. The integration of such covariate information requires a LMC model to be fitted jointly, which provides a licit representation of the auto and cross-covariances (variograms).

As mentioned in Chapter 2, various statistical methods have been attempted for studying atmospheric phenomena. In the space-time approach, ozone is treated as a random variable $Z(t)$, which can then be decomposed into a trend $M(t)$ and a residual $R(t)$ component. At a station location, the temporal trend is often modeled deterministically using a noise-filtering algorithm such as a Fourier series analysis due to the availability of environmental records. Such a technique will be performed next in the case of modeling seasonal and annual ozone trends. Another popular approach is to train a neural network, or more specifically the multilayer perceptron (MLP), using two independent data sets. The network weights obtained from such training procedure can be applied in predictive modes. As will be seen subsequently, the neural network is an alternative approach to reproducing the temporal patterns; however, it suffers from the requirement of having quasi-exhaustive data.

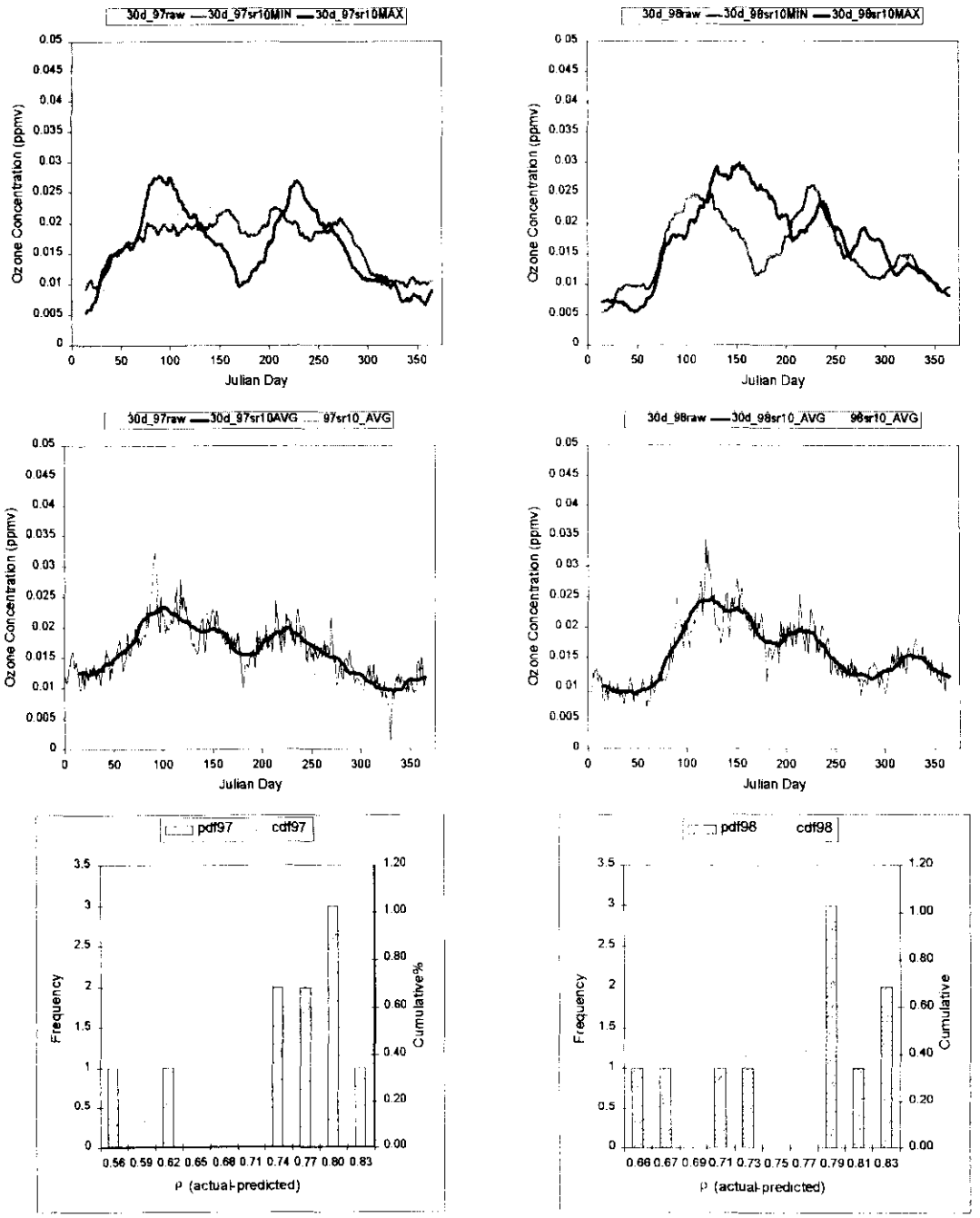


Figure 6.2 (i)

Sequential Gaussian simulation results over ten realizations for 1997 [**LEFT**] and 1998 [**RIGHT**] based on twelve data points, evenly spaced at every 30th Julian day of the year. In a least-square sense, the 30-day moving averages (30dMA) of the minimum (green) and maximum (red) [**top**] realizations, as well as the *daily* average fluctuations (blue) [**middle**] are superimposed on those of raw data (gray). The distributions of correlation coefficients between the 30dMA of raw data and the simulated realizations are also plotted [**bottom**].

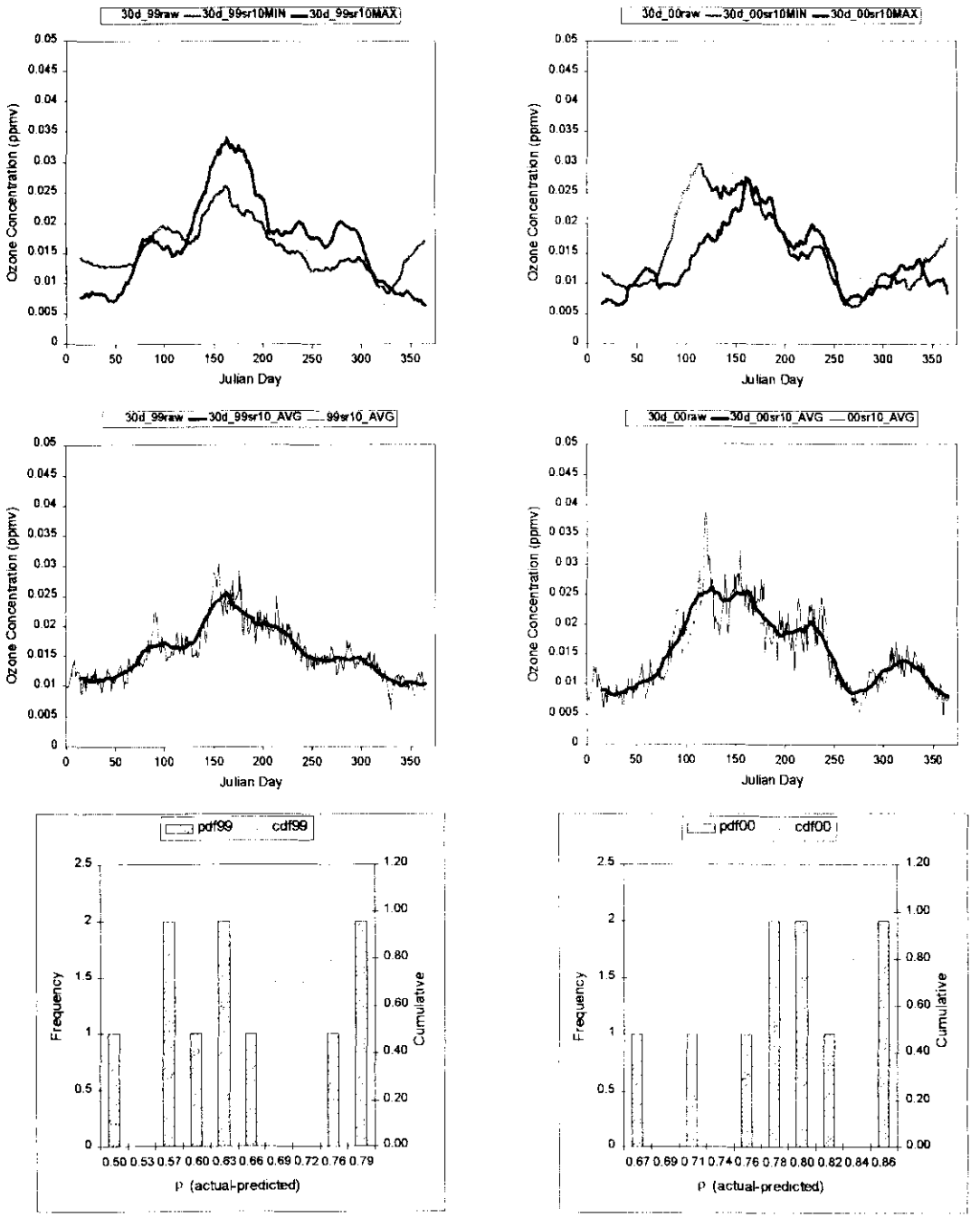


Figure 6.2 (ii)

Sequential Gaussian simulation results over ten realizations for 1999 [**LEFT**] and 2000 [**RIGHT**] based on twelve data points, evenly spaced at every 30th Julian day of the year. In a least-square sense, the 30-day moving averages (30dMA) of the minimum (green) and maximum (red) [**top**] realizations, as well as the *daily* average fluctuations (blue) [**middle**] are superimposed on those of raw data (gray). The distributions of correlation coefficients between the 30dMA of raw data and the simulated realizations are also plotted [**bottom**].

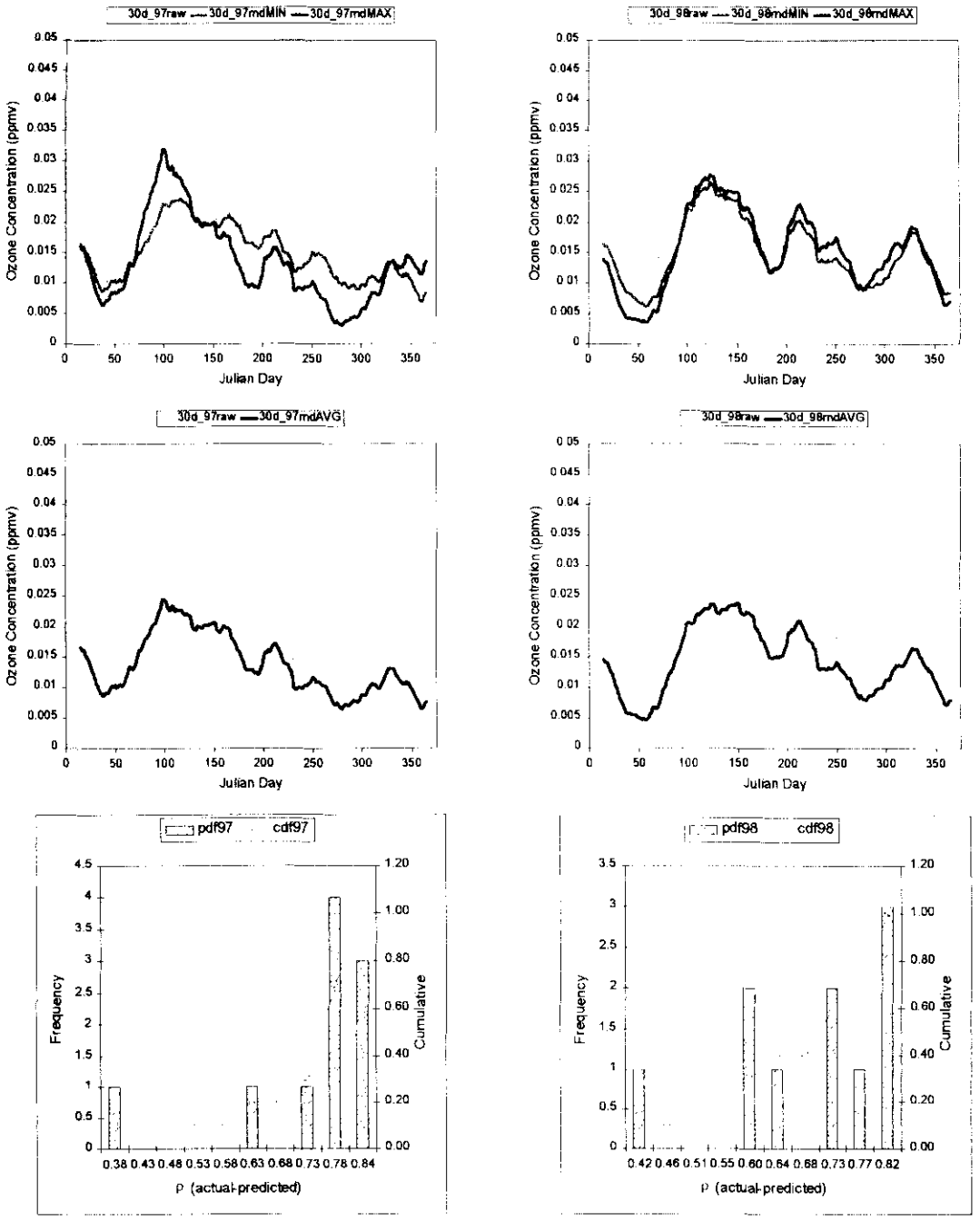


Figure 6.3 (i)

Sequential Gaussian simulation using ten sets of twelve randomly selected data between the 25th and 30th Julian day of the month for 1997 [**LEFT**] and 1998 [**RIGHT**]. In a least-square sense, the 30-day moving averages (30dMA) of the minimum (green) and maximum (red) [**top**], as well as the average (blue) [**middle**] of the ten results are superimposed on those of raw data (gray). The distributions of correlation coefficients between the 30dMA of raw data and the simulated results are also plotted [**bottom**].

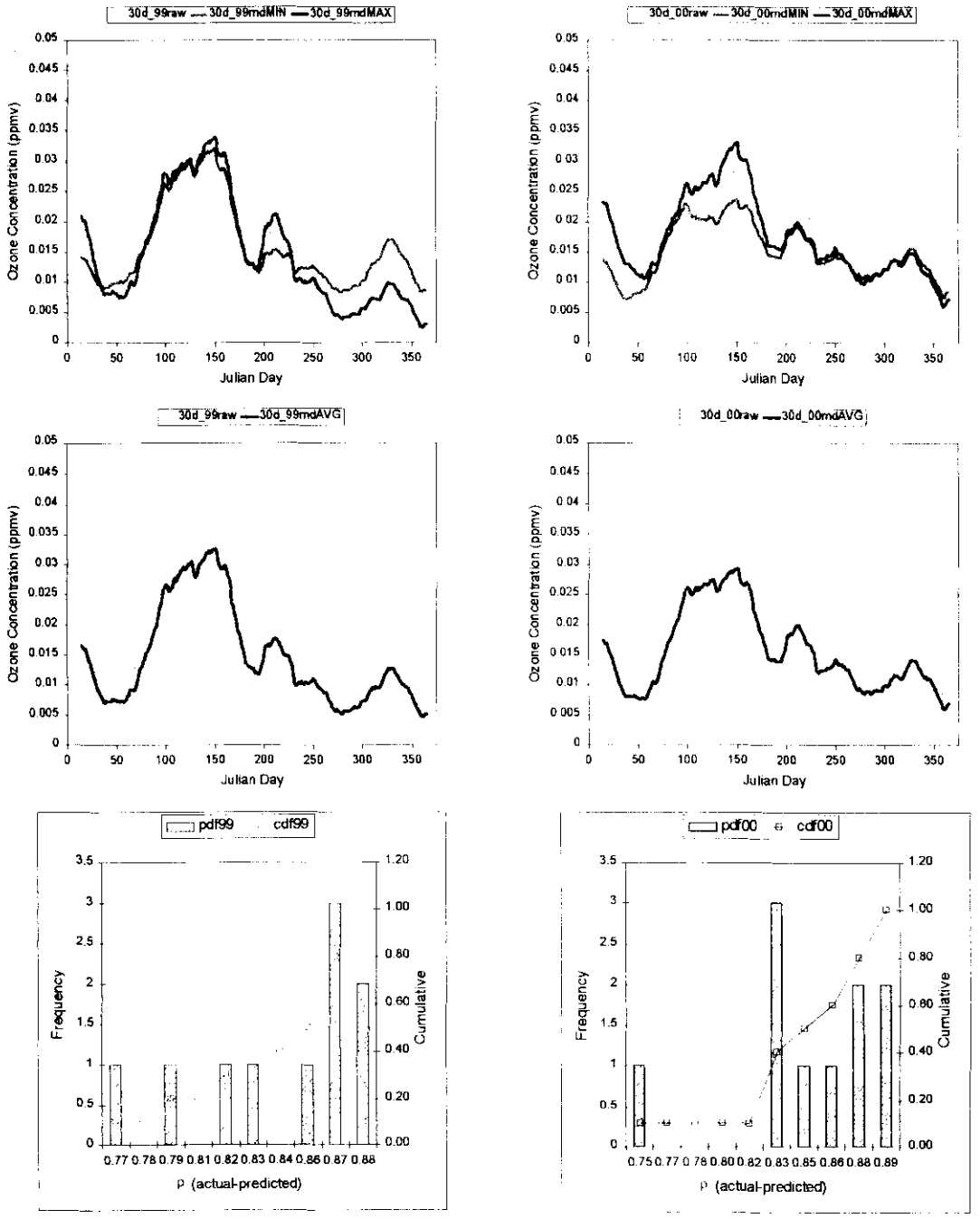
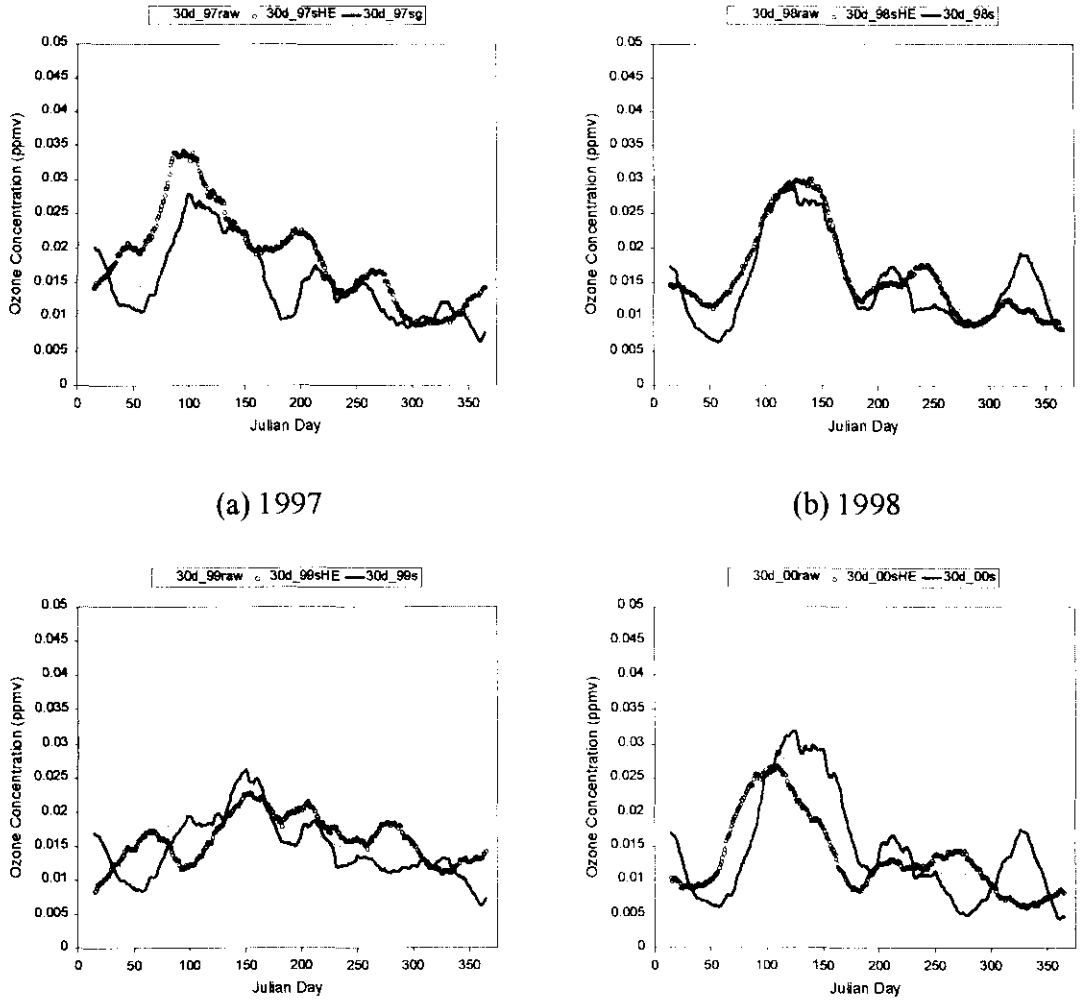


Figure 6.3 (ii)

Sequential Gaussian simulation using ten sets of twelve randomly selected data between the 25th and 30th Julian day of the month for 1999 [**LEFT**] and 2000 [**RIGHT**]. In a least-square sense, the 30-day moving averages (30dMA) of the minimum (green) and maximum (red) [**top**], as well as the average (blue) [**middle**] of the ten results are superimposed on those of raw data (gray). The distributions of correlation coefficients between the 30dMA of raw data and the simulated results are also plotted [**bottom**].



(a) 1997 (b) 1998

(c) 1999 (d) 2000

Figure 6.4
Sequential Gaussian simulation results based on the hole-effect (HE) variogram model $\gamma(\tau) = 1 - 0.50 \cdot \cos(2\pi\tau/365)$. The figure shows the 30-day moving averages (30dMA) computed on: raw data (thin gray line), hole-effect simulated realization (open blue circle), and one realization based on the two-structure variogram model $\gamma(\tau) = 0.50 \cdot Exp(\tau/5) + 0.50 \cdot Gauss(\tau/100)$ (thick pink line).

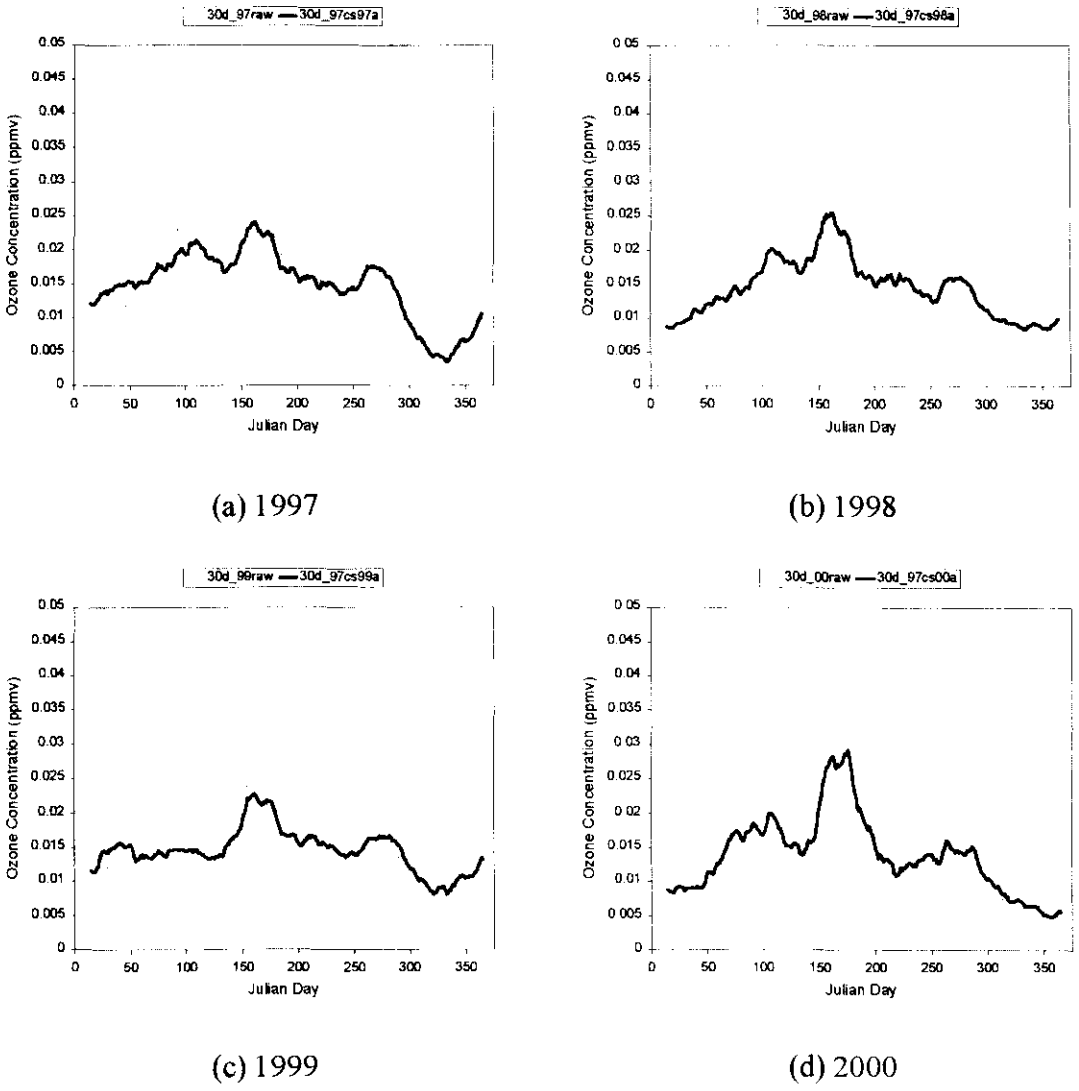
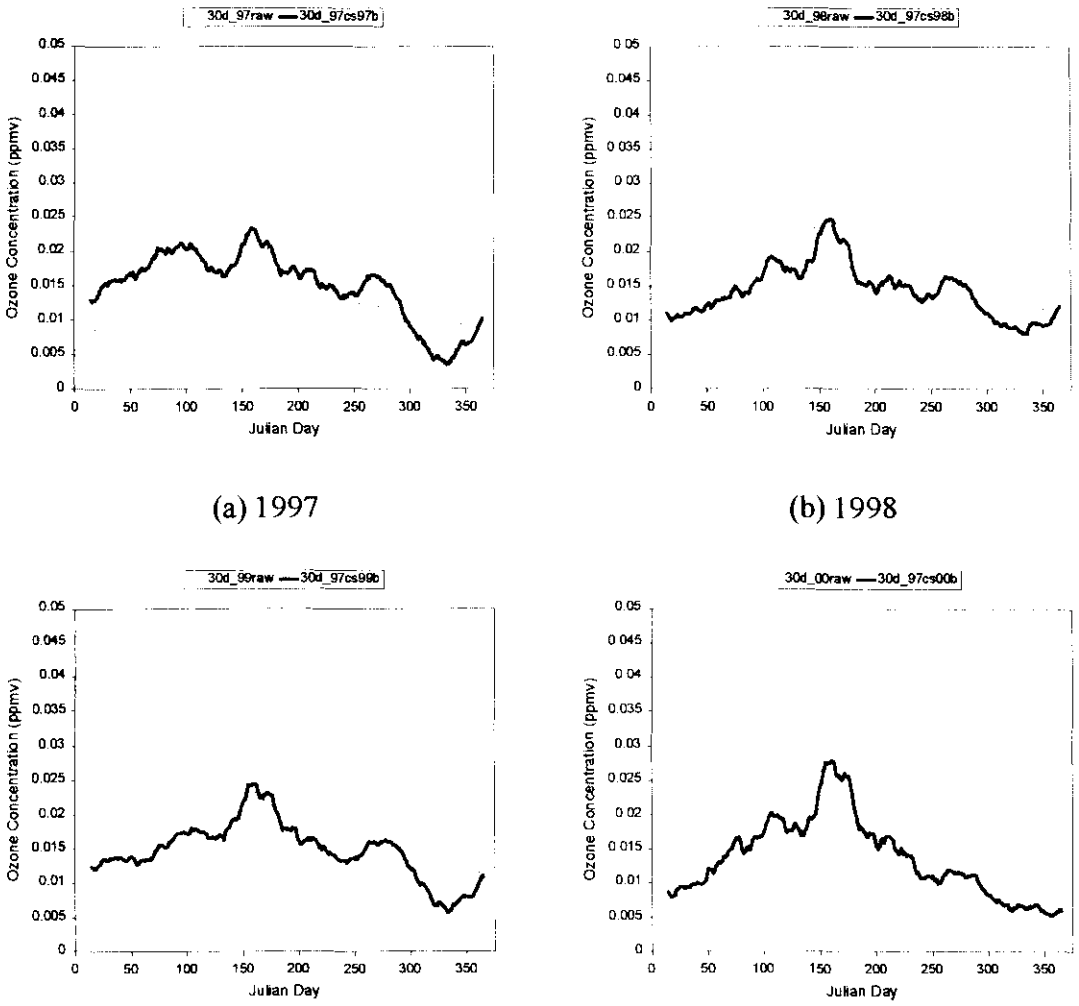


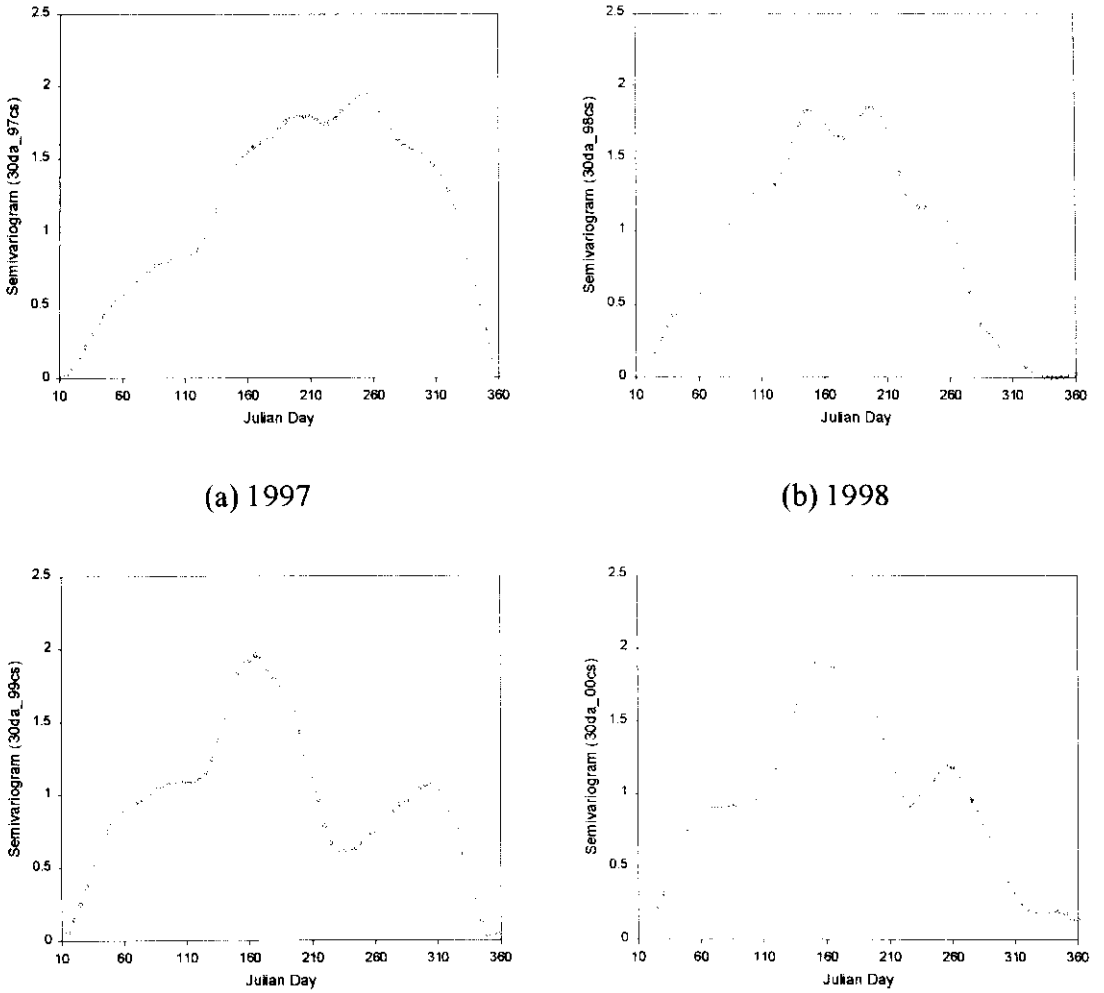
Figure 6.5
30-day moving averages (30dMA) calculated on one realization obtained by sequential Gaussian co-simulation (thick blue line), condition to twelve data evenly spaced on every 30th Julian day of the year. The corresponding "true" profiles are also shown (thin gray line). The covariate used for co-simulation is total hydrocarbon (THC).



(a) 1997 (b) 1998

(c) 1999 (d) 2000

Figure 6.6
30-day moving averages (30dMA) calculated on one realization obtained by sequential Gaussian co-simulation (thick blue line), condition to twelve data evenly spaced on every 30th Julian day of the year. The corresponding "true" profiles are also shown (thin gray line). The covariate used for co-simulation is nitric oxide (NO).



(a) 1997 (b) 1998

(c) 1999 (d) 2000

Figure 6.7
Sample variograms of the 30-day moving average values (30dMA) of the co-simulation outputs using total hydrocarbon (THC) as the covariate. Since the initial 30dMA value of the results and raw data are placed on the 15th Julian day of the year, the first time instant of the semivariogram must, by construction, also be placed on the same day.

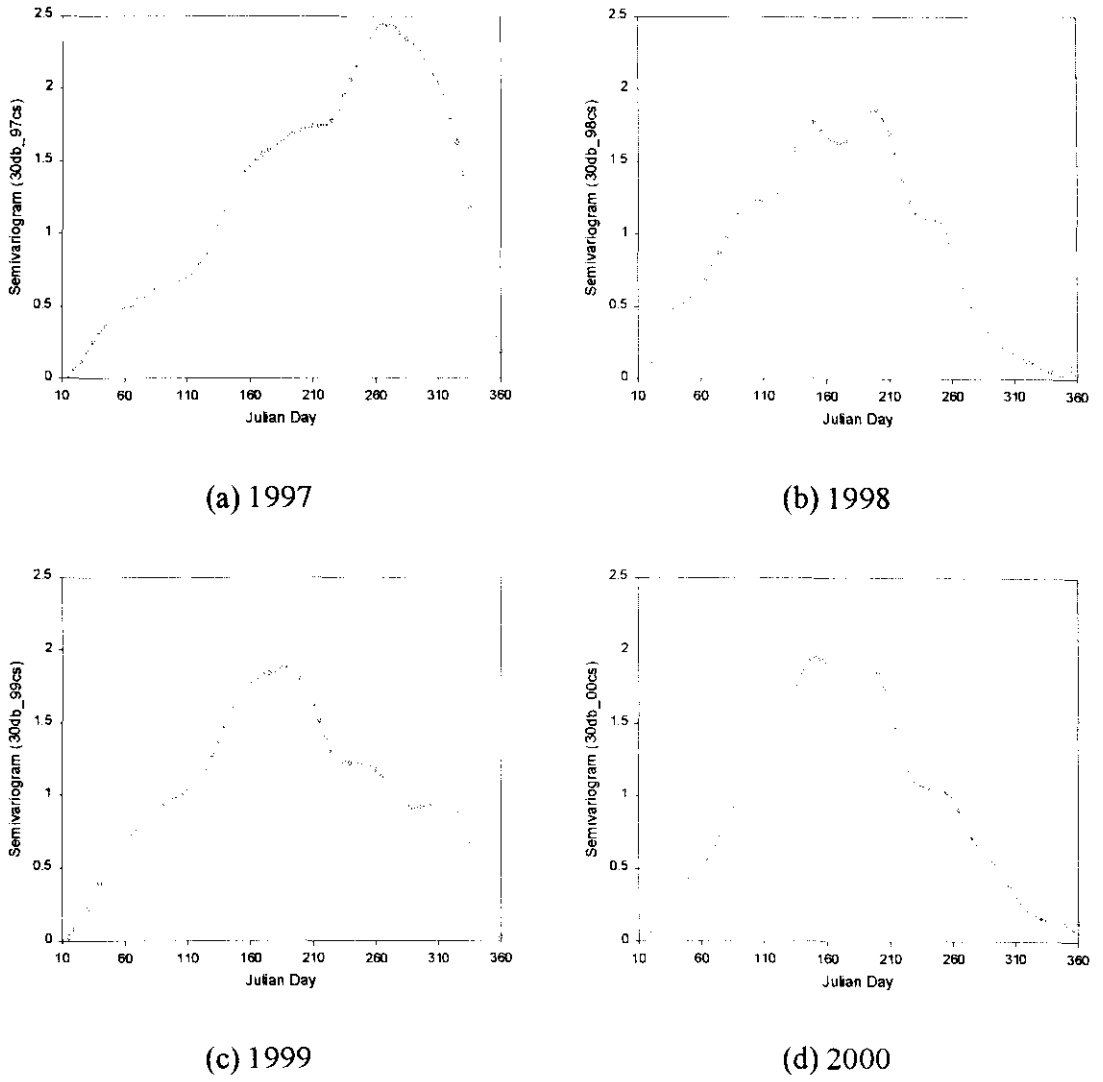


Figure 6.8
Sample variograms of the 30-day moving average values (30dMA) of the co-simulation ouputs using nitric oxide (NO) as the covariate. Since the initial 30dMA value of the results and raw data are placed on the 15th Julian day of the year, the first time instant of the semivariogram must, by construction, also be placed on the same day.

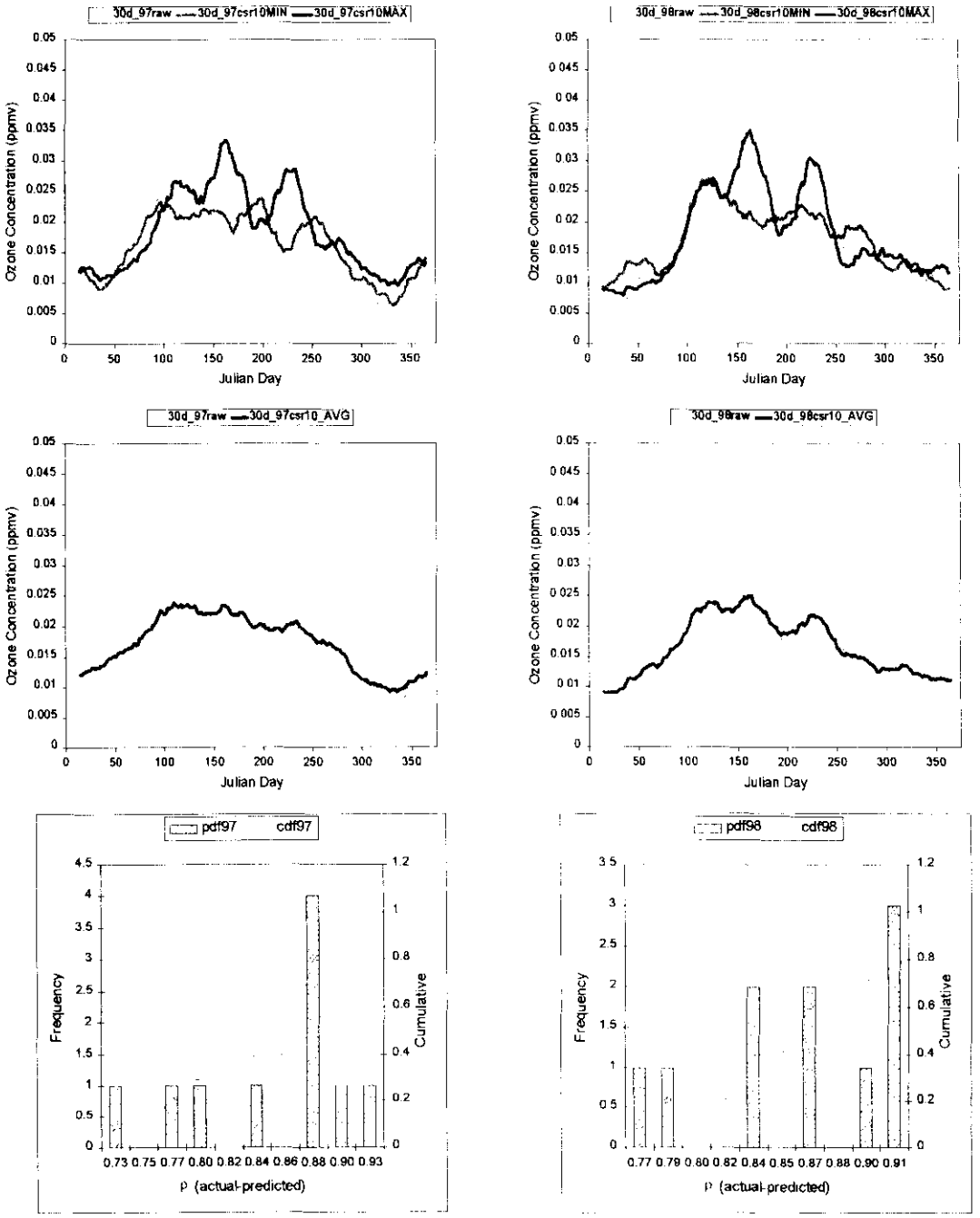


Figure 6.9 (i)

Sequential Gaussian co-simulation (using total hydrocarbon THC as the covariate) over ten realizations for 1997 [LEFT] and 1998 [**RIGHT**] based on twelve evenly spaced data. In a least-square sense, the 30-day moving averages (30dMA) of the minimum (green) and maximum (red) [**top**], as well as the average (blue) [**middle**] realizations are superimposed on those of raw data (gray). The distributions of correlation coefficients between 30dMA of the raw data and the simulated realizations are also plotted [**bottom**].

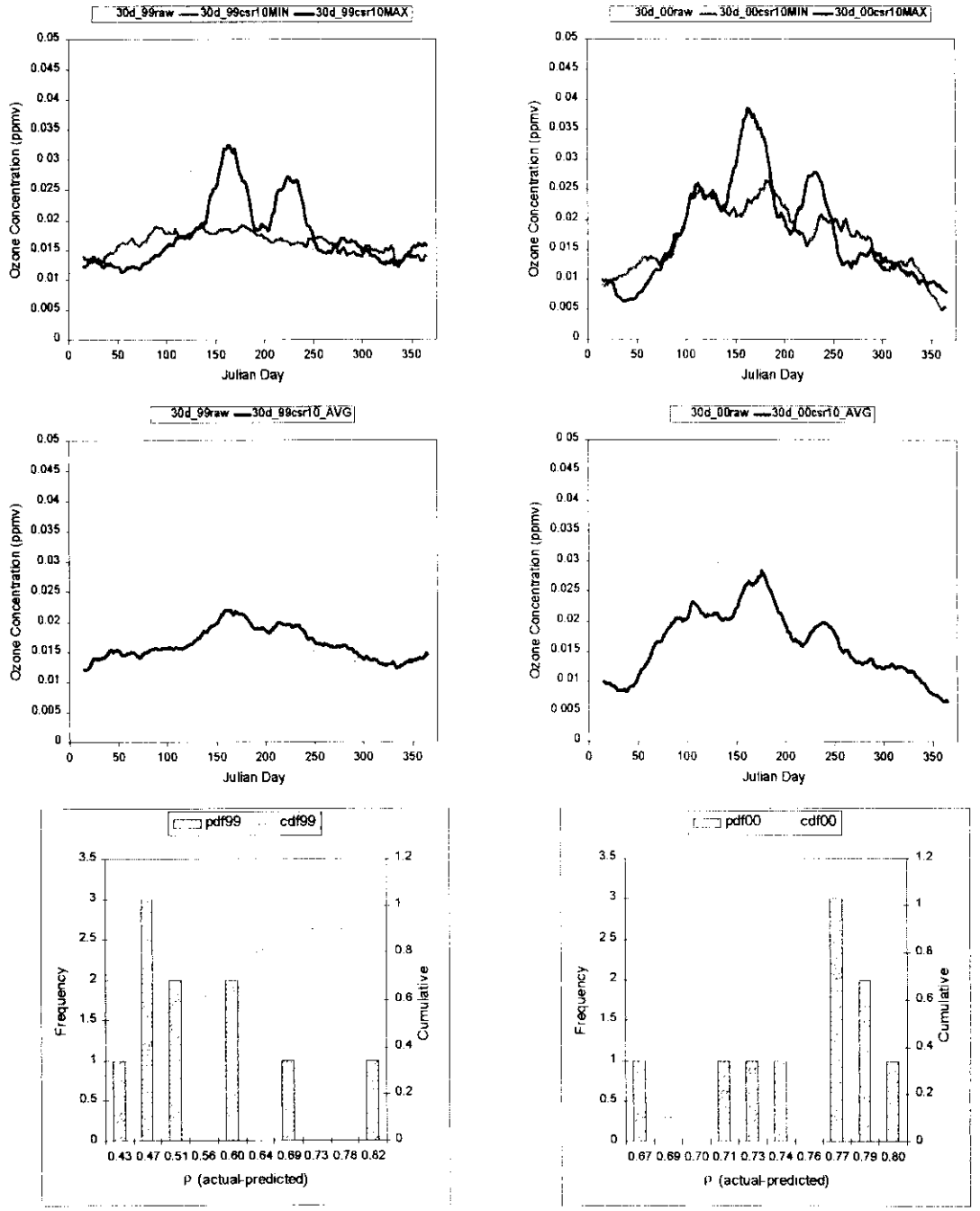


Figure 6.9 (ii)
Sequential Gaussian co-simulation (using total hydrocarbon THC as the covariate) over ten realizations for 1999 [LEFT] and 2000 [RIGHT] based on twelve evenly spaced data. In a least-square sense, the 30-day moving averages (30dMA) of the minimum (green) and maximum (red) **[top]**, as well as the average (blue) **[middle]** realizations are superimposed on those of raw data (gray). The distributions of correlation coefficients between 30dMA of the raw data and the simulated realizations are also plotted **[bottom]**.

CHAPTER 7

FOURIER SERIES AND NEURAL NETWORK ANALYSES

This chapter serves as a bridge between the deterministic and stochastic approaches. The Fourier series analysis is often utilized for filtering the 'noise' of the random data in order to better visualize the underlying temporal trend $m(t)$ deterministically. In this case, the singular value decomposition (SVD) technique is employed to determine the Fourier coefficients of the sine and cosine series. If the "true" trend can be reproduced, at least, in an ergodic sense, the residual component $R(t)$ can be obtained using such method as sequential simulation (Chapter 6). In the field of atmospheric science, the application of neural networks has gained momentum due to its ability to reproduce the complex patterns of the temporal profiles. Nonlinear associations between the inputs (predictors) and output (ozone) variables are tackled via activation functions at the hidden and, if necessary, at the output nodes.

7.1 Fourier Series

Historically, Fourier series fall into the category of orthogonal sets of functions, which by definition are sets of jointly perpendicular vectors (Churchill, 1963). Consider a three-dimensional (3D) vector $g(\mathbf{r})$, $\mathbf{r}$ = 1, 2, 3, in Euclidean space. This vector has a length $\| g \|$, called norm, and defined as:

$$\| g \| = \sqrt{[g(1)]^2 + [g(2)]^2 + [g(3)]^2}$$

or in words, the norm of a vector $g(\mathbf{r})$ is the quadratic summation of its components; the result is square rooted and taken as only the positive number. If $\| g \| = 1$, $g(\mathbf{r})$ is termed a unit vector or sometimes, normalized vector. On the other hand, $g(\mathbf{r})$ is labeled a zero vector if $\| g \| = 0$, and this can only happen when each of its component is zero.

Now consider two 3D vectors, $g_1(\mathbf{r})$ and $g_2(\mathbf{r})$, also in Euclidean space. Their algebraic operations, i.e., addition, subtraction and multiplication by a scalar, are linear. The scalar product, or inner product, of these vectors is denoted $\langle g_1, g_2 \rangle$ and written as:

$$\langle g_1, g_2 \rangle = \sum_{r=1}^{3} g_1(r) g_2(r) = \| g_1 \| \| g_2 \| \cos\theta$$

Note that the angle θ between $g_1(\mathbf{r})$ and $g_2(\mathbf{r})$ only exists when both these vectors are nonzero. Keeping this in mind, two nonzero vectors $g_1(\mathbf{r})$ and $g_2(\mathbf{r})$ can only be orthogonal (i.e., perpendicular to each other), if the following condition is satisfied:

$$\langle g_1, g_2 \rangle = 0$$

which can only occur when cos θ is zero, i.e., when $\theta = k\frac{\pi}{2}$, $| k | = 1, 3, \ldots, \infty$.

If there are n such orthogonal nonzero vectors $\{g_n(\mathbf{r}), n = 1, \ldots, \infty\}$ and each is divided by its norm, the results constitute a set of unit vectors ϕ_n. These unit-vectors ϕ_n are mutually perpendicular and therefore termed orthonormal. Analogous to orthogonal vectors, an orthonormal set $\{\phi_n\}$ may also be described by means of inner product:

$$\langle \phi_m, \phi_n \rangle = \delta_{mn}, \qquad m, n = 1, \ldots, \infty$$

where δ_{mn} is the Kronecker delta, whose value is one if $m = n$ and zero otherwise. The above condition is termed the property of orthogonal (or orthonormal) functions and regularly encountered in various subjects. For example, in the method of separation of variables commonly used in the boundary value problems, the auxiliary solution of a parabolic PDE (a Sturm-Liouville problem), in terms of infinite series, is obtained by utilizing the orthogonality of the eigenfunctions (e.g., Ozisik, 1993)

Taking advantage of this property, an arbitrary function $f(x)$ may be represented by a linear combination of orthonormal functions $\{\phi_n(x), n = 1, \ldots, \infty\}$ within the interval $(-L, L)$ and generalized as:

$$f(x) = c_1 \phi_1(x) + c_2 \phi_2(x) + \ldots + c_n \phi_n(x) + \ldots \quad \forall x \in (-L, L)$$

where c_n ($n = 1, \ldots, \infty$) are scalars, obtained by multiplying both sides of the equation by $\phi_m(x)$ and integrating over the interval $(-L, L)$ as follows:

$$\int_{-L}^{L} f(x)\phi_m(x)\,dx = \int_{-L}^{L} \left[c_1\langle\phi_1,\phi_m\rangle + c_2\langle\phi_2,\phi_m\rangle + \ldots + c_n\langle\phi_n,\phi_m\rangle + \ldots\right] dx$$

The inner product of the orthonormal functions can only assume a value when $m = n$, due to the fact that $\langle\phi_m,\phi_n\rangle = \delta_{mn}$ or when written in complete form:

$$\int_{-L}^{L} \phi_m(x)\phi_n(x)\,dx = \begin{cases} 0 & \text{if } m \neq n \\ 1 & \text{if } m = n \end{cases}$$

As a result, the only non-vanishing terms on the right side of the equation are c_n, i.e.,

$$c_n = \int_{-L}^{L} f(x)\phi_n(x)\,dx, \quad n = 1, \ldots, \infty$$

which are called Fourier constants for the function $f(x)$ corresponding to the orthonormal set $\{\phi_n(x)\}$. The infinite series

$$f(x) \approx \sum_{n=1}^{\infty} c_n\phi_n(x) = \sum_{n=1}^{\infty} \phi_n(x) \int_{-L}^{L} f(\xi)\phi_n(\xi)\,d\xi$$

is termed the generalized Fourier series, whose convergence property is established by the Fourier theorem (Churchill, 1963).

In practice, the orthonormal functions $\{\phi_n(x),\ n = 1,\ldots,\infty\}$ in the Fourier series are represented by a complex exponential form (e.g., Lighthill, 1958):

$$f(x) = \sum_{n=-\infty}^{\infty} c_n \exp\left(\frac{in\pi x}{L}\right); \quad c_n = \frac{1}{2}(a_n - ib_n), \forall x$$

where $i = \sqrt{-1}$; a_n and b_n are defined below. For an even (symmetric) Fourier series, each term of the series is periodic within $2L$. As a consequence, the series converges to a

periodic function $h(x)$, which coincides with $f(x)$, within the fundamental interval $(-L, L)$. This type of Fourier series is important because it serves two purposes: (1) to represent an orthonormal set of functions $\{\phi_n(x), n = 1,\ldots,\infty\}$ defined over the interval $(-L, L)$, for all values of x, and (2) to represent a periodic function $h(x)$ with period $2L$, for all values of x. For convenience, however, a Fourier series of period $2L$ is more commonly written in trigonometric form to avoid dealing with imaginary numbers (e.g., Mickley et al., 1957):

$$f(x) = \sum_{n=0}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=0}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) \tag{7.1a}$$

where the coefficients a_n and b_n are given as:

$$a_o = \frac{1}{2L}\int_{-L}^{L} f(x)\,dx$$

$$a_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{n\pi x}{L}\right)dx$$

$$b_o = 0$$

$$b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\left(\frac{n\pi x}{L}\right)dx \tag{7.1b}$$

Note that the integration of a temporal trend component in inter/extrapolation of time series data would require specification of seasonal variations in the response variable (i.e., ozone). The Fourier series expressed as trigonometric functions, specifically the half-range Fourier series in terms of the cosine functions only, is a viable way to model positive-definite covariance functions (e.g., hole-effect) as presented in Chapter 5.

Fourier series analysis of the form:

$$y(t) = a_o + \sum_{n=1}^{N} a_n \cos\left(\frac{2n\pi t}{M}\right) + \sum_{n=1}^{N} b_n \sin\left(\frac{2n\pi t}{M}\right) \tag{7.2}$$

is utilized in the following analysis. The time t is expressed in Julian day, e.g., January 1 and December 31 equal to 1 and 365, respectively, and M is now the number of days in a year (365). For 2000, the data on the last day of this leap year is neglected, without loss of generality, so that M is always equal to 365. In the case of daily-average data analysis

(as in this work), this step results in the elimination of one datum while that of hourly average, the same step results in the "loss" of twenty-four data points.

Theoretically, as the number of coefficients N approaches infinity, the original data are reproduced. However, the increment of N results in numerical error associated with the computation of the coefficients a_n and b_n; thus creating noise in the interpolation. A feasible way to reduce the number of coefficients N required to fit the temporal trend is through the use of singular value decomposition (SVD), which is discussed next.

7.1.1 Singular Value Decomposition

When dealing with large systems of equations or matrices, often the methods of Gaussian elimination, LU, QR and Cholesky decompositions fail to give satisfactory solutions due to the presence of singular or numerically close to singular matrices. One approach for solving such systems, at least, in the linear least squares sense is to perform a singular value decomposition (SVD) (e.g., Numerical Recipes).

The SVD technique is based on the following theorem of linear algebra, which states that any (M x N) matrix **A** can be decomposed as:

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^{\mathrm{T}} \tag{7.3}$$

or when expressed in matrix form,

$$\begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{1N} \\ \vdots & \ddots & \vdots \\ u_{M1} & \cdots & u_{MN} \end{bmatrix} \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_N \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & \cdots & v_{1N} \\ \vdots & v_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ v_{N1} & \cdots & \cdots & v_{NN} \end{bmatrix}^T$$

(M x N) (M x N) (N x N) (N x N)

The (M x N) matrix **U** is termed column-orthogonal, i.e. $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ for $i \neq j$, where $\mathbf{u}_i$ and $\mathbf{u}_j$ denote the i^{th} and j^{th} column vectors, respectively; column-orthogonal because its number of rows may exceed that of columns. Matrix **U** is also column-orthonormal in

the sense that $\sum_{i=1}^{M} u_{ki}u_{in} = \delta_{kn}$, $(1 \le k \le N), (1 \le n \le N)$ where δ_{kn} is the Kronecker delta, which equals to one if $k = n$ and zero otherwise. The (N x N) matrix **W** consists of only positive diagonal elements w_j (the singular values), and the transpose of the (N x N) square matrix **V** is also orthogonal. In addition, because **V** is an (N x N) square matrix, it is also row-orthonormal in the sense that $\mathbf{V}\cdot\mathbf{V}^T = \mathbf{I}$, where **I** is the identity matrix having the same dimension as **V**. Or using familiar notation, the condition for orthonormality can be written as $\sum_{i=1}^{M} v_{ki}v_{in} = \delta_{kn}$, $(1 \le k \le N), (1 \le n \le N)$, where δ_{kn} is the Kronecker delta, whose value is one if $k = n$ and zero otherwise. One unique property of the SVD is that the decomposition of an (M x N) matrix **A** remains true even if the columns of **U** and **V** (or rows of $\mathbf{V}^T$) as well as the elements of **W** are permutated in consistent manner.

To illustrate the usefulness of SVD, let us first consider an (N x N) square matrix **A**. The decomposition of **A** into the matrices **U**, **W** and $\mathbf{V}^T$ is relatively straightforward, partly, because the decomposed matrices are also square. Since **U** and **V** are orthogonal, the inverses of these matrices equal to their transposes, i.e., $\mathbf{U}^{-1} = \mathbf{U}^T$ and $\mathbf{V}^{-1} = \mathbf{V}^T$. As a corollary, an identity matrix **I** is obtained when **U** and **V** are multiplied by their respective transposes even in reverse order, i.e., $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$. The inverse of matrix **W** is also a diagonal matrix whose values are the reciprocals of the elements w_j. Hence, the inverse of **A** can be written as:

$$\mathbf{A}^{-1} = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T \tag{7.4}$$

Although trivially obtained, the solution to the inverse of matrix **A** may encounter a problem when one of the singular values w_j is zero or approaches the machine's floating-point precision. That is when any element of the inverse of matrix **W** is greater than 10^6 for single precision or 10^{12} for double precision, the element $1/w_j$ should be set to zero, not left to infinity.

It is perhaps informative to discuss several important concepts before going to the final stage of obtaining solutions using the SVD technique. First, a matrix **A** is deemed singular when its determinant is zero. In SVD approach to solving a linear system of

equations, i.e., $\mathbf{A}\cdot\mathbf{x} = \mathbf{b}$ where $\mathbf{A}$ is a square matrix (N x N), $\mathbf{x}$ and $\mathbf{b}$ are (N x 1) vectors, the matrix $\mathbf{A}$ is also termed singular when its condition number, the ratio of the maximum to the minimum singular values w_j, is infinite. Then there exists a nullspace $N(\mathbf{A}) = \{\mathbf{x} \mid \mathbf{A}\cdot\mathbf{x} = 0\}$, which maps vector $\mathbf{x}$ to zero. The dimension of the nullspace is called nullity of $\mathbf{A}$, which also quantifies the number of zero diagonal elements (singular values) in the matrix $\mathbf{W}$. If vector $\mathbf{x}$ can be mapped to a space $\mathbf{b}$, which in turn can also be reached by matrix $\mathbf{A}$, there exists a subspace of $\mathbf{b}$ termed the rangespace of $\mathbf{A}$, $R(\mathbf{A}) = \{\mathbf{A}\cdot\mathbf{y}, \forall\mathbf{y}\}$ where $\mathbf{y}$ is usually different from $\mathbf{x}$. The dimension of $R(\mathbf{A})$ is known as the rank of $\mathbf{A}$, whose value is less than N if $\mathbf{A}$ is singular and exactly N otherwise. More precisely, the nullity and rank of $\mathbf{A}$ sum up to N. To further illustrate these concepts, consider a (2 x 2) matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{bmatrix} 5 & 2 \\ 10 & 4 \end{bmatrix}$$

It is easy to determine that its determinant is zero; hence, $\mathbf{A}$ is singular. The nullspace of $\mathbf{A}$ can be found by utilizing the homogeneous system of equations, i.e., $\mathbf{A}\cdot\mathbf{x} = 0$ or simply setting the first row to zero and solving for the first element of vector $\mathbf{x}$. Therefore the nullspace of $\mathbf{A}$, $N(\mathbf{A}) = \lambda[2; -5]$ where $\lambda \in \mathbb{R}$ is a scalar multiplier and semicolon (;) inside the bracket represents a column vector. Note that it is a matter of choice to have a positive first element of $N(\mathbf{A})$ since the vector can also be $\lambda[-2; 5]$, which lies on the same line (nullspace) as the previous vector. Similarly, the rangespace of $\mathbf{A}$ can also be determined by recognizing that the elemental values in the second row are twice as large as those in the first row. Hence the rangespace of $\mathbf{A}$, $R(\mathbf{A}) = \lambda[1; 2]$ where $\lambda \in \mathbb{R}$ is a scalar multiplier and semicolon (;) represents a column vector.

Utilizing the above concepts, an (N x N) matrix $\mathbf{A}$ when decomposed by SVD will produce matrices $\mathbf{U}$, $\mathbf{W}^{-1}$ and $\mathbf{V}^{T}$. The columns of $\mathbf{U}$ whose same-numbered elements w_j are nonzero represent a set of orthonormal vectors in the range of $\mathbf{A}$. Conversely, the columns of $\mathbf{V}$ whose same-numbered elements w_j are zero form a set of orthonormal vectors of the nullspace. In the homogenous case, i.e., $\mathbf{b} = 0$, any vector in the nullspace

or linear combination thereof is a solution. When **b** is a nonzero vector, one must initially determine whether **b** is in the range of **A** or not. If it is in the range of **A**, then the solutions are simply **x** and its linear combination from the nullspace of **A**, i.e., $\mathbf{x} = \mathbf{x}_0$ + [any vector from the nullspace of **A**] where $\mathbf{x}_0 = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T\mathbf{b}$. However, if **b** is not in the range of **A**, the best approximated solution **x** is obtained by minimizing the residual error $|\mathbf{A}\cdot\mathbf{x} - \mathbf{b}|$, i.e., a set of {**x**} that best maps to the desired vector **b**. The set {**x**} that minimizes the residual error, in most cases is much smaller in length compared to the original vector **x**, hence resulting in dimensionality reduction.

The next step is to find solutions for the case of non-square matrix. If the number of rows are less than that of columns ($M < N$), i.e., an under-determined system, the particular solutions may be obtained by augmenting the matrix **A** by ($N - M$) zeros and vector **b** also by ($N - M$) zeros. The final solutions are just like those in the case of square matrix discussed above. On the contrary, if the number of rows are greater than that of columns ($M > N$), i.e., an over-determined system, the solution is simply written as $\mathbf{x} = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T\mathbf{b}$ because **b** is always in the range of **A**. Keeping this discussion in mind, it is perhaps better to summarize the SVD techniques and corresponding solutions in a tabulated form (Table 7.1).

Table 7.1
The formal solutions using SVD techniques.

No.	Case	Solution
1.	$M = N$ (square matrix)	
	(a) $\mathbf{b} = 0$	Any vector in the nullspace of **A**, i.e., columns of **V** whose same-numbered elements w_j are zero.
	(b) $\mathbf{b} \neq 0$ but in R(**A**)	$\mathbf{x} = \mathbf{x}_0$ + [any vector in the nullspace of **A**], where $\mathbf{x}_0 = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T\mathbf{b}$
	(c) $\mathbf{b} \neq 0$ but not in R(**A**)	No exact solution. Best approximated "solution" is $\mathbf{x} : \min\|\mathbf{A}\cdot\mathbf{x} - \mathbf{b}\|$.
2.	$M < N$	Augment matrix **A** and vector **b** by ($N - M$) rows of zeros. The solutions are analogous to case (1).
3.	$M > N$	$\mathbf{x} = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T\mathbf{b}$

7.1.2 Results and Discussion

Environmental and meteorological data sets are generally massive, and consist of observations taken at continuous intervals. The frequency of the data and its redundancy render the resultant correlation or covariance matrix strongly singular. The solution of equation involving such strongly singular matrices is possible using SVD, which can result in reduction in dimensionality and thereby facilitating solution. To illustrate how this is achieved, consider the solution of the matrix system $\mathbf{A}\cdot\mathbf{x} = \mathbf{b}$ where matrix **A** (365 x N) contains the values of the cosine and sine terms in a Fourier series; vector **b** (365 x 1) consists of daily average ozone values $Z_\alpha(t_j)$, $j = 1, \ldots, 365$ and α refers to the year (1) 1997, (2) 1998, (3) 1999, and (4) 2000. The solution is obtained by constraining vector **b** to within the range of matrix **A** and mapping it to an (N x 1) solution-vector **x**, from which the values of temporal trend coefficients $a_{n\alpha}$ and $b_{n\alpha}$, $n = 1, \ldots, N$ and $\alpha = 1, \ldots, 4$, are determined.

It has been shown earlier that the residual error $|\mathbf{A}\cdot\mathbf{x} - \mathbf{b}|$ is minimum when the number of terms retained in the Fourier series is large, i.e., the dimension N of the matrix **A** (M x N) approaches infinity. However, it is possible that matrix **A** is singular, i.e., some of the Fourier coefficients are redundant given the temporal correlations in ozone profiles. So, the first task is to decompose matrix **A** using the SVD technique, i.e., $\mathbf{A} = \mathbf{UWV}^T$, where **U** is a column-orthonormal matrix (365 x N), **W** is a diagonal matrix (N x N) whose nonzero elements are the singular values, and $\mathbf{V}^T$ is a row-orthonormal square matrix (N x N). The final step is easily accomplished by taking the inverse of the decomposed matrix **A** and multiplying (matrix operation) it with vector **b**, as summarized in Table 7.1 (case 3).

Figure 7.1 shows the results from the application of Fourier series analysis (coupled with SVD) for daily-average ozone values $Z_\alpha(t_j)$, $\alpha = 1, \ldots, 4$, $j = 1, \ldots, 365$; here only those for 1997 are shown for example. The $Z_\alpha(t_j)$ values for that year may be reproduced by retaining 320 (160 x 2) temporal trend coefficients $a_{n\alpha}$ and $b_{n\alpha}$, which is ~56% reduction from the original 730 values corresponding to 365 data. Note that $a_{o\alpha}$ is

merely a constant and therefore left outside of the summation. On the other hand, $b_{o\alpha}$ is often dropped from the Fourier series because the first sine term is zero.

The importance of Fourier series analysis is perhaps more obvious in the study of ozone seasonal trends (Figures 7.2 to 7.5). First, the yearly data is arbitrarily divided into four seasons, i.e., winter (Jday: 1-90), spring (Jday: 91-180), summer (Jday: 181-270), and fall (Jday: 271-365) where Jday denotes the date formatted in Julian day. Here only 8 (4 x 2) Fourier coefficients are needed to represent each seasonal trend as opposed to about 40 (20 x 2) when yearly data are used (Figure 7.1a). To illustrate an example, the seasonal trend in winter 1997 is obtained from:

$$\hat{Z}_1(t) = a_{01} + a_{11}\cos\left(\frac{2\pi t}{M}\right) + a_{21}\cos\left(\frac{4\pi t}{M}\right) + a_{31}\cos\left(\frac{6\pi t}{M}\right) + a_{41}\cos\left(\frac{8\pi t}{M}\right)$$

$$+ b_{11}\sin\left(\frac{2\pi t}{M}\right) + b_{21}\sin\left(\frac{4\pi t}{M}\right) + b_{31}\sin\left(\frac{6\pi t}{M}\right) + b_{41}\sin\left(\frac{8\pi t}{M}\right) \qquad (7.5)$$

where a_{o1} = 1; a_{n1} and b_{n1} values for 1997 are tabulated in Table 7.2 and for the rest, in Table 7.3 to 7.5; t refers to Julian day from 1 to 90; and M = 90 for the winter season of 1997. From Table 7.2 to 7.5, it is confirmed that the values of $a_{o\alpha}$ are always one whereas those of the other coefficients $a_{n\alpha}$ and $b_{n\alpha}$ vary, i.e., could be either positive or negative, depending on the yearly or, in this case, seasonal trend. It is also observed that the first two Fourier coefficients correlate for the corresponding seasons from year to year. This is intuitively expected since the first two coefficients correspond to the low frequency (or large temporal range) structure, and all the annual trends exhibit similar temporal profiles within corresponding seasons. However, the Fourier coefficients corresponding to the higher frequency structures are difficult to correlate.

In the case of the thirty-day moving average 30dMA values $W_\alpha(t_j)$:

$$W_\alpha(t_j) = \frac{1}{2k}\sum_{i=j-k+1}^{j+k} Z_\alpha(t_i)$$

$$\alpha = 1, \ldots, 4; j = 15, \ldots, 365; k = 15$$

where α refers to the yearly index as defined above and $2k$ denotes the size of the moving window. As shown in Figure 7.6, the number of coefficients $a_{n\alpha}$ and $b_{n\alpha}$ required for data reproduction using FSA (coupled with SVD) is only 8 (4 x 2), compared with 351 values (15 to 365) from the original 30dMA values $W_\alpha(t_j)$. In other words, the FSA (coupled with SVD) requires only 8 Fourier coefficients to be stored, upon which the temporal ozone trends can be reconstituted. Data compression of approximately 97% is thus achieved. Of course, the residual components (higher frequency structures) may be indicative of sudden shifts in environmental parameters, and it may be necessary to reproduce some of these residual components. The principles of stochastic simulation presented in Chapter 6 provide an avenue for such purpose. The reduced data set of 8 coefficients of the Fourier series, i.e., sine and cosine terms, can also be used in conjunction with the basis functions for stochastic interpolation methods such as universal kriging (e.g., Huijbregts and Matheron, 1971).

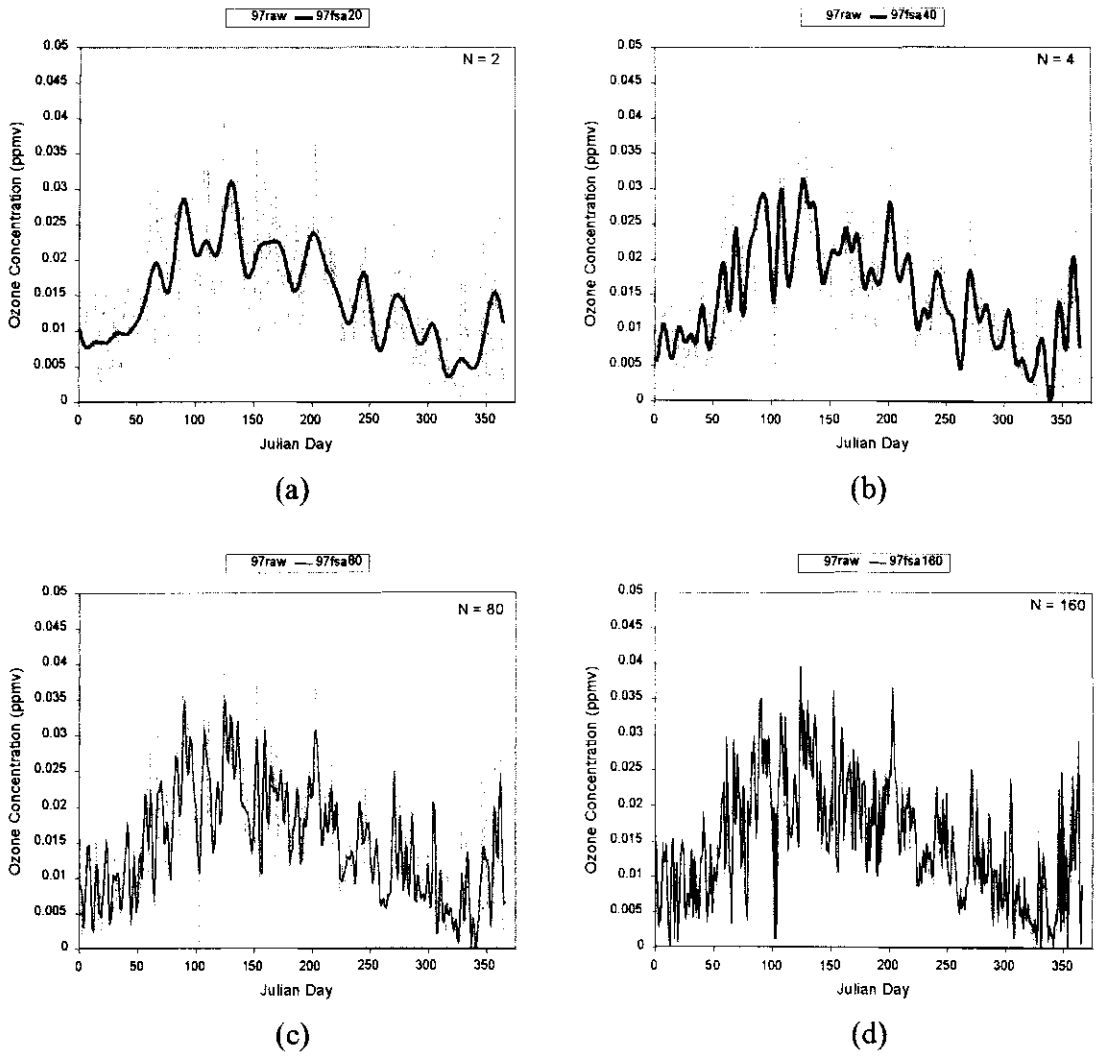
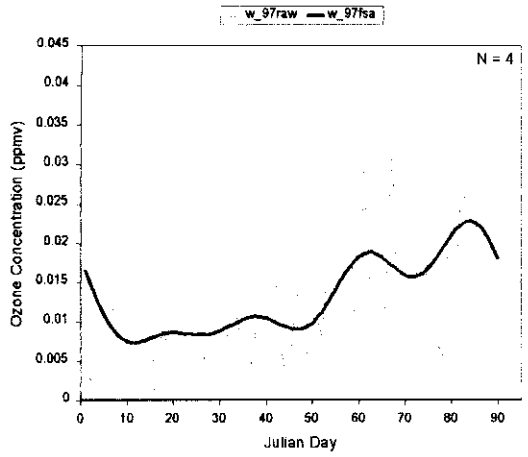


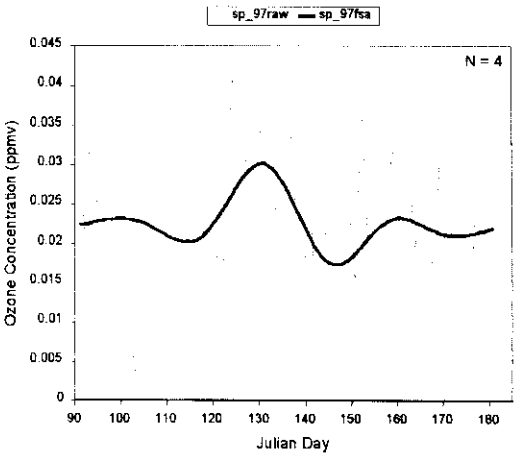
Figure 7.1
Results of Fourier series analysis (FSA) coupled with SVD for 1997 ozone daily values. Different numbers of coefficients N are tested to visualize the oscillations: (a) N=20, (b) N=40, (c) N=80, and (d) N=160. Note that an excellent match is obtained with N=160 Fourier coefficients. FSA alone will result in data identification as N approaches infinity.

Table 7.2
The values of Fourier coefficients for 1997 seasonal ozone data.

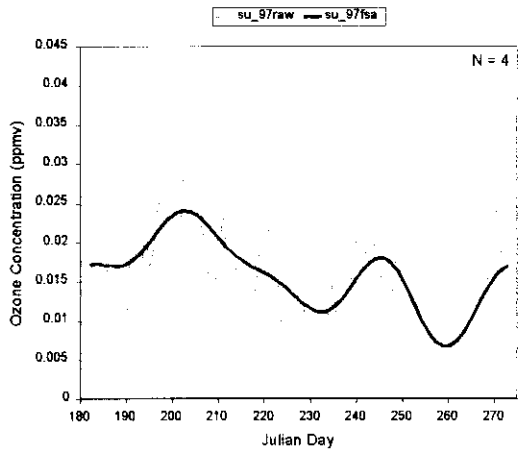
Winter		Spring		Summer		Fall	
a	**b**	**a**	**b**	**a**	**b**	**a**	**b**
1.0000	-0.0053	1.0000	0.0014	1.0000	0.0047	1.0000	0.0003
0.0025	-0.0018	-0.0010	-0.0021	0.0009	0.0027	0.0044	-0.0016
0.0004	-0.0011	0.0009	0.0029	-0.0014	-0.0008	0.0003	-0.0011
0.0019	-0.0020	-0.0001	-0.0010	0.0020	-0.0021	0.0003	0.0014
0.0010		0.0003		-0.0001		-0.0007	



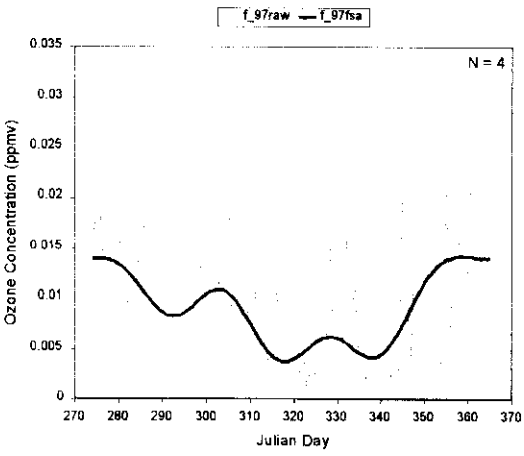
(a) Winter



(b) Spring



(c) Summer

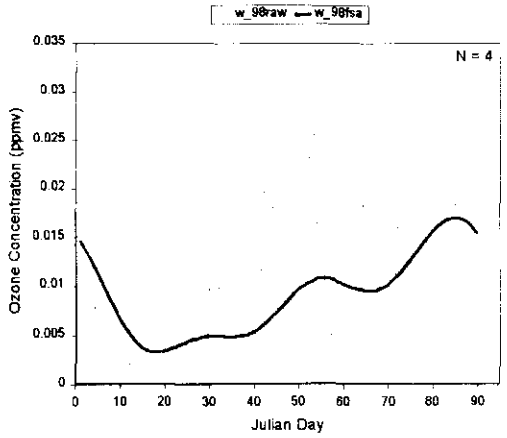


(d) Fall

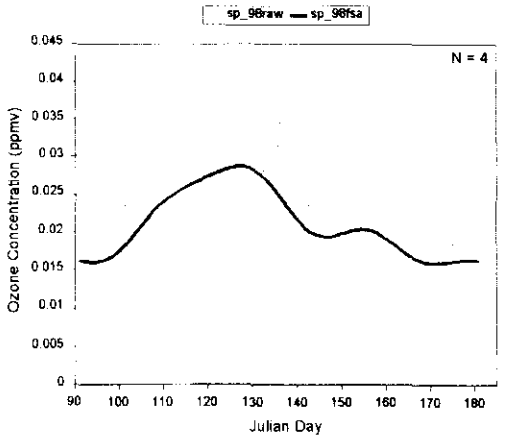
Figure 7.2
Fourier series analysis (FSA) for 1997 seasonal ozone data.

Table 7.3
The values of Fourier coefficients for 1998 seasonal ozone data.

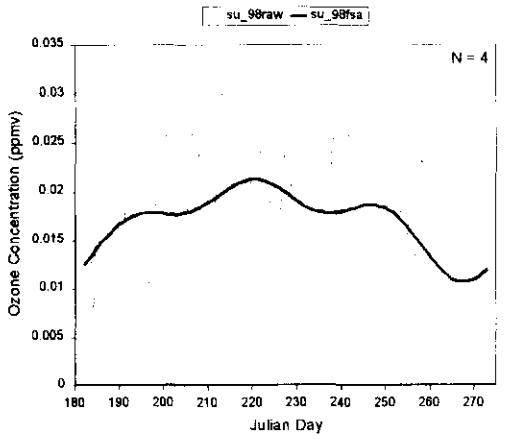
Winter		Spring		Summer		Fall	
a	**b**	**a**	**b**	**a**	**b**	**a**	**b**
1.0000	-0.0041	0.9971	0.0041	1.0000	0.0019	1.0000	-0.0019
0.0029	-0.0009	-0.0041	-0.0013	-0.0032	0.0009	0.0011	0.0001
0.0023	-0.0013	-0.0014	0.0004	-0.0007	0.0015	-0.0016	-0.0009
0.0011	0.0002	0.0005	-0.0009	0.0004	0.0000	-0.0016	0.0010
0.0009		-0.0006		0.0000		0.0008	



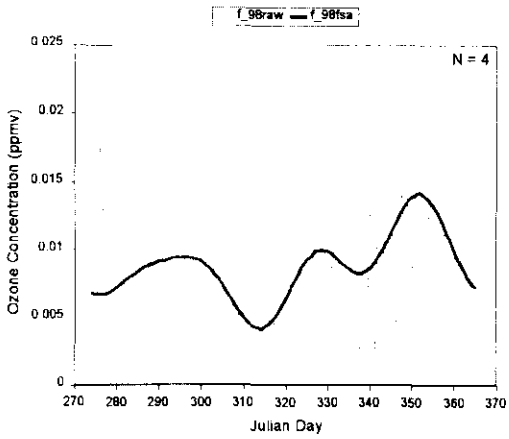
(a) Winter



(c) Spring



(c) Summer



(d) Fall

Figure 7.3
Fourier series analysis (FSA) for 1998 seasonal ozone data.

Table 7.4
The values of Fourier coefficients for 1999 seasonal ozone data.

Winter		Spring		Summer		Fall	
a	b	a	b	a	b	a	b
1.0000	-0.0019	0.9982	0.0023	1.0000	0.0019	1.0000	0.0022
0.0007	-0.0037	-0.0018	-0.0002	0.0004	0.0021	0.0028	-0.0003
0.0020	-0.0029	-0.0004	0.0022	0.0006	-0.0011	-0.0008	0.0005
0.0017	-0.0017	0.0007	0.0031	0.0001	0.0020	0.0000	-0.0001
-0.0014		-0.0027		0.0000		-0.0005	

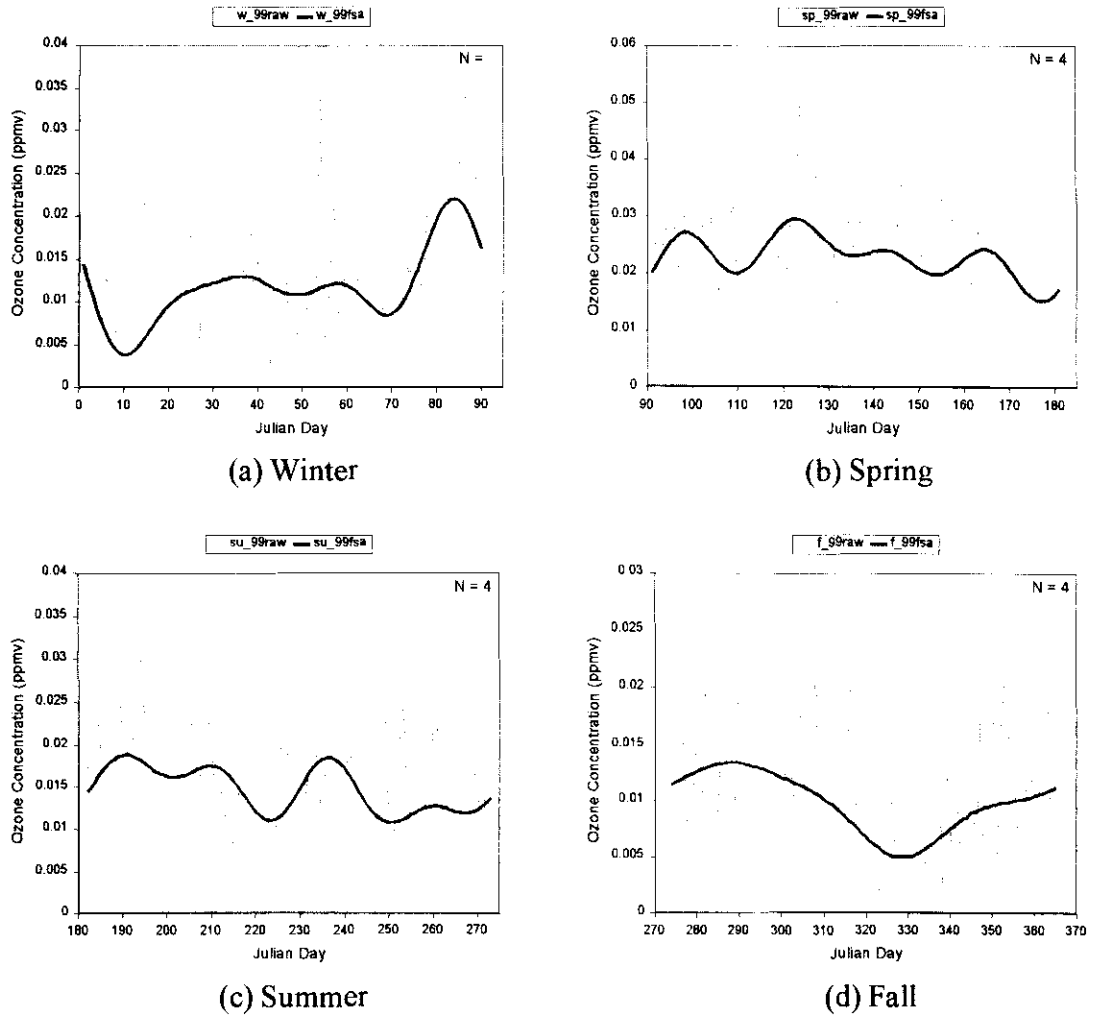
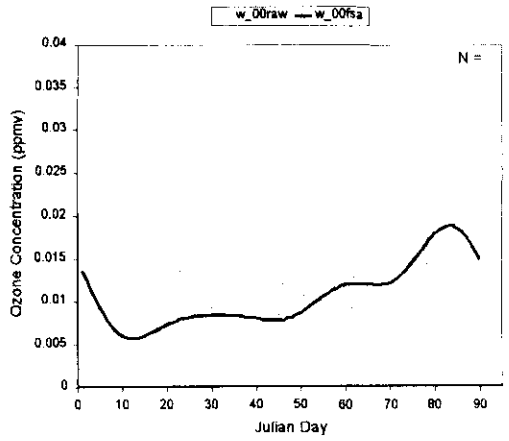


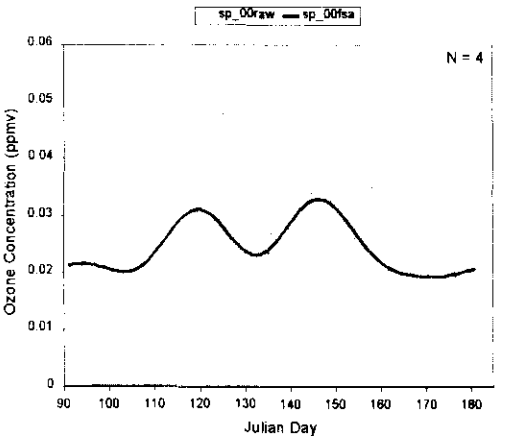
Figure 7.4
Fourier series analysis (FSA) for 1999 seasonal ozone data.

Table 7.5
The values of Fourier coefficients for 2000 seasonal ozone data.

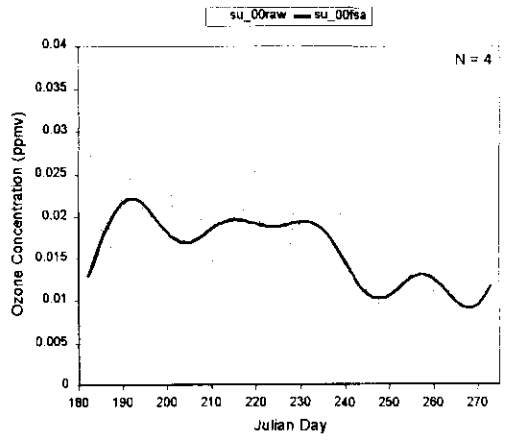
Winter		Spring		Summer		Fall	
a	**b**	**a**	**b**	**a**	**b**	**a**	**b**
1.0000	-0.0035	0.9986	0.0011	1.0000	0.0044	1.0000	0.0004
0.0024	-0.0022	-0.0046	0.0015	-0.0012	0.0012	0.0009	0.0015
0.0007	-0.0015	-0.0005	-0.0026	0.0014	0.0016	0.0004	0.0003
0.0011	-0.0008	0.0021	0.0015	-0.0007	0.0020	0.0003	0.0010
0.0003		0.0000		0.0019		0.0005	



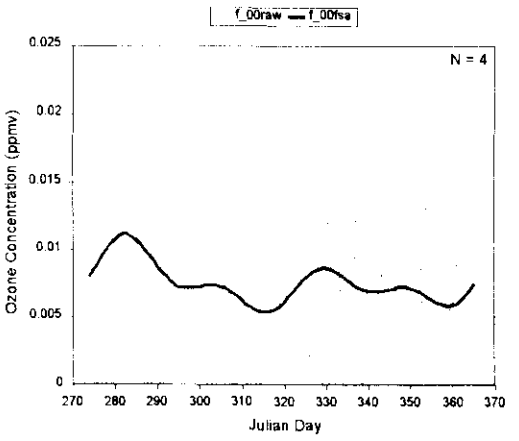
(a) Winter



(b) Spring

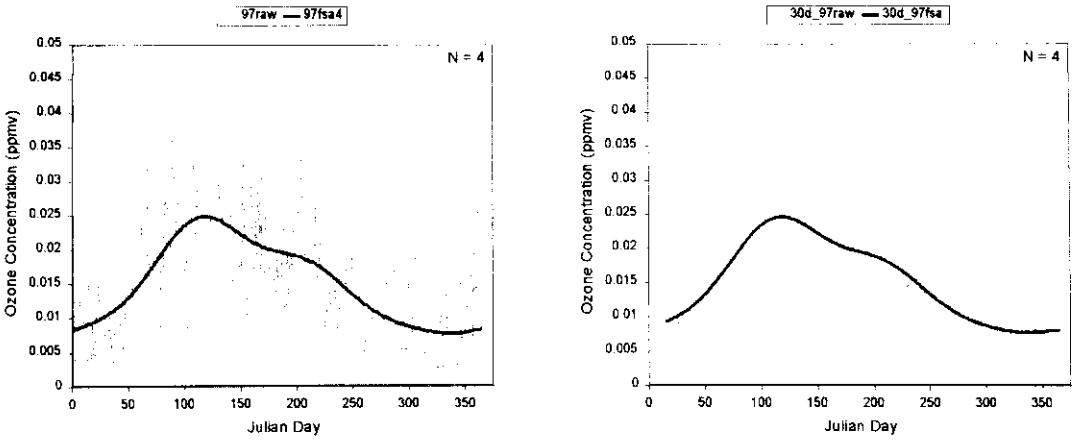


(c) Summer

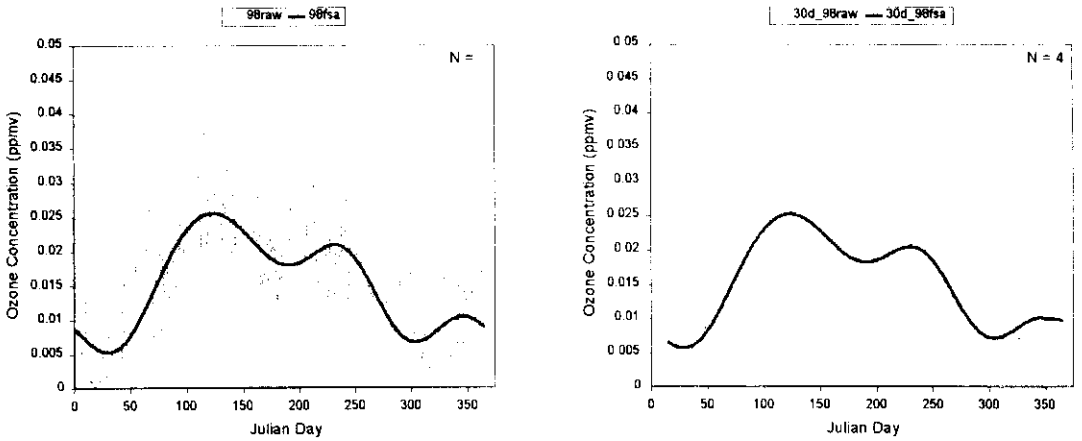


(d) Fall

Figure 7.5
Fourier series analysis (FSA) for 2000 seasonal ozone data.

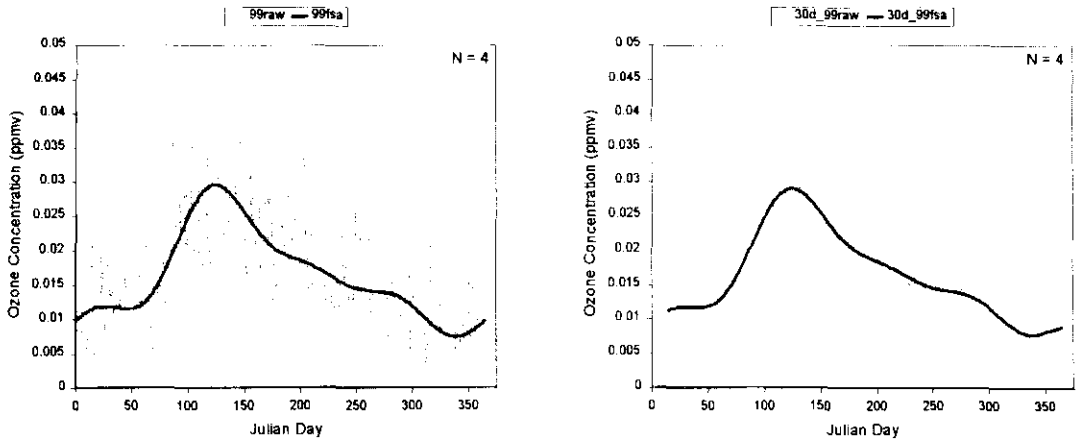


(a) 1997

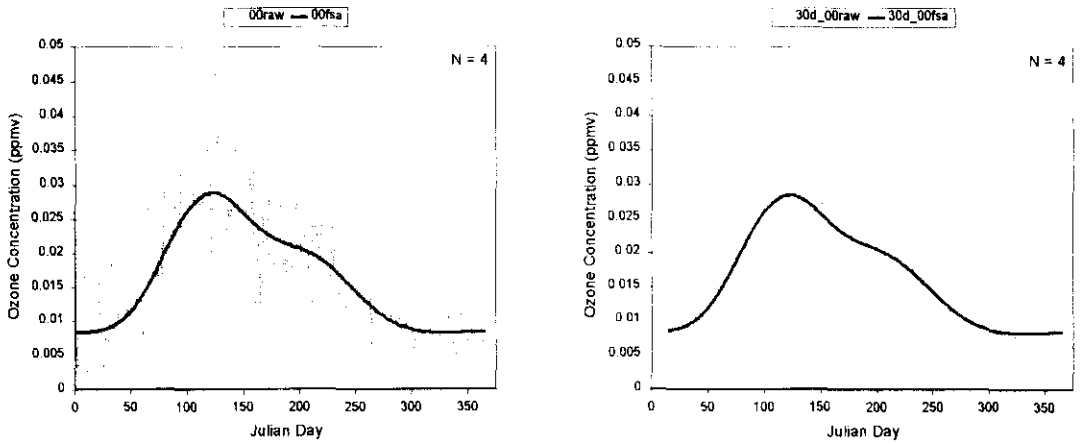


(b) 1998

Figure 7.6
Results of Fourier series analysis (FSA) coupled with SVD for 1997 and 1998. The resulting fits (thicker blue lines) are superimposed on raw daily ozone data [**LEFT**] and 30-day moving average ozone data [**RIGHT**].



(c) 1999



(d) 2000

Figure 7.6
Results of Fourier series analysis (FSA) coupled with SVD for 1999 and 2000. The resulting fits (thicker blue lines) are superimposed on raw daily ozone data [**LEFT**] and 30-day moving average ozone data [**RIGHT**].

7.2 Neural Network

Environmental pollution processes often involve highly nonlinear interactions between several meteorological and chemical variables. In Calgary, Alberta, nine daily average covariates, i.e., dust and smoke (COH), carbon monoxide (CO), nitric oxide (NO), nitrogen dioxide (NO2), total hydrocarbon (THC), wind speed (WSPD), temperature (Tavg), relative humidity (RHavg) and bright sunshine hours (bSUN), have been determined to be the key factors affecting ozone formation. Physicochemical models that are generally used for predicting ozone concentrations accounting for complex covariate interaction are cumbersome and computationally expensive. Most models can be utilized to predict ozone variations only within extremely short simulation intervals (around 2-5 days).

Neural networks (e.g., Bishops, 1995) offer a relatively straightforward approach to substitute for the complex physicochemical models, and can be applied for predicting ozone levels subject to nonlinear interaction between covariates. The most commonly implemented neural network in the study of atmospheric pollution is the multilayer perceptron (MLP) due to its ability to make accurate generalization when presented with new sets of input data. In contrast to other statistical approaches, MLP makes no prior assumption regarding the data distribution and can be trained to approximate any "smooth" pattern underlying the temporal process (Hornik et al., 1989). The multilayer perceptron shown in Figure 7.7 comprises a system of inter-connected neurons, which map the selected inputs to an output through multiple nodes in the hidden layer. Each value a_j at the hidden nodes is a linear combination of the inputs or covariates $X^{(m)}$ weighted with ω_{ij}, and can simply be written as follows:

$$a_j = \sum_{i=1}^{N} \omega_{ij} X^{(m)}(t_i), \; j = 1, \ldots, J; \; m = 1, \ldots, M \tag{7.6}$$

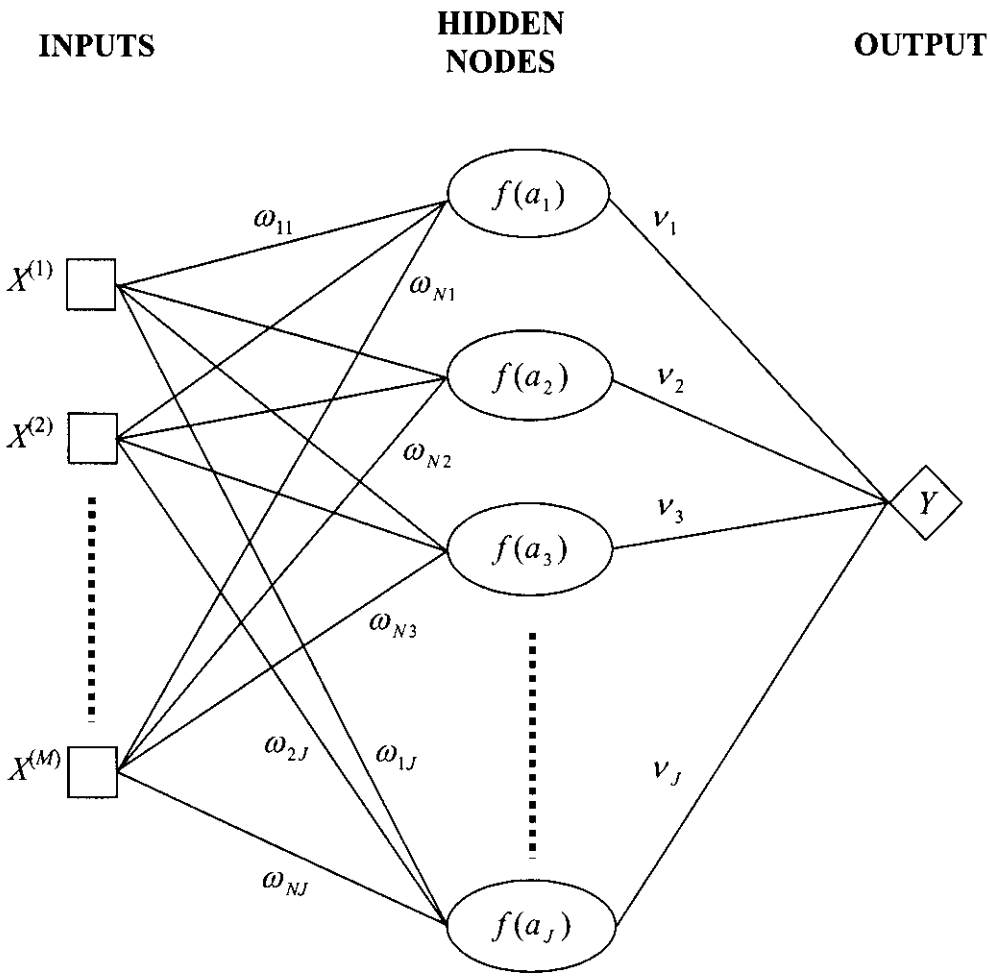


Figure 7.7
Schematic of neural network architecture (*M*:*J*:1). The covariates (meteorological and chemical variables) are mapped into a target output (ozone) through a single hidden layer neural network.

Here, a set corresponding to the i^{th} time instant is fed contemporaneously to a series of J hidden nodes. To add nonlinearity to the process, an activation (transfer) function $f(\cdot)$ is then applied to a_j:

$$h_j = f(a_j) \tag{7.7}$$

In theory, any nonlinear function may be used but it is more practical to apply an activation function, which is bounded and easy to compute. The boundedness of this function is preferred to avoid dealing with large weights, which may result in slow convergence during the training mode. For example, the hyperbolic tangent (tanh) function where λ replaces a_j is given as:

$$f(\lambda) = \frac{e^{\lambda} - e^{-\lambda}}{e^{\lambda} + e^{-\lambda}} \tag{7.8}$$

The above function has two distinctly nice features; it is bounded at [−1, 1], and its derivative is also easily obtained as:

$$f'(\lambda) = 1 - \left(\frac{e^{\lambda} - e^{-\lambda}}{e^{\lambda} + e^{-\lambda}}\right)^2 = 1 - [f(\lambda)]^2 \tag{7.9}$$

This relatively simple form of derivative is essential during the back propagation step where a gradient-based optimization technique is generally applied. The intermediate values h_j at the hidden nodes are linearly combined to yield:

$$\tilde{Y} = \sum_{j=1}^{J} v_j h_j \tag{7.10}$$

where v_j are the weights associated with the corresponding hidden-node values. To add further nonlinearity to the network system and thus increase accuracy of prediction, another activation function $g(\cdot)$ may be applied to the output $\tilde{Y}$. Again, there are no theoretical limitations imposed on this function. The function $g(\cdot)$ may even differ from

$f(\cdot)$ if the situation warrants but in general, the activation function should be a nonlinear expression with a simple derivative. The final output values are given as:

$$Y^{(l)} = g\left(\tilde{Y}\right) \tag{7.11a}$$

or when written in the complete form:

$$Y^{(l)}(t_i) = g\left(\sum_{j=1}^{J} v_j \cdot f\left[\sum_{i=1}^{N} \omega_{ij} X^{(m)}(t_i)\right]\right) \tag{7.11b}$$

$$\forall m = 1, \ldots, M$$

where superscript (l) refers to the result after l^{th} iteration step. This predicted output $Y^{(l)}$ is compared against the target output corresponding to a training set Y. The norm of the residual error $\| Y - Y^{(l)} \|$ is computed, and the iterations l are repeated until the error is minimized. This minimization is usually performed using a gradient-based optimization approach, or more precisely by back propagating the error through the networks. The general back propagation procedure is summarized below following Gardner and Dorling (1998):

1. Set the network weights ω_{ij} and v_j to small random values.
2. Compute $Y^{(l)}$ corresponding to the assumed weights.
3. Calculate the residual error, usually in the form of sum of squared error (SSE), by comparing the predicted outputs to the target values.
4. Adjust the weights using the computed error-gradient.

Steps 2-4 are repeated for the subsequent inputs until the overall error is sufficiently small. The resultant weights are applied to an independent data set to test for the generalization performance. The generalization step aids in detecting any "overfitting" problem, which usually occurs when dealing with noisy data as in the case of ozone phenomena. An overfitted network will yield very low SSE corresponding to the training

set but high SSE for the generalization set. Finally, the performance of the network is validated using an independent (validation) data set.

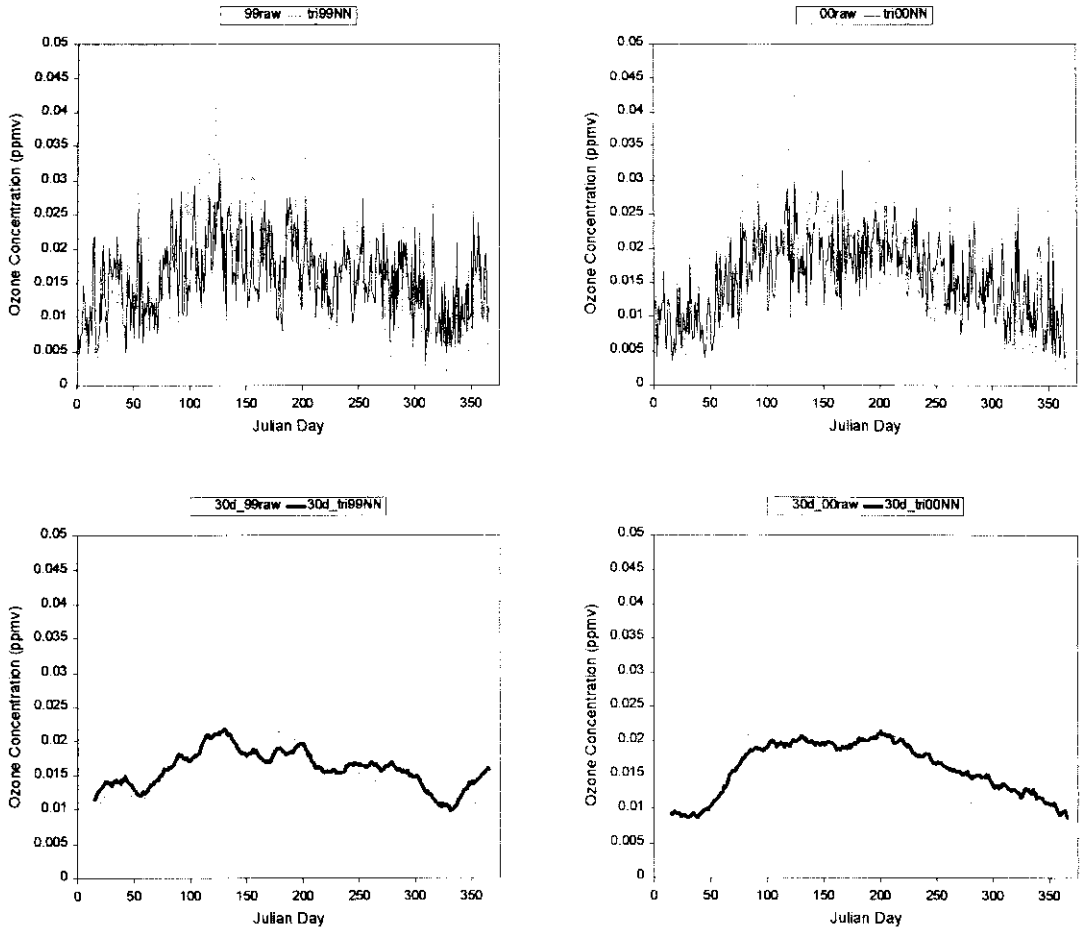
7.2.1 Results and Discussion

As a preliminary test of the neural network performance, three input variables (covariates), i.e., wind speed (WSPD), average temperature (Tavg) and bright sunshine hours (bSUN), are used for predicting ozone in 1999 and 2000. These covariates are chosen in order to obtain fair comparison between the previous regression results (Section 4.2) and those of the neural network. The learning data set is initially divided into a training set (1997) and a generalization set (1998). In order to obtain robust prediction using the neural network, the training must be performed using a representative data set. The complete annual profile for 1997 should be used for training, since if the data are randomly sampled, the training set may consist of extreme values and thus result in biased prediction. Figure 7.8 shows the daily predicted outputs compared to the raw data [**top**] as well as the 30-day moving averages (30dMA) of the predicted ozone values in 1999 and 2000 superimposed on those of the actual data [**bottom**]. The neural network manages to capture the dip occurring in the late fall of 1999 and also the second peak season in 2000. This is because the training and generalization data sets of 1997 and 1998, respectively, reflect these peaks and dips. However, the high ozone phenomena occurring in the late spring of all four years are unsuccessfully emulated because none of these covariates peak at the same period as ozone does.

The choice of covariates is important because ozone is formed through complex, nonlinear processes. The interactions between the nine variables, as previously tabulated, may influence the high and low values of ozone. For this reason, all nine covariates are then combined as the inputs to the neural network. The learning data set is divided such that the 1997 data are used for training the network and 1998 data for generalization. The 30dMA results of ozone predictions in 1999 and 2000 are plotted against those of the original values (Figure 7.9). Similar trends can be observed as in the case of three covariates above. The visible difference is that the predicted ozone trend in 1999 for the nine inputs is relatively smoother than that of three inputs. The magnitude of ozone peak

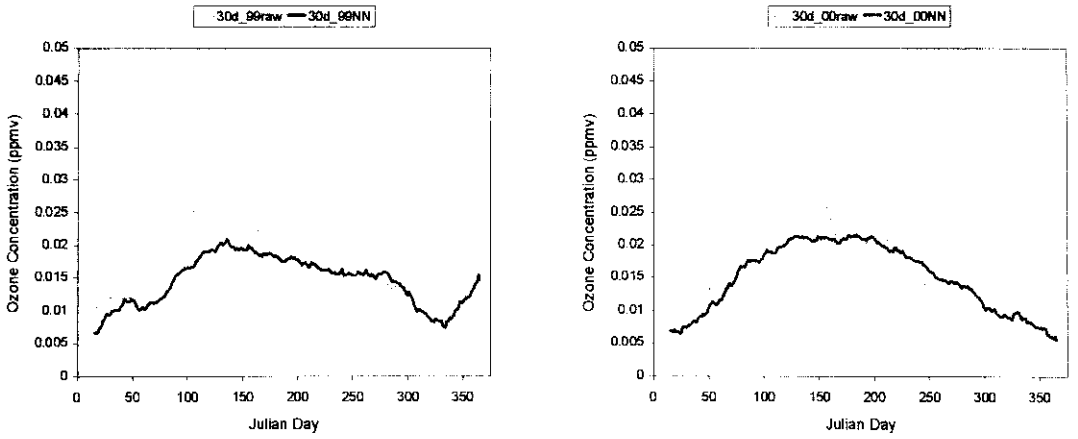
is underestimated for both years as a result of the compromise made in order to fit the valleys and peaks in the training data set.

The performance of the neural network is affected by the choice of the training data set. In order to study this influence, the training data set is switched from 1997 to 1998, and vice-versa for the generalization data set. The 30dMA of the results for ozone predictions in 1999 and 2000 are superimposed on those of the actual values in Figure 7.10. Notice the occurrence of another peak in the third quarter of 1999. This phenomenon is most likely due to the second peak of the temporal ozone trend in the training data set (1998). Hence it is obvious that the right choice of the training data set is an important measure of how accurate the neural network results will be. In essence, the neural network is a viable prediction tool in the presence of covariate information. Prediction using a neural network is likely to be improved by increasing the duration of the training data. A properly calibrated neural network can thus act as a surrogate for the physicochemical models provided the combinations of covariate and ozone phenomena have been exhaustively sampled. However, the single hidden-layer neural network as discussed in this chapter cannot capture the complex, nonlinear relationships that exist among the input variables. To rectify this shortcoming, we can resort to several hidden-layer networks, or employing a multipoint statistic calibrated using a probabilistic neural network. The latter approach will be proposed as an avenue of future research in the next chapter.



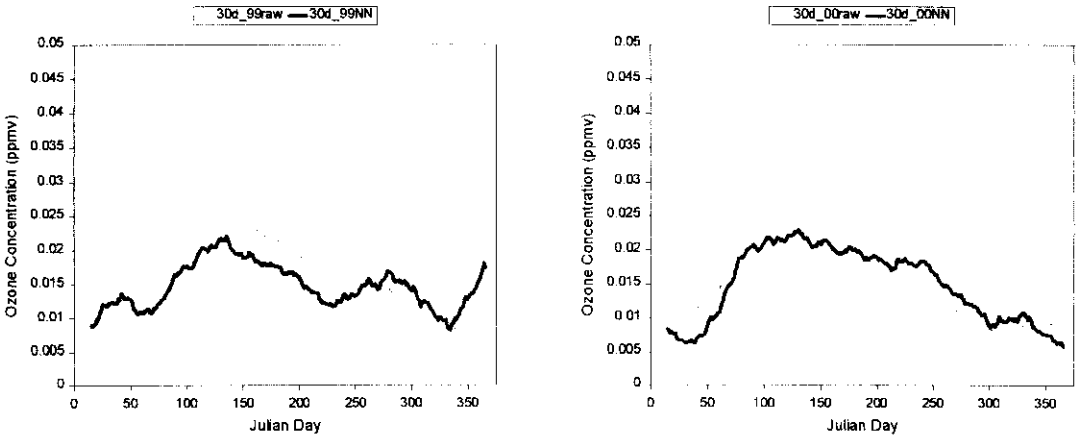
(a) 1999 (b) 2000

Figure 7.8
The daily average results of neural network predictions (blue) and those of actual ozone values (gray) in 1999 and 2000 using three covariates (inputs): WSPD, Tavg, bSUN **[top]**. The 30-day moving averages (30dMA) of the neural network results and actual values are also shown **[bottom]**. The 1997 data sets are used for training and 1998 for generalization.



(a) 1999 (b) 2000

Figure 7.9
The 30-day moving averages (30dMA) of the neural network predictions (blue) in 1999 and 2000 using nine covariates (inputs): COH, CO, NO, NO2, THC, WSPD, Tavg. RHavg, bSUN, and the corresponding actual ozone values (gray). The 1997 data sets are used for training and 1998 for generalization.



(a) 1999 (b) 2000

Figure 7.10
The 30-day moving averages (30dMA) of the neural network predictions (blue) in 1999 and 2000 using nine covariates (inputs): COH, CO, NO, NO2, THC, WSPD, Tavg. RHavg, bSUN, and the corresponding actual ozone values (gray). The 1998 data sets are used for training and 1997 for generalization.

CHAPTER 8

CONCLUSIONS AND FUTURE RESEARCH AVENUES

Space-time modeling of atmospheric pollutants has been actively attempted by many workers, including Kyriakidis (1999) who successfully integrated the deterministic trend $m(t)$ and the probabilistic residual $R(t)$ components of a random variable RV $Z(t)$ through a stochastic simulation approach. However, many of the spatiotemporal studies carried a major assumption that the temporal aspect is fully understood and thus focused primarily on spatial modeling. The main objective of this thesis is concerned with evaluating the accuracy and suitability of the techniques used for modeling the temporal phenomena. For this reason, various statistical methodologies, e.g., linear regression, kriging and stochastic simulation, were performed in the case of predicting tropospheric ozone concentrations in Calgary, Alberta for 1998-2000. The general conclusions of this study are highlighted below and new research avenues are then recommended as part of future works.

8.1 General Conclusions

° The formation of ozone via photochemical reaction is complex and highly nonlinear. Hence to describe the temporal phenomena using linear models is inappropriate as illustrated by the results of linear regression in Chapter 4. Here only the positively correlated variables, in bivariate sense, such as wind speed (WSPD), average temperature (Tavg) and bright sunshine hours (bSUN) were applied for predicting ozone concentrations in 1998-2000. This way, the "reducing effect" of the negatively correlated variables, i.e., the amount of dust and smoke (COH), the concentrations of carbon monoxide (CO), nitric oxide (NO), nitrogen dioxide (NO2), total hydrocarbon (THC) and average relative humidity (RHavg), could be circumvented. The linear

regression approach can be best implemented if the relationships between predictor variables and ozone were known *a priori*, either from historical/current records or equivalent scenarios. In addition, complete predictor data in the corresponding years must be available.

- The redundancy between data, not accounted for in the above approach, was captured via the kriging paradigm. The prediction proceeded through the use of a variogram, a "two-point" statistic. Once the variogram of the previous year was determined to be adequately representative for modeling the temporal variations in future years, ozone predictions could be performed based on "a few" sample data in the corresponding years. However, kriging suffered two shortcomings; the data-to-unknown covariance was identified but the unknown-to-unknown covariance was not, resulting in poor reproduction of the temporal ozone concentrations, and also the estimation variance was lower than the true value resulting in the smooth kriged profiles. Nevertheless, the exactitude of kriging at the data locations, as shown by the spikes, renders it valuable as the basis for stochastic simulation.

- The reduction in kriging variance was corrected by employing stochastic simulation. Here a temporal residual component $R(t)$ was added to the kriged estimate $Z_K^*(t)$. To ensure unbiasedness, the residual was assumed to satisfy certain criteria; its mean must vanish, its variance carried the value of kriging estimation variance σ_K^2, and it must be orthogonal to the kriged estimate. The simulated results reflect the temporal fluctuations of ozone profiles, as they should be. Based on the evenly spaced data at every 30^{th} Julian day of the respective years, the temporal ozone trends were only reproducible in an ergodic sense (average over several realizations). The uncertainty induced by the data sampling procedure was investigated by selecting samples randomly between the 25^{th} and 30^{th} Julian day of the month. To account for periodicity in the long-term temporal variations, the hole-effect model was implemented in the case of one realization, and the results represented the average trends over the four-year period of ozone study.

- The results from physicochemical modeling (e.g., Roelofs and Lelieveld, 2000) and experimentation (e.g., Campbell, 1986) suggested that air pollutants such as THC and NO were directly responsible for ozone formation/destruction and thus should be incorporated in statistical modeling. Following this observation, cokriging and co-simulation were performed using twelve evenly spaced data points at every 30^{th} Julian day of the respective years. The integration of the secondary information improved the prediction accuracy in the sense that the peak magnitudes of ozone trends for 1997, 1998 and 2000 were correctly placed. Fewer fluctuations in the temporal trends could also be observed suggesting that these covariates "added value" to simulation. Poor performance observed in 1999 was due to the influence of THC, whose annual trend peaked much later than that of ozone.

- The performance of linear regression (LR), ordinary kriging (OK) and sequential Gaussian simulation (SGSIM) can be assessed by comparing the correlation coefficients $\rho_{\alpha o}$ between the 30-day moving averages (30dMA) of actual values and results obtained from various cases of random data sampling and multiple realizations (Table 8.1). Recall that the primary goal is to predict future ozone concentrations using sparse data; hence there is likely to be uncertainty associated with those values. That uncertainty may be due to: (1) the variance of the samples themselves, which are unrepresentative of the annual trend, and (2) the sparse information itself, i.e., the complete 365 values were predicted using only twelve data points. Thus the smallest variation in $\rho_{\alpha o}$, i.e., [0.96, 0.98] obtained from the OK results for 2000, indicates that data sampling for that particular year is really not an issue, and that reliable prediction can be obtained by constraining the model to the appropriate variogram. Conversely, the highest variation in $\rho_{\alpha o}$, i.e., [0.25, 0.94] also obtained from the OK results but now for 1997, represents large uncertainty in the predicted ozone values attributed to the variance of the data. The uncertainty in the simulated result due to the sparse information can be investigated if several realizations are imposed on the modeling using the same data configuration; for example, the SRND result for 1999 corres-

ponding to $\rho_{\alpha o}$ = 0.77 could further yield a range of $\rho_{\alpha o} \in$ [0.50, 0.79] after ten simulated realizations were performed (SR10).

Table 8.1
The variations of correlation coefficients $\rho_{\alpha o}$ between the 30dMA of actual values and predicted results over various cases.

Cases	1997	1998	1999	2000
LR	[0.45, 0.89]	[0.57, 0.96]	[0.85, 0.94]	[0.94, 0.97]
OK	[0.25, 0.94]	[0.72, 0.94]	[0.66, 0.93]	[0.96, 0.98]
SR10	[0.56, 0.83]	[0.66, 0.83]	[0.50, 0.79]	[0.67, 0.86]
SRND	[0.38, 0.84]	[0.42, 0.82]	[0.77, 0.88]	[0.75, 0.89]

Note: LR = linear regression (random data sampling); OK = ordinary kriging (random data sampling); SR10 = sequential Gaussian simulation (SGSIM, 10 realizations); SRND = SGSIM (random data sampling).

- The $\rho_{\alpha o}$ tabulated above represent the prediction accuracy obtained by comparing one predicted value and the corresponding "true" value at the same instant in time t_i. A better comparison of model performance can be achieved by employing a variogram, a measure of joint variability of two temporal values contemporaneously. The sample variograms of the predicted results for 2000 were compared with that of the 1997 variogram, used as the basis for inference. From Figure 8.1, the variograms for the standardized results of linear regression not only failed to emulate the short-range structure (5 days) of the original 1997 variogram but also resulted in a higher range (~70 days), smooth trends and Gaussian behavior near the origin as opposed to the exponential shape in the original variogram (1997). The variograms corresponding to kriging begin to bracket the "true" temporal pattern; however, the kriged values were unable to reproduce the short-range structure and gave a higher range (~50 days), implying less variability in the predicted temporal profiles. The drawbacks of linear regression and kriging were significantly alleviated by SGSIM. Here the short-range structure was reproduced by both cases of ten realizations and randomly sampled data, which indicate that the correct patterns of temporal variability were successfully restored.

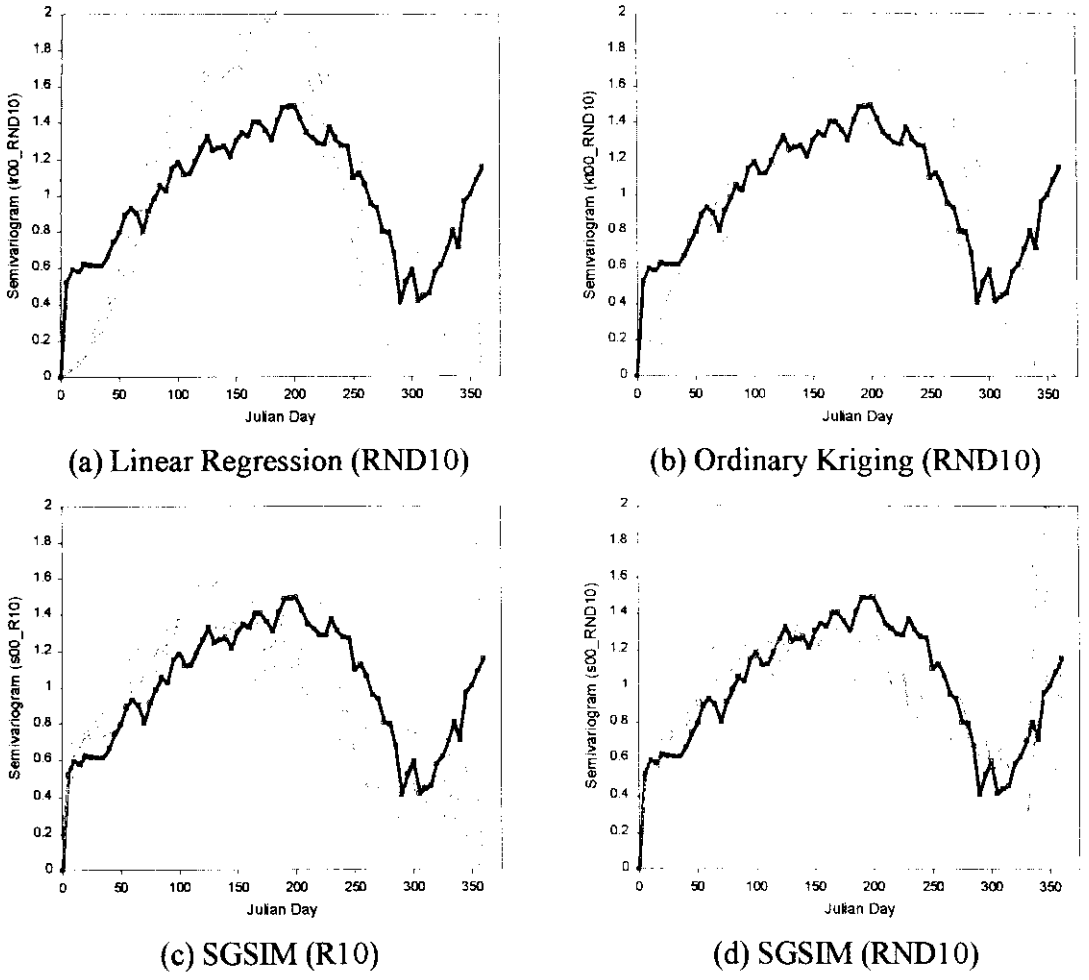


Figure 8.1
The reproduction of 1997 sample semivariogram (thick blue line with open rectangles) on the predicted results (thin gray lines) for different cases in 2000. Note: SGSIM = sequential Gaussian simulation; RND10 = 10 randomly selected data sets; R10 = 10 realizations.

8.2 Future Research Avenues

While this thesis work has presented a number of promising statistical techniques for predicting ozone profiles accurately, some important issues remain to be addressed. In the perspective of monitoring atmospheric pollutants, regulatory agencies such as the Alberta Environment are interested in determining the effects of certain policies on air pollution abatement. For example, can a stricter regulation imposed on the precursors (NO_x and VOCs) helps reducing tropospheric ozone concentrations within the same city? Or, how do regional phenomena play a role in influencing the pollutant levels in Calgary? Traditionally, this is accomplished through process-based modeling, which utilizes the transport and/or kinetic equations that govern the underlying physicochemical process. Such a deterministic model by itself requires massive amount of environmental and meteorological data that are not always available at the sites of interest. On the other hand, a purely statistical approach may be applied based on the mathematical correlations between relevant data but this may in turn render it less useful at other locations due to dissimilarity, especially in meteorological conditions. The following discussion attempts to solve these challenges by recommending the next plausible steps in tackling ozone space-time phenomena:

° Modeling the complex patterns of temporal variability in ozone concentrations using a "two-point" statistic (i.e., a variogram) is insufficient. For example, consider the following set of patterns comprising ozone values at four time instants $\{Z(t_1), Z(t_2), Z(t_3), Z(t_4)\}$ (Figure 8.2). The correct profiles can be restored if the variability of ozone concentrations at all 4 points is considered jointly. The variogram, being a "two-point" statistic, accounts for the variability at only two points and thus for example at temporal lag $\tau = 3$, the variability $Z(t_2)$ and $Z(t_3)$ at time instants t_2 and t_3 that also constitute the temporal patterns is ignored. It should be stressed that reproducing the correct temporal patterns requires the identification of a multipoint statistic, i.e., the joint variability at all four points.

In the modeling perspective, the implication is that the conditional probability of an ozone concentration at any unsampled time instant t_i has to be inferred based on

the pattern exhibited by the neighboring values. The inference of such a probability can be performed by calibrating a mixture density network (MDN). Once the probability is calibrated and inferred, it can be assimilated into a stochastic simulation procedure in order to reproduce the correct multipoint patterns signifying the ozone temporal phenomena.

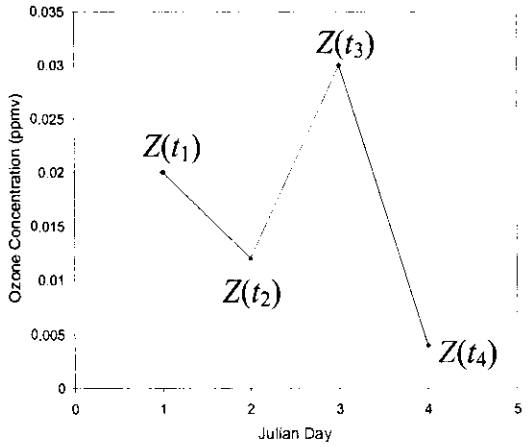


Figure 8.2
A set of ozone temporal patterns.

- The calibration of the temporal parameters via MDN, using only historical ozone data as mentioned above, can be extended to comprise the influence of predictor variables (covariates) on ozone phenomena. Recall that the incorporation of covariates in the prediction algorithms was established via cokriging (Chapter 5) and co-simulation (Chapter 6). Here the linear model of co-regionalization (LMC) was utilized to ensure the legitimacy of the joint auto and cross-variograms. However, cokriging and co-simulation were still carried out using a "two-point" (variogram) statistic, which was previously found to be inadequate for inferring the highly fluctuated profiles and hence resulting in the proposed implementation of a multipoint statistic.

The accuracy of the multipoint statistical approach can be enhanced by utilizing a physicochemical model, e.g., the Urban Airshed Model (UAM), as a way to detect the contemporaneous patterns of variability of ozone and covariate events. For example, the covariate data sets can be varied one at a time while keeping the others constant

and the corresponding ozone profile can be obtained by running the physicochemical model. This yields a training set that can be used to calibrate a MDN not only for the multipoint interactions between historical ozone data but also those between various covariates at different time instants. This difficult task would entail the cross-calibration of the temporal parameters to account for multipoint associations among the covariates. This procedure will again yield the probability of ozone concentration at the unsampled time instant t_i, conditioned on the pattern of covariate information as well as the pattern of available ozone data. This conditional probability can then be assimilated into a stochastic simulation algorithm to predict the ozone patterns at other unsampled time instants. Such a calibration procedure will inject the physics of the ozone phenomena into the statistical methodology.

- Remember that the primary goal in the study of atmospheric pollution is to obtain good knowledge of the space-time phenomena. This difficult task can be alleviated by first modeling the temporal profiles at independent geographical locations, e.g., Calgary, Lethbridge and Medicine Hat, in order to explore the air pollution problems in southern Alberta. Once successful, the spatiotemporal interpolation between these monitoring stations can be performed by recognizing that the sudden rise in ozone concentrations, termed ozone episodes, is likely due to complex nonlinear interactions between ozone and its covariates jointly occurring in space $\mathbf{u}$ and time t. That is to say that the covariate events, e.g., an increase in the precursor levels coupled with high southerly wind speed and bright sunshine in Lethbridge may directly affect the ozone episodes observed on the following day in Calgary.

  This space-time technique would entail the inference of parameters signifying the temporal patterns observed at stations and then subsequently regionalizing those parameters in space. It should be emphasized that the more accurate the temporal modeling is performed at various environmental monitoring stations, the higher the probability of success in estimating ozone values at unknown locations. For this reason, exhaustive studies of ozone phenomena must be carried out at as many cities in Canada as possible.

REFERENCES

1. Abdul-Wahab S, Bouhamra W, Ettouney H, Sowerby B and Crittenden BD (1996), "Predicting Ozone Levels: a Statistical Model for Predicting Ozone Levels," *Environmental Science and Pollution Research*, **3**, 195-204.
2. Alberta Environment (2000), "Alberta Ambient Air Quality Guidelines," Environmental Services Division.
3. Bergin MS, Russel AG and Milford JB (1998), "Effects of Chemical Mechanism Uncertainties on the Reactivity Quantification of Volatile Organic Compounds using a Three-Dimensional Air Quality Model," *Environmental Science and Technology*, **32**, 694-703.
4. Bishops CM (1995), **Neural Network for Pattern Recognition**, Clarendon Press, Oxford.
5. Bloomfield P, Royle JA, Steinberg LJ and Yang Q (1996), "Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends," *Atmospheric Environment*, **30** (17), 3067-3078.
6. Breiman L, Friedman JH, Olshen RA and Stone CJ (1984), **Classification and Regression Trees**, Wadsworth and Brooks/Cole, Belmont, California.
7. BC Ministry of Environment, Lands and Parks (BCMELP) (1992), "No Room to Breathe: Photochemical Smog and Ground-Level Ozone," Air Resources Branch, British Columbia Ministry of Environment, Lands and Parks. Available at http://www.env.gov.bc.ca/epd/epdpa/ar/vehicle/nrtbpsag.html.
8. Burrows WR, Benjamin M, Beauchamp S, Lord E, McCollor D and Thompson B (1995), "CART Decision-Tree Statistical Analysis and Prediction of Summer Season Maximum Surface Ozone for Vancouver, Montreal and Atlantic Regions of Canada," *Journal of Applied Meteorology*, **34**, 1848-1862.
9. Campbell IM (1986), **Energy and the Atmosphere: A Physical-Chemical Approach**, John Wiley and Sons, Inc., New York.

10. Carroll RJ, Chen R, George EI, Li TH, Newton HJ, Schmiediche H and Wang N (1997), "Ozone Exposure and Population Density in Harris County, Texas," *Journal of the American Statistical Association*, **92** (438), 392-404.

11. Chaloulakou A, Assimacopoulos D and Lekkas T (1999), "Forecasting Daily Maximum Ozone Concentrations in the Athens Basin," *Environmental Monitoring and Assessment*, **56** (1), 97-112.

12. Chang JS, Brost RA, Isaksen ISA, Madronich S, Middleton P, Stockwell WR and Walcek CJ (1987), "A Three Dimensional Eulerian and Acid Deposition Model: Physical Concepts and Formulation," *Journal of Geophysical Research - Part D: Atmospheres*, **97** (12), 14681-14700.

13. Chang ME and Cardelino C (2000), "Technical Papers: Application of the Urban Airshed Model to Forecasting Next-Day Peak Ozone Concentrations in Atlanta, Georgia" *Journal of the Air and Waste Management Association*, **50** (11), 2010-2024.

14. Churchill RV (1963), **Fourier Series and Boundary Value Problems**, McGraw Hill, Inc., New York.

15. Clark LA and Pregibon D (1990), "Tree Based Model," in **Statistical Models in S**, eds. JM Chambers and TJ Hastie, Wadsworth & Brooks/Cole, Pacific Grove, California, 377-425.

16. Cobourn WG and Hubbard MC (1999), "Enhanced Ozone Forecasting Model using Air Mass Trajectory Analysis," *Atmospheric Environment*, **33** (28), 4663-4674.

17. Comrie AC (1997), "Comparing Neural Networks and Regression Models for Ozone Forecasting," *Journal of the Air and Waste Management Association*, **47**, 653-663.

18. Cox WM and Chu S-H (1993), "Meteorologically Adjusted Ozone Trends in Urban Areas: A Probabilistic Approach," *Atmospheric Environment - Part B: Urban Atmosphere*, **27** (4), 425-434.

19. Cox WM and Chu S-H (1996), "Assessment of Interannual Ozone Variation in Urban Areas from a Climatological Perspective," *Atmospheric Environment*, **30** (14), 2615-2626.

20. Cramer H (1941), **Mathematical Methods of Statistics**, Princeton University Press.

21. Davis JM and Speckman P (1999), "Model for Predicting Maximum and 8-Hour Average Ozone in Houston," *Atmospheric Environment*, **33** (16), 2487-2500.

22. Davis JM, Eder BK, Nychka D and Yang Q (1998), "Modeling the Effects of Meteorology on Ozone in Houston using Cluster Analysis and Generalized Additive Models," *Atmospheric Environment*, **32** (14), 2505-2620.

23. de Grandpre J, Beagley SR, Fomichev VI, Griffioen E, McConnell JC, Medvedev AS and Shepherd TG (2000), "Composition and Chemistry - Ozone Climatology Using Interactive Chemistry: Results from the Canadian Middle Atmosphere Model (Paper 2000JD900427)," *Journal of Geophysical Research - Part D: Atmospheres*, **105** (21), 26475-26492.

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

25. Derwent DG and Davies TJ (1994), "Modeling the Impact of NOx or Hydrocarbons Control on Photochemical Ozone in Europe," *Atmospheric Environment*, **28** (12), 2039-2052.

26. Deutsch CV and Journel AG (1998), **GSLIB: Geostatistical Software Library and User's Guide**, Oxford University Press, Inc., Oxford.

27. Dickerson RR, Kondragunta S, Stenchikov G, Civerto KL, Doddridge BG and Holben BN (1997), "The Impact of Aerosols on Solar Ultraviolet Radiation and Photochemical Smog," *Science*, **278**, 827-830.

28. Elsner JB, Lehmiller GS and Kimberlain TB (1996), "Objective Classification of Atlantic Hurricanes," *Journal of Climate*, **11**, 2880-2889.

29. Environment Canada (1980), "Guidelines for a Short Term Air Quality Index," Federal-Provincial Committee on Air Pollution, Ottawa, Canada.

30. --- (1993), "Guidelines for the Index of the Quality of Air," Report EPS 1/AP/3.

31. Eskridge RE, Ku JY, Rao ST, Porter PS and Zurbenko IG (1997), "Separating Different Scales of Motion in Time Series of Meteorological Variables," *Bulletin of the American Meteorological Society*, **78**, 1473-1483.

32. Feister U and Balzer K (1991), "Surface Ozone and Meteorological Predictors on a Subregional Scale," *Atmospheric Environment*, **25**, 1781-1790.

33. Fiore AM, Jacob DJ, Logan JA and Yin JH (1998), "Long-Term Trends in Ground Level Ozone over the Contiguous United States (1980-1995)," *Journal of Geophysical Research*, **103**, 1471-1480.

34. Fishman JS, Solomon S and Crutzen PJ (1979), "Observational and Theoretical Support of a Significant In-Situ Photochemical Source of Tropospheric Ozone," *Tellus*, **31**, 432-446.

35. Friedman JH (1990), "Multivariate Adaptive Regression Splines," Technical Report 102 Rev., Department of Statistics, Stanford University, California.

36. Gandin L (1963), **Objective Analysis of Meteorological Fields**, Gidrometeo-rologicheskoe Izdatel'stvo (GIMEZ), Leningrad. Reprinted by Israel Program for Scientific Translations, Jerusalem, 1965.

37. Gao D, Stockwell WR and Milford JB (1995), "First-Order Sensitivity and Uncertainty Analysis for a Regional-Scale Gas-Phase Chemical Mechanism," *Journal of Geophysical Research*, **100** (23), 23153-23166.

38. Gao D, Stockwell WR and Milford JB (1996), "Global Uncertainty Analysis of a Regional-Scale Gas-Phase Chemical Mechanism," *Journal of Geophysical Research*, **101**, 9107-9119.

39. Gardner MW and Dorling SR (1998), "Artificial Neural Networks (the Multilayer Perceptron): A Review of Applications in the Atmospheric Sciences," *Atmospheric Environment*, **32** (14-15), 2627-2636.

40. Gardner MW and Dorling SR (1999a), " Neural Network Modeling of Hourly NO_x and NO_2 Concentrations in Urban Air in London," *Atmospheric Environ-ment*, **33** (5), 709-719.

41. Gardner MW and Dorling SR (2000a), "Statistical Surface Ozone Models: An Improved Methodology to Account for Non-Linear Behaviour," *Atmospheric Environment*, **34** (1), 21-34.

42. Gardner MW and Dorling SR (2000b), "Meteorologically Adjusted Trends in UK Daily Surface Ozone Concentrations," *Atmospheric Environment*, **34**, 171-176.

43. Gery MW, Whitten GZ, Killus JP and Dodge MC (1989), "A Photochemical Kinetics Mechanism for Urban and Regional Scale Computer Modeling," *Journal of Geophysical Research*, **94 (D10)**, 12925-12956.

44. Goovaerts P (1997), **Geostatistics for Natural Resources Evaluation**, Oxford University Press, New York.

45. Guardani R, Nascimento CAO, Guardani MLG, Martins MHRB and Romano J (1999), "Study of Atmospheric Ozone Formation by Means of a Neural Network-Based Model," *Journal of the Air & Waste Management Association*, **49** (3), 316-323.

46. Hanna SR, Lu Z, Christopher-Frey H, Wheeler N, Vukovich J, Arunachalam S, Fernau M, Alan-Hansen D (2001), "Uncertainties in Predicted Ozone Concentrations Due to Input Uncertainties for the UAM-V Photochemical Grid Model Applied to the July 1995 OTAG Domain," *Atmospheric Environment*, **35** (5), 891-904.

47. Haslett J and Raftery AE (1989), "Space-Time Modeling with Long-Memory Dependence: Assessing Ireland's Wind Power Resource," *Applied Statistics*, **38**, 1-21.

48. Hastie TJ and Pregibon D (1992), "Generalized Linear Model," in **Statistical Models in S**, eds. JM Chambers and TJ Hastie, Wadsworth & Brooks/Cole, Pacific Grove, California, 195-247.

49. Hornik K, Stinchcombe M and White H (1989), "Multilayer Feedforward Networks Are Universal Approximators," *Neural Networks*, **2**, 359-366.

50. Huang L and Smith RL (1999), "Meteorologically-Dependent Trends in Urban Ozone," *Environmetrics*, **10** (1), 103-118.

51. IMSL (1992), "FORTRAN Subroutine for Statistical Analysis," User's Manual Version 3.0, IMSL, Houston, Texas.

52. Isaaks EH and Srivastava RM (1989), **An Introduction to Applied Geostatistics**, Oxford University Press, Inc., Oxford.

53. Jonson JE, Sundet JK and Tarrason L (2001), "Model Calculations of Present and Future Levels of Ozone and Ozone Precursors with a Global and a Regional Model," *Atmospheric Environment*, **35** (3), 525-538.

54. Journel AG and Huijbregts C (1972), "Estimation of Lateritic-Type Ore-bodies," in *Proceedings of the 10th International APCOM Symposium*, Society of Mining Engineers, Johannesburg, 202-212.

55. Journel AG and Huijbregts C (1978), **Mining Geostatistics**, Academic Press, New York.

56. Katsoulis BD (1996), "The Relationship between Synoptic, Mesoscale and Microscale Meteorological Parameters during Poor Air Quality Events in Athens, Greece," *Science of the Total Environment*, **181**, 13-24.

57. Kajii Y, Someno K, Tanimoto H, Hirokawa J, Akimoto H, Katsuno T and Kawara J (1998), "Evidence for the Seasonal Variation of Photochemical Activity of Tropospheric Ozone: Continuous Observation of Ozone and CO at Happo, Japan," *Geophysical Research Letters*, **25** (18), 3505-3508.

58. Korsog PE and Wolff GT (1991), "An Examination of Urban Ozone Trends in the Northeastern US (1973-1983) using a Robust Statistical Method," *Atmospheric Environment*, **25**, 47-57.

59. Kyriakidis PC (1998), "Stochastic Models for Spatiotemporal Distributions: Application to Sulfate Deposition over Europe," in Report 11, Stanford Center for Reservoir Forecasting, Vol. 2, Stanford, California.

60. Kyriakidis PC (1999), "Stochastic Simulation of Spatiotemporal Phenomena," **Ph.D. Thesis**, Stanford University, California.

61. LeBlanc M and Crowley J (1993), "Survival Trees by Goodness of Fit," *Journal of the American Statistical Association*, **88**, 457-467.

62. Leone JA and Seinfeld JH (1985), "Comparative Analysis of Chemical Reaction Mechanisms for Photochemical Smog," *Atmospheric Environment*, **19**, 437-464.

63. Lighthill MJ, F.R.S. (1959), **Introduction to Fourier Analysis and Generalized Functions**, Cambridge University Press, Cambridge.

64. Liu MK and Seinfeld JH (1975), "On the Validity of Grid and Trajectory Models of Urban Air Pollution," *Atmospheric Environment*, **9**, 555-574.

65. Liu SC, Trainer M, Fehsenffeld FC, Parrish DD, Williams EJ, Fahey DW, Hubler G and Murphy PC (1987), "Ozone Production in the Rural Troposphere and the Implications for Regional and Global Distributions," *Journal of Geophysical Research*, **92**, 4191-4207.

66. Matheron G (1962), Traité de Géostatistique Appliquée, Vol. 1 (1962), Vol. 2 (1963), ed. Technip, Paris.

67. Matthijsen J, Builtjes PJH and Meyer EW (1996), "Modeling of Cloud Effects on Ozone over Europe," in *Air Pollution Modeling and Its Application XI*, eds., Gryning and Schiermeier, Plenum Press, New York.

68. McRae GJ and Seinfeld JH (1984), "Development of a Second-Generation Mathematical Model for Urban Pollution II: Model Performance Evaluation," *Atmospheric Environment* **17**, 501-523.

69. Milanchus ML, Rao ST and Zurbenko IG (1998), "Evaluating the Effectiveness of Ozone Management Efforts in the Presence of Meteorological Variability," *Journal of the Air and Waste Management Association*, **48** (3), 201-215.

70. Milford JB, Gao D, Russell AG and McRae GJ (1992), "Use of Sensitivity Analysis to Compare Chemical Mechanisms for Air-Quality Modeling," *Environmental Science and Technology*, **26**, 1179-1189.

71. Morris RE and Myers TC (1990), **Volume I: User's Manual for UAM (CB-IV)**, EPA-450/4-90-007A, Research Triangle Park, North Carolina.

72. Nkemdirim LC (1988), "An Assessment of the Relationship between Functional Groups of Weather Elements and Atmospheric Pollution in Calgary, Canada." *Atmospheric Environment*, **22** (10), 2287-2296.

73. Niu X-F (1996), "Nonlinear Additive Models for Environmental Time Series, with Applications to Ground-Level Ozone Data Analysis," *Journal of the American Statistical Association*, **91** (435), 1310-1321.

74. NRC (National Research Council) (1991), "Rethinking the Ozone Problem in Urban and Regional Air Pollution," National Academy Press, Washington, DC.

75. **Numerical Recipes in C: The Art of Scientific Computing**, 2nd Ed., Numerical Recipes Software; available online at www.nr.com.

76. Ozisik MN (1993), **Heat Conduction**, John Wiley and Sons, Inc., New York.

77. Papoulis A (1991), **Probability Random Variables and Stochastic Processes**, McGraw-Hill, New York.

78. Prybutok VR, 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** (1), 31-40.

79. Rao ST, Sistia G and Henry R (1992), "Statistical Analysis of Trends in Urban Ozone Air Quality," *Journal of the Air and Waste Management Association*, **42**, 1204-1211.

80. Rao ST, Zalewsky E and Zurbenko IG (1995), "Determining Temporal and Spatial Variations in Ozone Air Quality," *Journal of the Air and Waste Management Association*, **45** (1), 57-61.

81. Rao ST, Zurbenko IG, Neagu R, Porter PS, Ku JY and Henry RF (1997), "Space and Time Scales in Ambient Ozone Data," *Bulletin of the American Meteorological Society*, **78** (10), 2153-2166.

82. Roelofs G-J and Lelieveld J (2000), "Composition and Chemistry - Tropospheric Ozone Simulation with a Chemistry-General Circulation Model: Influence of Higher Hydrocarbon Chemistry (Paper 2000JD900316)," *Journal of Geophysical Research - Part D: Atmospheres*, **105** (18), 22697-22712.

83. SAI (System Applications International) (1999), **User's Guide to the Variable-Grid Urban Airshed Model (UAM-V)**, SYSAPP-99-95/27r2.

84. Seinfeld JH (1985), **Atmospheric Physics and Chemistry of Air Pollution**, Wiley, New York.

85. Seinfeld JH (1988), "Ozone Air Quality Model: A Critical Review," *Journal of Air Pollution and Control Association*, **38**, 616-645.

86. Smith RL (1989), "Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-level Ozone (with discussion)," *Statistical Science*, **4**, 367-393.

87. Smith RL and Huang L (1993), "Modeling High Threshold Exceedances of Urban Ozone," National Institute for Statistical Science Technical Report #6.

88. Smith RL and Shively TS (1995), "Point Process Approach to Modeling Trends in Tropospheric Ozone Based on Exceedances of a High Threshold," *Atmospheric Environment*, **29** (23), 3489-3500.

89. Smolarkiewicz PK (1983), "A Simple Positive Definite Advection Scheme with Small Implicit Diffusion," *Monthly Weather Review*, **111**, 479-486.

90. Soja G and Soja A-M (1999), "Ozone Indices Based on Simple Meteorological Parameters: Potentials and Limitations of Regression and Neural Network Models," *Atmospheric Environment*, **33** (26), 4299-4307.

91. Sportisse B (2001), "Box Models Versus Eulerian Models in Air Pollution Modeling," *Atmospheric Environment*, **35** (1), 173-178.

92. Stull R (1988), **An Introduction to Boundary Layer Meteorology**, Kluwer Academic Publishers, Dordrecht.

93. Theodoridis S and Koutroumbas K (1999), **Pattern Recognition**, Academic Press, San Diego, California.

94. Thompson ML, Reynolds J, Cox LH, Guttorp P and Sampson PD (2001), "A Review of Statistical Methods for the Meteorological Adjustment of Tropospheric Ozone," *Atmospheric Environment*, **35** (3), 617-630.

95. Turner DB (1970), "Workbook of Atmospheric Dispersion Estimates," U.S. Environment Protection Agency (EPA) Report AP-26, U.S. Government Printing Office, Washington, DC.

96. van Loon M (1996), "Numerical Methods in Smog Prediction," **Ph.D. Thesis**, University of Amsterdam, The Netherlands.

97. Voltz M and Goulard M (1994), "Spatio Interpolation of Soil Moisture Retention Curves," *Geoderma*, **62**, 109-123.

98. Vuilleumeir L, Harley RA and Brown NJ (1997), "First- and Second-Order Sensitivity Analysis of a Photochemically Reactive System (a Green's Function Approach),"*Environmental Science and Technology*, **31**, 1206-1217.

99. Vuilleumier L, Harley RA, Brown NJ, Slusser JR, Kolinski D and Bigelow DS (2001), "Variability in Ultraviolet Total Optical Depth During the Southern California Ozone Study (SCOS97)," *Atmospheric Environment*, **35** (6), 1111-1122.

100. Vukovich FM (1995), "Regional-Scale Boundary Layer Ozone Variations in the Eastern United States and Their Association with Meteorological Variations," *Atmospheric Environment*, **29** (17), 2259-2273.

101. Wakamatsu S, Uno I, Ueda H and Uehara K (1989), "Observational Study of Stratospheric Ozone Intrusions into the Lower Troposphere," *Atmospheric Environment*, **23**, 1815-1826.

102. Walters TS (1969), "The Importance of Diffusion along the Mean Wind Direction for a Ground-level Crosswind Line Source," *Atmospheric Environment*, **3**, 461-466.

103. Wilmott CJ (1982), "Some Comments on the Evaluation of Model Perfor-mance," *Bulletin of the American Meteorological Society*, **63** (11), 1309-1313.

104. Yang YJ, Stockwell WR and Milford JB (1995), "Uncertainties in Incremental Reactivities of Volatile Organic Compounds," *Environmental Science and Technology*, **29**, 1336-1345.

105. Yang YJ, Stockwell WR and Milford JB (1996), "Effect of Chemical Product Yield Uncertainties on Reactivities of VOCs and Emissions from Reformulated Gasoline and Methanol Fuels," *Environmental Science and Technology*, **30**, 1392-1397.

106. Zannetti P (1990), **Air Pollution: Theories, Computational Methods and Available Software**, Computational Mechanics Publications, Van Nostrand Rheinhold, New York.

APPENDICES

APPENDIX A

GSLIB PARAMETER FILES

```
                  Parameters for GAM
                  ******************

START OF PARAMETERS:
d97std.dat                -file with data
1   8   3                 -   number of variables, column numbers
-1.0e21     1.0e21        -   trimming limits
O397.out                  -file for variogram output
1                         -grid or realization number
365  0.5   1.0              -nx, xmn, xsiz
 1   0.5   1.0              -ny, ymn, ysiz
 1   0.5   1.0              -nz, zmn, zsiz
1  73                     -number of directions, number of lags
 5  0  0                  -ixd(1),iyd(1),izd(1)
0                         -standardize sill? (0=no, 1=yes)
1                         -number of variograms
1  1   1             -tail variable, head variable, variogram type


type 1 = traditional semivariogram
     2 = traditional cross semivariogram
     3 = covariance
     4 = correlogram
     5 = general relative semivariogram
     6 = pairwise relative semivariogram
     7 = semivariogram of logarithms
     8 = semimadogram
     9 = indicator semivariogram - continuous
    10 = indicator semivariogram - categorical
```

Appendix A.1
Parameter file `gam.par` that is applied for generating experimental (sample) variogram from the 1997 ozone data. Here the standardized (at zero mean and unit variance) ozone concentrations are used to enhance the quality of the sample variogram values.

```
                    Parameters for VMODEL
                    *********************

START OF PARAMETERS:
O397.var                          -file for variogram output
1   73                            -number of directions and lags
 0.0   0.0   5                    -azm, dip, lag distance
2    0.0                          -nst, nugget effect
2    0.5  0.0   0.0   0.0         -it,cc,ang1,ang2,ang3
          5.0   5.0  10.0         -a_hmax, a_hmin, a_vert
3    0.5  0.0   0.0   0.0         -it,cc,ang1,ang2,ang3
        100.0   5.0  10.0         -a_hmax, a_hmin, a_vert
```

Appendix A.2

Parameter file `vmodel.par` that is applied for modeling the sample variogram, generated by the `gam.par` above. This variogram model has 73 temporal lags, 5 days apart. Two basic variogram structures are implemented: (1) Exponential model (**it** = 2) with 0 nugget, 0.5 sill contribution and range of 5 days, and (2) Gaussian model (**it** = 3) with 0 nugget, 0.5 sill contribution and range of 100 days.

```
                      Parameters for KT3D
                      *******************

START OF PARAMETERS:
O3_12k.dat                    -file with data
1   10   0    5    0         -   columns for X, Y, Z, var, sec var
-1.0e21   1.0e21              -   trimming limits
0                             -option: 0=grid, 1=cross, 2=jackknife
xvk.dat                       -file with jackknife data
1   2   0    3    0           -   columns for X,Y,Z,vr and sec var
3                             -debugging level: 0,1,2,3
O397kt_98.dbg                 -file for debugging output
O397kt_98.out                 -file for kriged output
365  1.0    1.0                  -nx,xmn,xsiz
1    1.0    1.0                  -ny,ymn,ysiz
1    0.5    1.0                  -nz,zmn,zsiz
1    1      1                 -x,y and z block discretization
1    8                        -min, max data for kriging
0                             -max per octant (0-> not used)
250.0  1.0  1.0               -maximum search radii
 90.0  0.0  0.0               -angles for search ellipsoid
1     2.302                   -0=SK,1=OK,2=non-st SK,3=exdrift
0 0 0 0 0 0 0 0 0             -drift: x,y,z,xx,yy,zz,xy,xz,zy
0                             -0, variable; 1, estimate trend
extdrift.dat                  -gridded file with drift/mean
4                             -   column number in gridded file
2    0                            -nst, nugget effect
2    0.5  90.0   0.0   0.0   -it,cc,ang1,ang2,ang3
          5.0   5.0  10.0       -a_hmax, a_hmin, a_vert
3    0.5  90.0   0.0   0.0   -it,cc,ang1,ang2,ang3
        100.0   5.0  10.0       -a_hmax, a_hmin, a_vert
```

Appendix A.3

Parameter file `kt3d.par` that is applied in the ordinary kriging (OK) algorithm for predicting ozone concentrations in 1998 based on the 1997 variogram model. The variogram is modeled using `vmodel.par` prior to kriging. The `O3_12k.dat` is the data file consisting of 12 standardized (at zero mean and unit variance) sample values, evenly spaced at every 30th Julian day of 1998.

```
                    Parameters for COKB3D
                    *********************

START OF PARAMETERS:
12ck98.dat                  -file with data
3                           -number of variables primary + other
1   5   0   2   3   4       -   columns for X,Y,Z and variables
-10.01      1.0e21          -   trimming limits
0                           -co-located cokriging? (0=no, 1=yes)
somedata.dat                -   file with single gridded covariate
4                           -   column for covariate
0                           - local varying mean (0=no, 1=yes)
lvmfl.dat                   -   file with local varying mean
4                           -   column for local varying mean
3                           -debugging level: 0,1,2,3
O397ck_98.dbg               -file for debugging output
O397ck_98.out               -file for output
365   1.0   1.0               -nx,xmn,xsiz
  1   1.0   1.0               -ny,ymn,ysiz
  1   0.5   1.0               -nz,zmn,zsiz
1    1     1                 -x, y, and z block discretization
1   12     8                 -min primary,max primary,max all sec
250.0  1.0  1.0               -maximum search radii: primary
250.0  1.0  1.0               -maximum search radii: all secondary
 90.0  0.0  0.0               -angles for search ellipsoid
1                           -kriging type (0=SK, 1=OK, 2=OK-trad)
3.38  2.32  0.00  0.00      -mean(i),i=1,nvar
3                         - model type (1=MM1, 2=MM2, 3=LMC)
0.50                      - correlation coefficient for MM1 or MM2
10.0                      - variance of secondary variable for MM1
5.0                       - variance of primary variable for MM2
          :
(continued into the next page)
```

Appendix A.4

Parameter file `cokb3d.par` that is applied in the ordinary cokriging (COK) algorithm for predicting ozone concentrations in 1998 based on the 1997 variogram model, and using two covariates: total hydrocarbon (THC) and nitric oxide (NO). The auto and cross-variogram models have been ensured positive-definite via the linear model of coregionalization (LMC) prior to cokriging. The `12ck98.dat` is the data file comprising three sets of 12 standardized (at zero mean and unit variance) sample values of ozone, THC and NO, evenly spaced at every 30th Julian day of 1998.

```
                :
(continued from the previous page)

1     1                  -semivariogram for "i" and "j" (O3)
2   1e-10                              -   nst, nugget effect
2   0.56  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
           5.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
3   0.44  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
         120.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
1     2                  -semivariogram for "i" and "j" (O3-THCm)
2   1e-10                              -   nst, nugget effect
2   0.24  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
           5.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
3   0.44  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3n
         120.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
1     3                  -semivariogram for "i" and "j" (O3-NOm)
2   1e-10                              -   nst, nugget effect
2   0.26  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
           5.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
3   0.41  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
         120.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
2     2                  -semivariogram for "i" and "j" (THCm)
2   1e-10                              -   nst, nugget effect
2   0.46  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
           5.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
3   0.54  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
         120.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
2     3                  -semivariogram for "i" and "j" (THCmNOm)
2   1e-10                              -   nst, nugget effect
2   0.39  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
           5.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
3   0.50  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
         120.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
3     3                  -semivariogram for "i" and "j" (NOm)
2   1e-10                              -   nst, nugget effect
2   0.50  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
           5.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
3   0.50  90.0   0.0  0.0              -   it,cc,ang1,ang2,ang3
         120.0   0.0  0.0              -   a_hmax, a_hmin, a_vert
```

Appendix A.4

Parameter file `cokb3d.par` that is applied in the ordinary cokriging (COK) algorithm for predicting ozone concentrations in 1998 based on the 1997 variogram model, and using two covariates: total hydrocarbon (THC) and nitric oxide (NO). The auto and cross-variogram models have been ensured positive-definite via the linear model of coregionalization (LMC) prior to cokriging. The `12ck98.dat` is the data file comprising three sets of 12 standardized (at zero mean and unit variance) sample values of ozone, THC and NO, evenly spaced at every 30th Julian day of 1998.

```
                    Parameters for SGSIM
                    ********************

START OF PARAMETERS:
O3_12k.dat                  -file with data
1  10   0   4   0   0       -  columns for X,Y,Z,vr,wt,sec.var.
-1.0         1.0e21         -  trimming limits
1                           -transform the data (0=no, 1=yes)
O397_s98.trn                -  file for output trans table
0                           -  consider ref. dist (0=no, 1=yes)
histsmth.out                -  file with ref. dist distribution
1  2                        -  columns for vr and wt
0.0    0.05                 -  zmin,zmax(tail extrapolation)
1      0.0                  -  lower tail option, parameter
1     15                    -  upper tail option, parameter
1                           -debugging level: 0,1,2,3
O397_s98.dbg                -file for debugging output
O397_s98.out                -file for simulation output
10                          -number of realizations to generate
365     1     1.0             -nx,xmn,xsiz
1       1     1.0             -ny,ymn,ysiz
1     0.5     1.0             -nz,zmn,zsiz
2374321                     -random number seed
1     8                     -min and max original data for sim
12                          -number of simulated nodes to use
1                           -assign data to nodes (0=no, 1=yes)
1     3                     -multiple grid search (0=no, 1=yes)
0                           -maximum data per octant (0=not used)
250.0  1.0  1.0             -maximum search radii (hmax,hmin,vert)
 90.0   0.0   0.0           -angles for search ellipsoid
1   0.60  1.0               -ktype: 0=SK,1=OK,2=LVM,3=EXDR,4=COLC
../data/ydata.dat           -file with LVM, EXDR, or COLC variable
4                           -  column for secondary variable
2    0                          -nst, nugget effect
2    0.5  90.0   0.0   0.0     -it,cc,ang1,ang2,ang3
            5.0   5.0  10.0     -a_hmax, a_hmin, a_vert
1    0.5  90.0   0.0   0.0     -it,cc,ang1,ang2,ang3
          100.0   5.0  10.0     -a_hmax, a_hmin, a_vert
```

Appendix A.5

Parameter file `sgsim.par` that is applied in the sequential Gaussian simulation (SGSIM) algorithm for predicting ozone concentrations in 1998 based on the 1997 variogram model. The variogram is modeled using `vmodel.par` prior to simulation. The `O3_12k.dat` is the data file consisting of 12 raw sample values, evenly spaced at every 30^{th} Julian day of 1998.

```
                     Parameters for SGSIM
                     ********************

START OF PARAMETERS:
12cs98.dat                        -file with data
1  5  0  2  0  3                  -  columns for X,Y,Z,vr,wt,sec.var.
-100.0         1.0e21             -  trimming limits
1                                 -transform the data (0=no, 1=yes)
O397_cs98.trn                     -  file for output trans table
0                                 -  consider ref. dist (0=no, 1=yes)
histsmth.out                      -  file with ref. dist distribution
1  2                              -  columns for vr and wt
0.0     0.05                      -  zmin,zmax(tail extrapolation)
1       0.0                       -  lower tail option, parameter
1       0.0                       -  upper tail option, parameter
1                                 -debugging level: 0,1,2,3
O397_cs98.dbg                     -file for debugging output
O397_cs98.out                     -file for simulation output
10                                -number of realizations to generate
365    1.0     1.0                   -nx,xmn,xsiz
1      1.0     1.0                   -ny,ymn,ysiz
1      0.5     1.0                   -nz,zmn,zsiz
2374321                           -random number seed
1      8                          -min and max original data for sim
12                                -number of simulated nodes to use
1                                 -assign data to nodes (0=no, 1=yes)
1      3                          -multiple grid search (0=no, 1=yes)
0                                 -maximum data per octant (0=not used)
250.0    1.0     1.0              -maximum search radii (hmax,hmin,vert)
 90.0    0.0     0.0              -angles for search ellipsoid
6                -ktype: 0=SK,1=OK,2=LVM,3=EXDR,4=MMI,5=MMII,6=LMC
0.7                               -correl. coeff. if MMI or MMII
12                                - number of secondary data to use if LMC
d97rawX.dat                       -  file with LVM, EXDR, or COLC variable
2                                 -  column for secondary variable
            :
(continued into the next page)
```

Appendix A.6

Parameter file `sgsim.par` that is applied in the sequential Gaussian co-simulation algorithm for predicting ozone concentrations in 1998 based on the 1997 variogram model, and using one covariate: total hydrocarbon (THC). The auto and cross-variogram models are obtained using `vmodel.par` and have been ensured positive-definite via the linear model of coregionalization (LMC) prior to co-simulation. The `12cs98.dat` is the data file comprising two sets of 12 raw ozone and THC values, evenly spaced at every 30^{th} Julian day of 1998.

```
        ⋮
(continued from the previous page)

2    1e-10              -nst, nugget effect: primary for LMC
2    0.56  90.0   0.0   0.0   -it,cc,ang1,ang2,ang3
         5.0  1.0  1.0         -a_hmax, a_hmin, a_vert
3    0.44  90.0   0.0   0.0   -it,cc,ang1,ang2,ang3
       120.0  1.0  1.0         -a_hmax, a_hmin, a_vert
2    1e-10              -nst, nugget effect: cross for LMC
2    0.24  90.0   0.0   0.0   -it,cc,ang1,ang2,ang3
         5.0  1.0  1.0         -a_hmax, a_hmin, a_vert
3    0.44  90.0   0.0   0.0   -it,cc,ang1,ang2,ang3
       120.0  1.0  1.0         -a_hmax, a_hmin, a_vert
2    1e-10              -nst, nugget effect: secondary for LMC
2    0.46  90.0   0.0   0.0   -it,cc,ang1,ang2,ang3
         5.0  1.0  1.0         -a_hmax, a_hmin, a_vert
3    0.54  90.0   0.0   0.0   -it,cc,ang1,ang2,ang3
       120.0  1.0  1.0         -a_hmax, a_hmin, a_vert
```

Appendix A.6

Parameter file `sgsim.par` that is applied in the sequential Gaussian co-simulation algorithm for predicting ozone concentrations in 1998 based on the 1997 variogram model, and using one covariate: total hydrocarbon (THC). The auto and cross-variogram models are obtained using `vmodel.par` and have been ensured positive-definite via the linear model of coregionalization (LMC) prior to co-simulation. The `12cs98.dat` is the data file comprising two sets of 12 raw ozone and THC values, evenly spaced at every 30th Julian day of 1998.

APPENDIX B

THE URBAN AIRSHED MODEL (UAM)

Statistical methods, though useful in the study of ozone phenomena, may not be adequate for predicting or estimating ground-level ozone episodes that frequently appear, especially in large cities like Chicago and Toronto. During an ozone episode, one would like to identify what factors (meteorological, chemical, etc.) cause the concentration of ozone to suddenly rise above the normal levels. Of course the statistical methods (e.g., artificial neural network), through improved learning process, can predict this occurrence to within tolerable error provided that historical records exist. However, the application of these methods for reliably predicting the variations in ozone concentration and the associated adverse effects to humans and their welfare requires a good understanding of the atmospheric ozone transport mechanisms. Therefore, if a physicochemical (physical-chemical) model can be coupled with any of these statistical methods, the end results would be more convincing and meaningful. Although detailed numerical simulation using physicochemical models such as the Urban Airshed Model (UAM) was not performed, relevant information was collected and succinct synopsis (overview) of the physicochemical processes, i.e., material balance and ozone chemistry, as well as a three-dimensional (3D) model that is commonly used by the regulatory agencies, e.g., EPA, in ozone prediction are discussed for future research work.

A good physicochemical model should be able to simulate as many chemical reactions involved in the formation and destruction of ozone in the atmosphere as possible. Relevant models for pollutant emissions, transport and removal processes must also be included. Some of the important factors that should be considered are:

- Anthropogenic (man-made) and biogenic (natural) spatiotemporal emissions (area and point sources) of nitrogen oxides (NO_x), volatile organic compounds (VOCs) and/or other relevant chemical species,

- Chemical reactions and kinetics (e.g., NO_2 photolysis rate) involving NO_x, VOCs and/or other species,
- Background concentrations of NO_x, VOCs and other species in upwind location proximity and at the upper atmosphere,
- Space-time variations of the wind fields – direction and speed,
- Stability of the atmosphere (i.e., heights of "diffusion break," diurnally and/or nocturnally) in the region of interest,
- Intensities and abnormalities of solar insolation (e.g., global UV radiation) and temperature (e.g., hourly-maximum or -average, surface and vertical T gradients),
- Removal of ozone and its precursors by dry (e.g., absorption by leaves' stomata) and wet depositions (e.g., rainfall),
- Effects of terrain (surface roughness and deposition factor), and
- Other meteorological factors like atmospheric pressure, opacity/turbidity (e.g., haze and cloud cover) and amount of water vapor (e.g., relative humidity),

Among the many publicly available software packages for modeling physicochemical reactions is the Urban Airshed Model (UAM), developed by the Systems Applications International, Inc. (SAI). This 3D Eulerian grid model is recommended for the simulation of 48- to 120-hour period (2 to 5 days) during an ozone episode. The core principle of the UAM is the material balance of chemical species i in the atmosphere that can be mathematically expressed as a set of partial differential equations (PDEs) of individual species concentration C_i:

$$\frac{\partial C_i}{\partial t} + \nabla\left(v_j C_i\right) = \nabla^2\left(\mathrm{K}_j C_i\right) + R_i + S_i + D_i + W_i, \qquad j = 1,2,3$$

which when expanded in familiar notations becomes (SAI, 1999):

$$\underset{[Time\ Dependence]}{\frac{\partial C_i}{\partial t}} + \underset{[Advection]}{\frac{\partial(uC_i)}{\partial x} + \frac{\partial(vC_i)}{\partial y} + \frac{\partial(wC_i)}{\partial z}} = \underset{[Turbulent\ Diffusion]}{\frac{\partial}{\partial x}\left(K_X \frac{\partial C_i}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_Y \frac{\partial C_i}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_Z \frac{\partial C_i}{\partial z}\right)}$$

$$+ \underset{[Chemical\ Reaction]}{R_i} + \underset{[Emissions]}{S_i} + \underset{[Dry\ and\ Wet\ Depositions]}{D_i} + W_i$$

The remaining variables are defined as:

u, v, w	:	Horizontal and vertical wind speed components
K_X, K_X	:	Horizontal turbulent diffusion coefficients (dispersivities)
K_Z		Vertical turbulent exchange coefficient (dispersivity)
R_i	:	Net production rate of species i by chemical reactions
S_i	:	Emission rate of species i
D_i	:	Net removal rate of species i by surface uptake processes
W_i	:	Net removal rate of species i by wet deposition processes

Note that C_i is a spatiotemporal variable, i.e., it varies jointly in space (x, y, z) and time (t), and must be solved simultaneously from all concentrations of reactive species (pollutants) involved in this process.

There are hundreds of species (mostly organic compounds) involved in complex chemical reactions to produce ozone. However, the main ones are the reactions of NO_x, VOCs and their derivatives in the presence of "bright" sunlight. Depending on the right kinetics, the same species, e.g., atomic oxygen O, may sometime participate in the formation and destruction of ozone (Table B.1). With thousands of chemical reactions occurring almost concurrently in this process, one would expect wide temporal variations in the reaction rate constants. As a result, the PDEs form a system of "stiff" equations, for which explicit solutions demand substantial amount of work. Highly accurate solutions require expensive computing time; hence a more suitable approach is to split

the family of species *i* into two groups: (1) the major species, and (2) the remainders or "state" species. The UAM solves for the former using a quasi-steady-state assumption, and the accompanying variations in the "state" species are obtained by an optimization algorithm. This simplification improves the numerical stability of the method and results in accurate solutions.

Table B.1
The Carbon Bond Mechanism IV (CB-IV).

Reactions			Rate Constants (cm^3/molecule/s)
$NO_2 + h\nu$	→	NO + O	Radiation dependent
$O + O_2$	→	O_3	1.4E3 * exp(1175/T)
O_3 + NO	→	NO_2	1.8E-12 * exp(-1370/T)
O + NO_2	→	NO	9.3E-12
O + NO_2	→	NO_3	1.6E-13 * exp(687/T)
O + NO	→	NO_2	2.2E-13 * exp(602/T)
O_3 + NO_2	→	NO_3	1.2E-13 * exp(-2450/T)
$O_3 + h\nu$	→	O	Radiation dependent

Adapted from Gery et al (1989). $h\nu$ = solar radiation.

The first step is to divide the simulated region into a coarse 3D grid (10-20 km), with a rectangular shape and constant lengths in the horizontal directions (x and y). Finer nested grids (1-2 km each) may then be imposed within the coarse grid to enhance resolution and thereby improve analysis on the more complex surface transport phenomena caused by, e.g., topography (terrain roughness) and oceanic breeze. The arbitrarily structure of the vertical layers is user-defined. The diffusion break is often determined from the bottom of the inversion layer, i.e., either an unstable diurnal convective layer (mixing height) or a stable nocturnal layer at nighttime. Other factors influencing the vertical layers are the number of mixing heights and the bottom layer

thickness. To elucidate the interpretation of these layers, an example of vertical cell structure with accompanying processes in each layer is illustrated in Figure B.1.

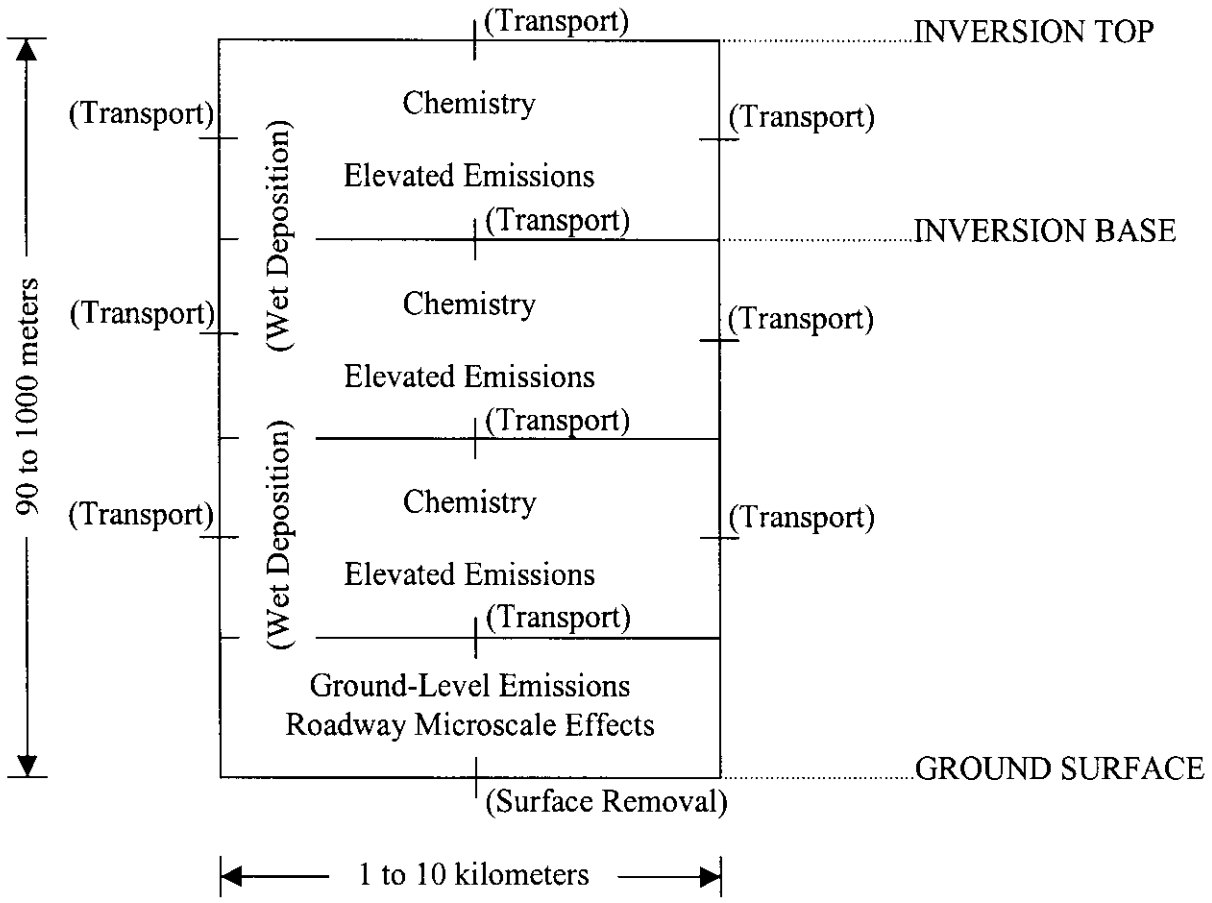


Figure B.1
Schematic diagram of the vertical layers used in the Urban Airshed Model (UAM). Adapted from Morris and Myers (1990).

After the grids are fully defined, each term in the PDE, i.e., advection, turbulent diffusion, chemical reaction (atmospheric chemistry) and removal mechanisms, is solved separately based on the following order: (1) treatment of the horizontal advective-diffusive processes in the x-(east-west), and (2) in the y-(north-south) directions, (3) in the z-(vertical) direction after the "injection" of pollutants, and finally (4) treatment of

the reactive chemical reactions. This four-step operation is performed using optimal algorithms at particular time interval, usually in the order of three to six minutes depending on the grid size and the maximum wind velocity.

The transport of pollutants in the atmosphere is mainly by advection, a process in which a species (or more) is entrained in a bulk fluid (in this case, air) and hence carried along when the fluid moves due to external forces (e.g., wind). The UAM treats advection from the perspective of the wind fields (direction and speed) and components (u, v and w). The horizontal wind components (u and v in each grid cell) are initially specified and the vertical component (w in the same cell) is determined from the terrain-factor and material balance of all species i in each grid cell. The process is repeated until all wind fields are simulated. The advective terms can then be solved in timely interval using the method proposed by Smolarkiewicz (1983).

Turbulent diffusion (or dispersion) is handled by assuming proportionality of the dispersivities (K_X, K_Y, K_Z) to the spatial concentration gradient $C_i(x, y, z)$. Despite this modification, the exact values of the dispersivities are still difficult to be obtained experimentally. Hence the UAM employs theoretical estimates based on the method suggested by Smagorinsky (1963). In essence, the horizontal dispersivities (K_X, K_Y) are inferred by applying scaling factors to some deformation characteristics of the horizontal wind fields (u, v). The vertical diffusivitiy (K_Z) is estimated from the vertical wind component (w) and temperature field, solved by the UAM meteorological preprocessor programs WIND and TEMPERATUR, respectively. Except for the surface (bottom layer), which is specified as dry deposition flux, other boundary conditions (lateral and top layer) are assumed zero mass flux, i.e., no flow of materials in or out.

The removal processes are divided into two major categories - dry and wet depositions. While the latter is primarily due to rainfall (i.e., scrubbing), the former includes several processes. The most important one occurring in the daytime is the absorption of gaseous pollutants (e.g., ozone) into vegetation via stomata, a hole-like feature that controls the pore openings in the leaves. These pollutants, once deposited into vegetated surfaces, are immediately converted to different chemical compounds, hence

reducing their concentrations in the atmosphere. Another process that is also classified as dry involves the absorption of the gaseous compounds into water surfaces, e.g., lakes or oceans. The solubility of these gases in water of different salinity controls this removal process. At nighttime, surface moisture also considered a dry deposition process is important. The extent of the dew-wetted surface is estimated from the relative humidity and wind velocity data. If rainfall rates are available, the rain-wetted surface effect can also be accounted for.

The final step in the UAM modeling is to calculate the species mass balance due to chemical reactions. Because there are thousands of possible reaction mechanisms, it would be prudent (in term of execution time) to treat the chemical species and respective kinetics according to their reactive functional groups. More specifically, the organic compounds are classified based on their carbon bonds. For example, propylene (C_3H_6), butene (C_4H_8) and acetaldehyde (CH_3CHO) are divided into three functional groups, comprising three paraffinic bonds (PAR), three olefinic bonds (OLE) and one higher molecular weight aldehyde (AD2). It may seem that the carbon bond approach increases the computing time but when there are thousands of parallel chemical reactions involved, this is the "fastest" way to obtain fairly accurate results. The extended version of this approach called the Carbon Bond (CB-IV, version 4) Mechanism is still employed in the UAM-V® (version 5) software package. For convenience, several important representations of the chemical species and functional groups employed by the UAM are listed in Table B.2.

Table B.2
Definition of the UAM (CB-IV) chemical species.

UAM Species	Species Name
NO	Nitric oxide
NO2	Nitrogen dioxide
O3	Ozone
OLE	Olefinic carbon bond (C=C)
PAR	Paraffinic carbon bond (C–C)
TOL	Toluene (C_6H_5-CH_3)
XYL	Xylene (C_6H_6-$(CH_3)_2$)
FORM	Formaldehyde (HCHO)
ALD2	High molecular weight aldehydes (RCHO, R > H)
ETH	Ethene (CH_2=CH_2)
CRES	Cresols and higher molecular weight phenols
MGLY	Methyl glyoxal ($CH_3C(O)C(O)H$)
OPEN	High molecular weight aromatic oxidation ring fragment
PNA	Peroxynitric acid (HO_2NO_2)
NXOY	Total of nitrogen compounds ($NO + NO_2 + N_2O_5 + NO_3$)
PAN	Peroxyacyl nitrate ($CH_3C(O)O_2NO_2$)
CO	Carbon monoxide
HONO	Nitrous acid
H2O2	Hydrogen peroxide
HNO3	Nitric acid
MEOH	Methanol (optional)
ETOH	Ethanol (optional)
ISOP	Isoprene (optional)

Adapted from Morris and Myers (1990).

To summarize, the physicochemical models, e.g., the UAM, provide a useful insight in a manner where spatiotemporal variations of the predictor variables such as carbon monoxide, nitrogen oxides and wind speed affect the variation of ozone. These models also form the scientific basis for the air pollution standards adopted by the regulatory agencies. However, such physicochemical models embrace several simplifying assumptions that might render them unrealistic. They are also computationally expensive to implement, and cannot be readily used in predictive mode. Combining physico-chemical models with stochastic approaches can improve prediction accuracy, provided the calibration between the physicochemical model outputs and the stochastic variables is rigorously performed. Statistical tools such as disjunctive kriging, neural networks, Bayesian estimation can also accomplished such calibration through training and likelihood functions. The resultant predictions can be constructed to be locally accurate (i.e., data exactitude) and that yielding an assessment of global uncertainty (i.e., annual patterns of spatiotemporal variation of ozone).