Air Quality Model Evaluation Through the Analysis of Spatial-Temporal Ozone Features

by

Tianji Shi

B.Sc., University of Massachusetts Amherst, 2006
M.Sc., University of Massachusetts Amherst, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2015

Abstract

Legislative actions regarding ozone pollution use air quality models (AQMs) such as the Community Multiscale Air Quality (CMAQ) model for scientific guidance, hence the evaluation of AQM is an important subject. Traditional point-to-point comparisons between AQM outputs and physical observations can be uninformative or even misleading since the two datasets are generated by discrepant stochastic spatial processes. I propose an alternative model evaluation approach that is based on the comparison of spatial-temporal ozone features, where I compare the dominant space-time structures between AQM ozone and observations. To successfully implement feature-based AQM evaluation, I further developed a statistical framework of analyzing and modelling space-time ozone using ozone features. Rather than working directly with raw data, I analyze the spatial-temporal variability of ozone fields by extracting data features using Principal Component Analysis (PCA). These features are then modelled as Gaussian Processes (GPs) driven by various atmospheric conditions and chemical precursor pollution. My method is implemented on CMAQ outputs during several ozone episodes in the Lower Fraser Valley (LFV), BC. I found that the feature-based ozone model is an efficient way of emulating and forecasting a complex space-time ozone field. The framework of ozone feature analysis is then applied to evaluate CMAQ outputs against the observations. Here, I found that CMAQ persistently over-estimates the observed spatial ozone pollution. Through the modelling of feature differences, I identified their associations with the computer model's estimates of ozone precursor emissions, and this CMAQ deficiency is focused on LFV regions where the pollution process transitions from NOx-sensitive to VOC-sensitive. Through the comparison of dynamic ozone features, I found that the CMAQ's over-prediction is also

connect to the model producing higher than observed ozone plume in daytime. However, the computer model did capture the observed pattern of diurnal ozone advection across LFV. Lastly, individual modelling of CMAQ and observed ozone features revealed that even under the same atmospheric conditions, CMAQ tends to significantly over-estimate the ozone pollution during the early morning. In the end, I demonstrated that the AQM evaluation methods developed in this thesis can provide informative assessments of an AQM's capability.

Preface

The statistical methods of air quality model (AQM) evaluation, as well as the particular framework of space-time ozone modelling presented in this thesis are the products of my original ideas, with extensive guidance and motivation from my supervisors: Prof. Douw G. Steyn and Prof. William J. Welch. The statistical AQM evaluation methods and the ozone modelling framework are essentially a collection of mostly existing statistical methodologies combined and applied in a novel way. The source of existing methodologies and results are cited and discussed at the appropriate places in the main text.

The entire research is originally motivated by a scientific question posed by Prof. Steyn. The data are provided by Prof. Steyn, Dr. Bruce Ainslie from Environment Canada, and Metro Vancouver. All computer codes, unless noted in the thesis, are programmed by me.

Selected materials in Chapters 3, 4 and Appendix B.1 were summarized into a 30-minute presentation at the 33rd International Technical Meeting (ITM) on Air Pollution Modelling and its Application, which took place August 26th to 30th, 2013 in Miami, Florida. The presented materials were further prepared for a chapter in the book "Air Pollution Modelling and its Application, XXIII" (Copyright 2014, Springer). Additional manuscripts based on the research in this thesis are in preparation for submission to peer-reviewed journals.

Table of Contents

List of Tables

List of Figures

List of Mathematical Notations, Acronyms and Abbreviations

Notations and Representations	
$\mathbf{Y}$	Generic notation for a matrix or a vector of a random response
$\mathbf{X}$	A matrix or a vector of model covariates
$\mathbf{O}$	The general notation for a matrix of space-time ozone data
$\mathbf{O}^c$	A data matrix of CMAQ outputs
$\mathbf{O}^o$	A data matrix of ozone physical observations
${}^t\mathbf{O}$	Simulated *true* ozone field
${}^c\mathbf{O}$	Simulated CMAQ output
${}^o\mathbf{O}$	Simulated ozone observations
t	The number of time points (usually in hourly intervals) in space-time data
n	The number of locations in space-time data
$\mathbf{E}$	A matrix of Empirical Orthogonal Functions extracted from space-time data, where each column of $\mathbf{E}$ usually captures spatial ozone feature
$\mathbf{P}$	A matrix of Principal Components, where each column of $\mathbf{P}$ usually captures temporal ozone feature
i, j	Row and column indexes of $\mathbf{O}$, $\mathbf{E}$ and $\mathbf{P}$
$\mathbf{E}_j$	The general notation of multivariate random process representing j-th column of $\mathbf{E}$

Notations and Representations (continued)	
$\mathbf{P}_j$	The general notation of multivariate random process representing j-th column of $\mathbf{P}$
$\mathbf{E}_j^c$	$\mathbf{E}_j$ from a CMAQ modelled ozone process
$\mathbf{P}_j^c$	$\mathbf{P}_j$ from a CMAQ modelled ozone process
$\mathbf{E}_j^o$	$\mathbf{E}_j$ from an observed ozone process
$\mathbf{P}_j^o$	$\mathbf{P}_j$ from an observed ozone process
$\mathbf{E}_j^d$	The j-th order feature difference of the EOF
$\mathbf{P}_j^d$	The j-th order feature difference of the PC
$\mathbf{x}_{\mathbf{E}\_j}$	Model covariates of $\mathbf{E}_j$
$\mathbf{x}_{\mathbf{P}\_j}$	Model covariates of $\mathbf{P}_j$
p	The number of spatio-temporal ozone features or the number of EOF-PC components used to model space-time ozone processes $\mathbf{O}$
$\mathbf{D}$	An ozone "decenter" matrix containing spatial means of $\mathbf{O}$
$\mathbf{R}$	The correlation matrix of a Gaussian Process
$\mathbf{Z}$	The zero-mean stochastic component of a Gaussian Process
$\mathbf{f}$	A vector of Gaussian Process regression functions
$\mathbf{F}$	The regression design matrix of a Gaussian Process
$\boldsymbol{\beta}$	A vector of regression coefficients
σ	Model standard deviation
$\boldsymbol{\theta}$	A vector of Gaussian Process correlation parameters
$\boldsymbol{\alpha}$	A vector of Gaussian Process smoothness parameters
$\boldsymbol{\xi}$	A vector of GP model parameters $\boldsymbol{\beta}$, σ, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$
h	Variable denoting the hour of the day
$\{x, y\}$	Specifically for synthetic ozone data, the horizontal and vertical location index within a 2-D geographical coordinate system.

Acronyms and Abbreviations	
GP	Gaussian Process
PCA	Principal Component Analysis
EOF	Empirical Orthogonal Function
PC	Principal Component
BLUP	Best Linear Unbiased Predictor
CMAQ	Community Multiscale Air Quality
WRF	Weather Research and Forecast
SMOKE	Sparse Matrix Operator Kernel Emissions
MCIP	Meteorology-Chemistry Interface Processor
AQM	Air Quality Model
VM model	*Variable Mean* model: An ozone feature model whose covariates are the spatial or temporal means of the model variables
CII model	*Covariate Iterative Improvement* model: An ozone feature model whose covariates are the PCA outputs of the model variables selected through the *Iterative Improvement* algorithm
CSC model	*Covariate Subset Combination* model: An ozone feature model whose covariates are the PCA outputs of the model variables selected through the *Subset Combination* procedure
GASP	The computer program used in this thesis to optimize the GP models.
Temp	Temperature measured in *Kelvin*
Wind	Wind speed measured in *meters per second*
BL	Planetary boundary layer height in meters
NOX-lag	The antecedent or lagged NOx concentration
VOC	Volatile Organic Compound
VOC-lag	The antecedent or lagged VOC concentration

Acknowledgements

I would like to thank Prof. Welch and Prof. Steyn for years of patient guidance, support and allowing an uncountable number of meetings. This research project and thesis would not be possible without you.

A hearty thank you to Dr. Bruce Ainslie for teaching me the complicated science behind ozone modelling, and helping me understand all the data that he made available to me. My research would not be the same without your help.

I would like to thank NSERC for the funding the research grants of Prof. Welch and Prof. Steyn, some of which provided me with financial support over the years.

I would also like to thank Metro Vancouver for providing me with even more data, and for making sure Vancouver stays a great city for living and studying.

Thank you to the department of statistics: especially Prof. Zidek and Prof. Joe for years of interesting conversations and inspirations, the department head and staff for making our department a cozy supportive family.

Last but not least, I would like to thank my friends for all the support, both emotionally and financially.

Dedication

Dedicated to my parents and the rest of my wonderful family.

Chapter 1

Introduction

Ozone is an oxygen compound with chemical composition O_3. In the gaseous form at ground (surface) level, ozone is considered a harmful pollutant especially on the lung function of people with respiratory conditions. In more severe instances, prolonged excessive exposure to ozone is linked to asthma, heart attack and premature death (WHO, 2003; Lippmann, 1989). Over the years, government agencies around the world have drafted and instituted standards defining the maximum ozone threshold deemed to be harmful to humans. In Canada, it is called the Canada Wide Standard (CWS): the fourth highest annual ozone measurement should not exceed the level of 65 *parts-per-billion* (ppb) over an "8-hour averaging time" (CCME, 2000). An ozone standard is enforced through the continuous monitoring and analysis of surface-level ozone (and other air pollutant) concentrations (JAICC, 2005; Reuten et al., 2012).

Ozone formation and destruction is a part of a complex system of interlinked photochemical reactions. The precursor chemicals are the photochemical compounds that once released into the atmosphere, trigger a new chain of ozone reactions. Precursor chemicals are introduced into the atmosphere through human activities; examples of such compounds include NOx (the generic term for NO and NO_2) and Volatile Organic Compounds (VOC) (Boubel et al., 1994, Chapter 12). Therefore, an ozone standard is implemented through the reduction of emission level (Reuten et al., 2012). From the perspectives of the regulatory agencies, it is important to study and understand the effect of precursor emission on ambient ozone concentration. The Community Multiscale Air Quality (CMAQ) modelling system is a useful scientific tool for this very purpose.

CMAQ is a process-based Air Quality Model (AQM) used to model the

spatial-temporal ozone distribution under given meteorological and emission conditions (Byun and Schere, 2006). CMAQ enables atmospheric researchers and managers to model ozone variation over a range of background conditions, making forecasts or conducting retrospective analyses. In other words, CMAQ is useful for estimating the effect of changing weather and precursor emissions on surface-level ozone. Reuten et al. (2012), Steyn et al. (2013) and Ainslie et al. (2013) are some of the most recent, extensive implementations of these types of analyses.

As with any modelling system that aims to simulate a real-life event, especially one as complex as the ozone process, CMAQ users need to have an informed, "big picture" view of its modelling capability. Hence the evaluation of the CMAQ model (and AQMs in general) is an inherently important research subject (Dennis et al., 2010; Galmarini and Steyn, 2010). The topic of CMAQ evaluation is the initial scientific motivation that started the statistical research in this thesis.

1.1 Overview of Ozone Process and CMAQ Modelling System

Before discussing CMAQ evaluation it is useful to describe in detail the science behind the ozone process and CMAQ modelling system.

1.1.1 Tropospheric Ozone Formation Processes

When pollutants are released into the atmosphere, chemical reactions subsequently occur to form new pollutants, one of which is ozone. In the simplest terms, the formation of ozone can be defined as a function of the existing hydrocarbon mixture and the concentration of NO_X plus the intensity of solar radiation and temperature. The hydrocarbon mixture could comprise many types of hydrocarbon compounds. For example, about 43 hydrocarbon compounds have been identified in the air of St. Petersburg, Florida in the 1970s (Boubel et al., 1994, Section 12.3). In addition, the precursors for the creation of ozone would undergo various chemical transformations

of their own, including reactions with ozone, resulting in a highly complex system of atmospheric chemical processes.

The process is further complicated by the fact that these gases are being transported through the atmosphere. Thus, an important aspect of an ozone modelling system is a meteorology model that forecasts the atmospheric circulation over a regional scale relevant to the transportation of pollutants. Also, the meteorology model provides information on ambient conditions like temperature and humidity, which are in turn necessary to model the chemical reactions.

In a nutshell, meteorological and chemical reaction models are essentially complex systems of dynamic functions that work jointly to model the creation and transportation of air pollution. The mechanisms can be summarized into two fundamental steps: (1) Pollutants are released into the atmosphere, (2) Subsequent atmospheric chemical transformations occur, both near the pollution source and over a wide geographical region due to being transported by atmospheric circulation, mixing and reacting with ambient gases along the way.

Space-time Aspects of Ozone Processes

An ozone process has important spatial and temporal structure. Human activities determine the types and intensities of pollutants entering the atmosphere (Steyn et al., 2013; Ainslie et al., 2013). Solar radiation and ambient weather determine the conditions under which the atmospheric reactions occur. Since these factors differ across geographical locations and time periods (hour of the day, season, etc.), pollution concentration naturally varies across space and time. Furthermore, ozone and other air pollution processes do not simply occur independently over space and time. Because of atmospheric transport the pollution level at one location depends on the pollution at other locations and their proximities with this location. In other words, pollutant concentration of location s at time t could significantly influence the pollution at nearby location s' at some time t' in the future (Le and Zidek, 2006). This dynamic aspect of space-time correlation structure

requires modelling in our statistical methods.

For instance, Figure 1.1 shows the diurnal (daily) trend for the four sum-

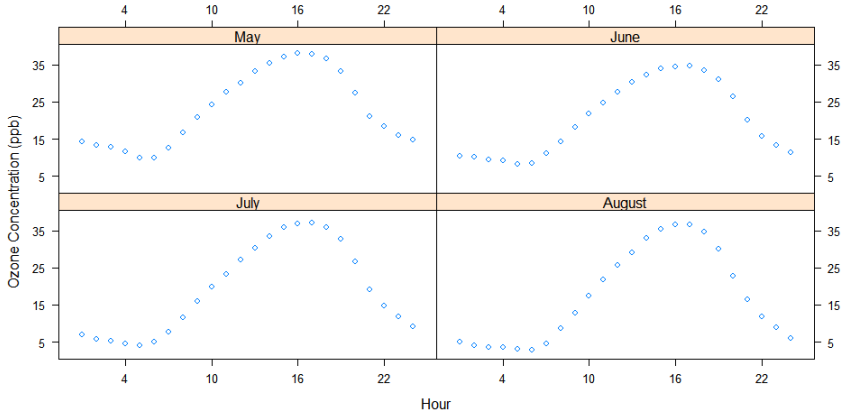


Figure 1.1: Diurnal trend during summer months of 2004–2008 at Chilliwack (BC, Canada). Each hourly concentration is averaged over the 5 years and the days of the month.

mer months in Chilliwack, BC. Each hourly ozone value is the average ozone measurement for that hour in the entire month over the years 2004–2008. The ozone data are recorded in units of parts-per-billion (ppb). Missing hourly data are filled with the average value of available data for that particular hour, e.g. the missing 2 p.m. value of August 3rd, 2005 is filled with the average of the available 2 p.m. values of August 3rd from other years. Data for any particular hour on a specific date are available for at least 2 years among 2004 to 2008. The same plot based on data from 1984–2008 shows a very similar diurnal trend. As another example, Figure 1.2 shows daily day-time (8 a.m. to 8 p.m.) mean ozone level for the years 2004–2008 and the average of these 5 years. One may notice that for Chilliwack, the annual day-time average peaks during the spring in some years, but when averaging the daily values over the 5 years, the diurnal fluctuations are smoothed out and the annual peak is evident during summer.

Figures 1.3 and 1.4 are time series plots of ozone concentrations during an ozone episode in the summer of 2006. The locations in Figure 1.3 are

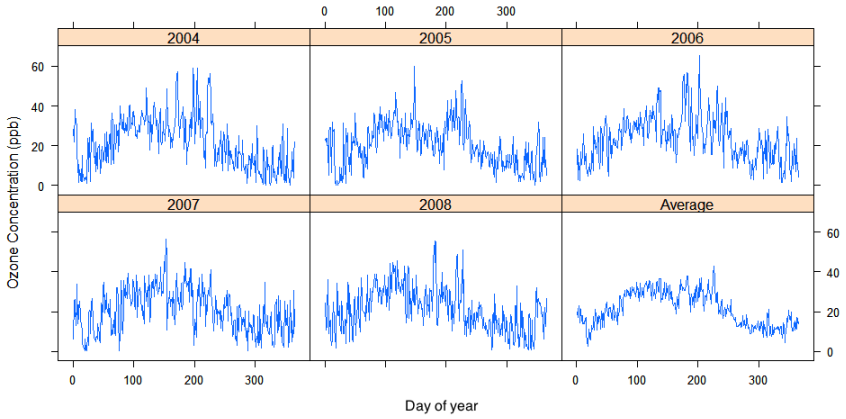


Figure 1.2: Day-time mean ozone level against day of the year at Chilliwack. The mean is averaged over values from 8 a.m. to 8 p.m.

contained within an area of 8km radius, while the locations in Figure 1.4 are much further apart. This type of plot is useful to visually assess how closely correlated are the ozone concentration levels at different locations.

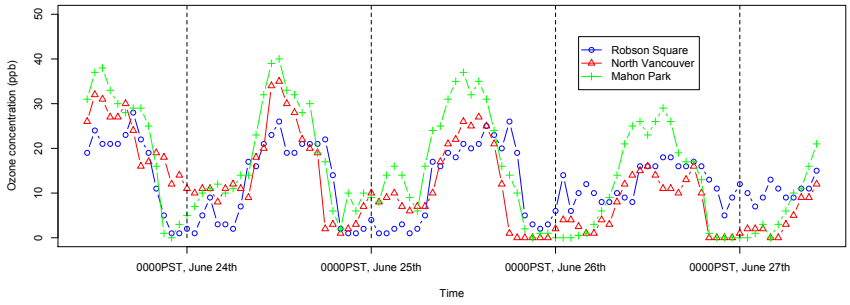


Figure 1.3: Time-series of observed ozone concentrations at 3 close locations during the time-period 1100PST, June 23rd to 1000PST, June 27th of 2006. The vertical dashed-lines indicate the hour 0000PST of each day.

1.1.2 The CMAQ Modelling System

This section describes in simple and general terms, the inner working of the CMAQ modelling system. In Chapter 2, I will provide a more detailed

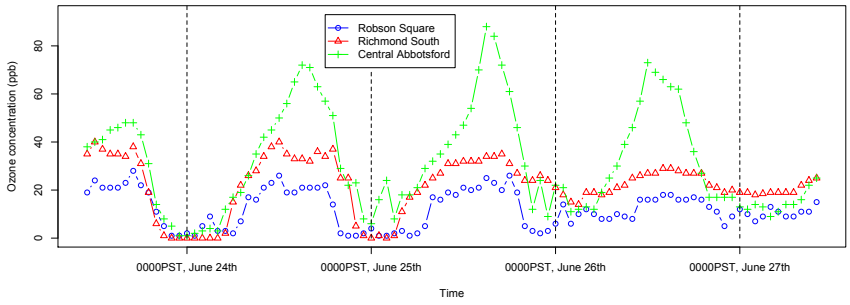


Figure 1.4: Time-series of observed ozone concentrations at 3 well-separated locations during the time-period 1100PST, June 23rd to 1000PST, June 27th of 2006. The vertical dashed-lines indicate the hour 0000PST of each day.

description of the conditions and settings of CMAQ runs that are relevant to this research.

As a numerical model, CMAQ is essentially a system of differential equations which are integrated given initial and boundary conditions. Hence to implement a CMAQ model run, one requires various inputs for these initial and boundary conditions of the pollution process. One important input is provided by the emission model Sparse Matrix Operator Kernel Emission (SMOKE) (SMOKE v2.5, University of North Carolina, 2012). Given an annual total emission for a geographical region, the SMOKE modelling system distributes this emission figure into spatial grids and time periods: it provides estimates of the types and amounts of pollutants (reaction precursors) released into the atmosphere, and this information is listed for each relevant geographical grid cell, at varying height, over given time periods.

For example, when we wish to estimate the pollutant types and amounts released by household heating during winter, we would first apply appropriate sampling methods to obtain the number of residences within a geographical region. This is our pollution source, then the follow-up procedure is described in Boubel et al. (1994, Section 6.4):

1. Identify what gases are produced from home heating, typically CO, CO_2, NO_X and CH_4.

2. Collect household fuel consumption figures from dealers, utility providers

and so forth. Without such data, one may apply established distribution models of fuel type and consumption amount to produce an estimate.

3. Examine the reference literature to decide on relevant emission factors for the given consumption figure, such as weight of pollutant per volume of fuel burned. A popular reference on emission factors is "Compilation of Air Pollution Emission Factors" by the U.S. Environmental Protection Agency[1].

4. There exist models that simulate consumption behaviour over time, allowing us to estimate the time of emission.

5. Calculate the total emission produced within this region at a certain time period.

The CMAQ system has modelling uncertainties from various sources; the emission inventories provided by SMOKE account for a large portion of said uncertainties. This is perhaps unavoidable given its task of estimating results of human activities.

Other inputs feeding into the CMAQ model include land surface information and meteorological conditions such as wind speed and direction, ambient temperature, pressure and solar radiation intensity. These input data are produced by the Weather Research and Forecasting (WRF) model (WRF v3.1, Skamarock et al. (2008)).

Finally, the role of the "chemical component" in the air pollution model is to simulate atmospheric chemical reactions. The emission model tells us the types and amounts of gases entering into the chemical reactions. Using the given inputs and working in conjunction with the meteorological model, the chemical reaction model simulates the systems of chemical transformations in the ozone process. At the end, a CMAQ model run produces output in the form of an average value over a geographical grid cell at a point in time. Thus, hourly averages need to be computed, usually by simple averaging of a series of outputs. Moreover, the model typically requires input every

[1] http://www.epa.gov/otaq/ap42.htm

10 minutes (other time intervals are also possible). Initial and boundary conditions may be updated every hour, and inputs for shorter time intervals would be interpolated within the model during simulation.

In summary, CMAQ is a complex machine comprised of purpose-specific models (WRF, SMOKE, etc.) that operate interactively to model an atmospheric air pollution process. Given its complexity, model uncertainties and errors are an unfortunate but inherent reality of CMAQ. In view of the use of CMAQ as a "reference guide" to the design of ozone-related policies (introductory paragraphs of the chapter), it is imperative to evaluate the modelling capabilities of CMAQ.

The topic of AQM evaluation has garnered substantial attention over the past few years. Dennis et al. (2010) and Galmarini and Steyn (2010) contain extensive overview and discussion of this topic. The book series "Air Pollution Modelling and its Application" (Springer Books) provides up-to-date summaries of new research relating to all aspects air-quality modelling every 18 months. One research area covered in this book series is the topic of AQM evaluation.

1.2 The Problems with "Usual" Means of AQM Evaluation

The most popular means of AQM evaluation is to analyze the statistics of point-to-point differences between the AQM output and corresponding (in space and time) physical observations. Dennis et al. (2010) listed common evaluation statistics such as *Mean Bias Error*, *Root Mean Squared Error* and *Correlation*. Willmot et al. (1985) suggested that bootstrap methods can be applied to obtain the confidence interval and assess the significance of observation-model difference statistics such as RMSE. Under the context of climate model evaluation, Preisendorfer and Barnett (1983) compared model outputs and physical data as two "swarms" of data points in a common euclidean space. Geometric properties such as the distance between two swarms' centroids, differences in their radial scales and space-time evo-

lutions are proposed as viable measures of a model's accuracy. Authors then outlined sampling procedures to obtain the statistical significance and power of proposed model accuracy measures.

However, Dennis et al. (2010) pointed out that while direct data comparisons can be useful to a certain degree, they "provide little insight" on the deficiency and behaviour of AQM simulated pollution fields. There are further problems stemming from the point-to-point comparisons.

AQM output and ozone observations are products of different processes driven by their own space-time mechanisms. AQM ozone is driven by a system of differential equations that describe specific ozone-related photochemical and atmospheric processes. In case of CMAQ, these equations are integrated over the inputs from WRF (meteorology) and SMOKE (emission). Ozone observations are measurements of the real-life ozone processes that occur during observed meteorological and chemical precursor conditions (Dennis et al., 2010). The physical monitoring stations are sparsely and irregularly located across a large spatial domain (details in Chapter 2). Hence observations do not necessarily represent initial and boundary conditions in the same way as an AQM modelled process.

The AQM ozone and observations are further differentiated by the fact that the computer model outputs are spatial-averaged concentrations (discussed in Section 1.1), whereas the physical observations are ozone measurements taken at point locations. In other words, AQM output and observations each capture ozone process on a different spatial scale.

The quality of physical observations is invariably degraded by stochastic measurement errors. The error source of AQM outputs are characterized by inadequacies in model inputs (WRF and SMOKE outputs) and its deficiencies in emulating the behaviour of key atmospheric processes. This implies that AQM modelled ozone and observations have different stochastic structures in relation to the underlying *true* ozone. For example, let s and t denote a location and time, x denotes a set of atmospheric conditions at s and t, and let the *unknown* true ozone at (s, t, x) be denoted by $O^t(s, t, x)$. Then the observed ozone is $O^t(s, t, x) + \varepsilon$: the true underlying ozone plus random measurement error, whereas the AQM modelled ozone $O^c(s, t, x)$

relates to true ozone by $O^t(s,t,x) = O^c(s,t,x) + \delta(s,t,x)$, where $\delta(s,t,x)$ is a random process representing the AQM modelling deficiency and often a nonlinear function of (s,t,x). Such statistical formulation of the AQM ozone and observations are based on the more general model-measurement relationship proposed by Kennedy and O'Hagan (2001), which is applied in later works such as Fuentes and Raftery (2005) and Higdon et al. (2008).

Given the above formulation, the observation-model difference is a random non-linear process $\delta(s,t,x)+\varepsilon$. Hence, an informative model evaluation should tell us something about the pattern and behaviour of random processes like $\delta(s,t,x)+\varepsilon$, these are information that point-to-point comparison summaries such as RMSE cannot provide.

In summary, AQM outputs and observational data are generated by two space-time ozone processes with discrepant physical and stochastic structures. The point-to-point comparison of two datasets only serves to inform the deviations in their output values, not their difference as individual ozone processes. Without a more insightful process-level understanding of the two ozone processes, close agreement based on point-to-point comparison should be deemed "fortuitous" (Dennis et al., 2010).

This research is motivated by the concerns and topics proposed by the Air Quality Model Evaluation International Initiative (AQMEII, Galmarini and Steyn (2010)) and Dennis et al. (2010). The literature discussed extensively the aforementioned problems of direct output-observation comparison, and highlighted the importance of evaluating the ability of a computer model such as CMAQ to emulate the interacting atmospheric processes within a space-time air pollution system.

1.3 Research Topics and Objectives

In the most general terms, my research objective is to develop statistical methods of AQM evaluation that are more informative than direct observation-model comparison. These "informative" methods should be able to provide useful insights into the way AQM and physical observations differ as ozone processes, and to identify possible sources of AQM deficiency. In addition,

the evaluation methods should correct or at least account for the fact that model output and observation are generated from individual processes with discrepant physical and statistical properties. The statistical analyses in this thesis are based on data (model output and physical observations) related to ozone pollution and the CMAQ system. However, the general ideas and statistical methods are intended to be applicable to other air pollution problems and AQM evaluations.

The AQM evaluation methods proposed in this thesis are based on the *analysis and modelling of spatial-temporal ozone features*. An *ozone feature* is a data component/mode that captures certain space-time structure of the underlying ozone process and/or recover non-trivial amount of data variation. One example of ozone features is the spatial or temporal ozone means. Suppose there is a space-time ozone dataset $\mathbf{O}$ of dimension $t \times n$, t and n being the number of hours and locations in a dataset. The column means of $\mathbf{O}_{t\times n}$ are the *spatial field of temporal ozone means*: ozone averaged across time at each location. The row means of $\mathbf{O}_{t\times n}$ are the *time series of spatial ozone means*: ozone averaged across space at each hour.

Analyses in later chapters will show that ozone features allow for statistical or real-world interpretability. In addition to space-time ozone mean, an ozone feature may also capture some dynamic patterns of ozone advection (atmospheric ozone transport).

In this thesis, I propose and implement two general methods of *feature based* evaluation of AQM against the observations:

1. The first approach is to compare the ozone features between AQM outputs and physical measurements, and analyze how the two ozone data differ in their underlying space-time structures. As mentioned, an ozone feature captures either the mean structure or some dynamic patterns of ozone advection. The advection patterns reveal the most fundamental mechanisms of the underlying atmospheric process, hence the comparison of such advection features is a means of *process level* AQM evaluation.

   I also propose to model the ozone feature differences to identify and

analyze the statistical associations between specific AQM inputs and feature differences. Here, the AQM inputs are variables representing the background meteorology, ozone precursor emission rates and atmospheric concentrations. This is a way of understanding how the components of an AQM (weather, emission, etc.) are associated with its deficiencies in capturing the physical field.

2. The second approach is to build statistical ozone feature models for both AQM and observations, then compare the two statistical models as another means of process-level AQM evaluation. For instance, one can use the fitted models to produce the AQM features and the observed features *given the same* regional meteorology and ozone precursor pollution. These features can then be compared and checked for the significance of their differences in space and time. Another evaluation can analyze how two ozone features react to the same variations in background atmospheric conditions. This is a process level comparison of the stochastic properties of AQM ozone and physical process.

This thesis will present a coherent set of statistical analyses that argue the following claim: an informative and "big-picture" AQM evaluation can be achieved through the analysis and modelling of ozone features.

1.3.1 Topics of Ozone Feature Analysis and Modelling

To successfully implement the proposed evaluation approaches, I will need to develop the necessary statistical tools and framework that: (1) extract ozone features from space-time ozone data, (2) model individual ozone features and model the complete space-time ozone field through these features. These are my other research topics in addition to the statistical AQM evaluation.

In this research, methods of Principal Component Analysis (PCA) are used to decompose space-time ozone data into ozone features. One PCA-related topic is to determine the number of ozone features that are *meaningful* for statistical analysis. A "meaningful" ozone feature should be a data

mode that either can be interpreted statistically, or captures the underlying physical process that created the ozone field. At the very least, a "meaningful" feature should recover a non-trivial amount of data variation. By determining the number of meaningful ozone features, one is able to answer whether a complex space-time ozone data can be analyzed using a few ozone features, i.e., simpler data components.

Furthermore, there are several PCA-related complications that should be addressed to ensure convincing implementation of the proposed feature-based AQM evaluation. One such complication is the ozone feature "degeneracy" (North et al., 1982), where a feature's order of extraction and/or its mode of variation are "mixed" with other features. Failure to address this and other PCA-related complications may result in the event where one feature from AQM is evaluated against an entirely different feature from the observations, resulting in erroneous conclusion about the AQM performance. Thus, it is important to determine a specific PCA procedure that ensures, or at least maximizes *feature correspondence* during model evaluation.

Aforementioned PCA related topics will be studied in Chapter 3 using both the CMAQ outputs and synthetic ozone data. These topics are all part of the first step in the framework of ozone feature analysis and modelling: the extraction of ozone features. The second step is to develop statistical models for individual spatial-temporal ozone features, and this will be done in Chapter 4.

The proposed ozone feature models are variations of Gaussian Processes (GPs) driven by the background atmospheric and chemical precursor conditions. As the reader will see, the structures of ozone features can be highly non-linear, making the task of model estimation an interesting and challenging one. One such challenge is to determine the ideal design and composition of model covariates; a significant portion of my modelling effort is focused on this topic. Typically for a statistical research, once the model is estimated, one need to implement the proposed model using appropriate data, which in this case are CMAQ-WRF-SMOKE outputs. The purpose is to carefully scrutinize the modelling and forecasting capabilities of individual ozone feature models through a series of goodness-of-fit tests, model diagnostics and

exercises in ozone forecasting.

This framework of *ozone feature extraction, analysis and modelling* forms the core statistical methods on which my proposed AQM evaluation is based. Although ozone feature models are developed as a means to AQM evaluation, the developed methodology is potentially an useful contribution in itself: it is a novel and computationally efficient means of modelling a complex space-time air pollution field.

1.3.2 Relation to Existing Model Evaluation Projects

As discussed earlier in this section, Galmarini and Steyn (2010) and Dennis et al. (2010) pointed out the importance of an air quality modelling system to emulate the interacting atmospheric processes within a space-time air pollution system. They term this type of model evaluation as *Diagnostic Evaluation*. The authors also categorized three more types of AQM evaluation. *Dynamic Evaluation* analyzes an AQM's ability to model the changes in pollution concentration due to the fluctuations in meteorological conditions and emissions. *Operational Evaluation* refers to "generating statistics of the deviations" between the AQM outputs and corresponding observations, and examination of the results based on a few "selected criteria". *Probabilistic Evaluation* proposes to model the AQM outputs and/or observations as random processes following certain probability density functions (pdfs), then the estimated pdfs are used to carry out various evaluations of AQM against observation. My proposed AQM evaluation approach may be viewed as a mixture of above categories of model evaluation done though a probabilistic framework.

This research is also closely related to the works of Steyn et al. (2011) and Steyn et al. (2013). In these works, the authors used CMAQ to simulate the space-time ozone fields during ozone episodes in the Lower Fraser Valley (LFV), British Columbia (BC). Among other things, the authors carried out point-to-point comparison of CMAQ outputs with available observations, and identified the precursor-sensitivities of local ozone fields within LFV. The statistical analyses in this thesis deals exclusively with ozone data

during LFV ozone episodes, and all my dataset are those used in Steyn et al. (2013). More detailed discussions of Steyn et al. (2011), Steyn et al. (2013) and other related LFV ozone papers will be done throughout the thesis.

In the end, my goal is to generate useful statistical methodologies and analyses that can be incorporated into the overall body of works in the fields of AQM evaluation and LFV regional air quality study.

1.4 Literature Review

The literature review is organized according to the main steps of the thesis research: (1) PCA of space-time processes, (2) methods of modelling the non-linear processes that produce both computer model outputs and physical observations, and (3) existing AQM evaluations based on the comparison of data features.

1.4.1 PCA and Extraction of Data Features

Principal Component Analysis (PCA) was originally introduced by Pearson (1902). One earliest application of PCA in the field of atmospheric and climate science can be traced to Lorenz (1956), where sea level pressure (SLP) data were decomposed into Empirical Orthogonal Functions (EOFs) in space and time. The term "EOF" used in this paper is widely adopted in atmospheric science.

The application of PCA/EOF has since gained prominence for the purpose of data decomposition, where the central idea is to decompose a non-linear climate system (sea level pressure, surface temperature, etc.) into independent physical modes of variation. A rich body of literature explores various topics stemming from the PCA of space-time physical processes. The general topics relevant to this research include (1) the interpretation and selection of data features (Richman, 1986; Preisendorfer, 1988), and (2) estimate error associated with EOF-decomposition, which relates to the separability and identifiability of data features (North et al., 1982; Monahan et al., 2009). These two topics and further references will be exten-

sively discussed in Chapter 3 on PCA. Furthermore, a useful summary of PCA/EOF-decomposition, their related issues, extended methodologies, and applications in atmospheric science can be found in the books by (Storch and Zwiers, 1999, Chapter 13) and Jolliffe (2002)) and review articles (Bjornsson and Venegas, 1997; Hannachi et al., 2007).

More recently, methods of PCA have been applied to study the spatial and temporal patterns of ozone process. These papers deal with an ozone process at a global or continental scale. Orsolini and Doblas-Reyes (2003) studied the spatial pattern of leading EOFs decomposed from monthly column ozone data observed by Total Ozone Mapping Spectrometer (TOMS) during the spring months between 1948-2000. The spatial field covers the Euro-Atlantic sector (20^o to 90^o latitude and 60^o to -90^o longitude) and the data are measured at 500mb geopotential height. From the patterns of leading EOFs, the authors were able to identify the pressure system that is associated with the pattern of each EOF. They found that the most important pressure system is the North Atlantic Oscillation (NAO), and other leading weather patterns are the Scandinavian, east Atlantic and European Block. Principal Component (PC) time-series were also studied for long-term trends. The purpose of this paper is to study the link between ozone EOFs and known climate patterns.

Jrrar et al. (2006) carried out similar analysis using outputs from Chemical Transport Model (CTM) SLIMCAT and identified 5 climate patterns from the spatial plots of 5 leading ozone EOFs. Camp et al. (2003) implemented PCA on column ozone data across the tropics, which are measured by both TOMS and Solar Backscatter Ultraviolet (SBUV) instruments. The key tropical oscillation patterns are identified from the ozone EOFs. The above mentioned ozone PCA literature has since been extensively cited over the past 10 years.

Similar ideas have been used in other fields. For example, Liu et al. (2003) implemented a modified Singular Value Decomposition (SVD) technique on an affymetrix microarray, or gene expression data. By analyzing the decomposition, they isolate a vector of data features that enables them to order gene types according to the level of data variation attributed to

each genome.

There are certainly other ways of decomposing and visualizing high-dimensional data: spectral decompositions (Fourier transform, etc.), and neural networking techniques like self-organizing maps (Kohonen, 1982, 1990) and multidimensional scaling (Borg and Groenen, 2005). Methods of PCA or EOF-decomposition are use here because they are widely adopted in both statistics and atmospheric science. Aforementioned literature are simply well-known examples among an extensive collection of works that demonstrate the use of PCA for extracting structure/dynamic modes from high-dimensional datasets. Furthermore, the usefulness of PCA methods are also reiterated in this particular research.

1.4.2 Emulation of Non-linear Computer Models and Physical Processes

My proposed statistical analysis involves the modelling of AQM outputs. AQMs such as CMAQ are referred to as "numerical models" or "computer models" in the sense that the mathematics involves numerical solutions of governing atmospheric dynamic equations and chemical kinetic equations, and only the input conditions are required to implement a model run that produces a deterministic output. In contrast, a "statistical model" is built around the idea of modelling randomness with probability functions, and sample data are needed to fit a model to output. On the surface, the nature of CMAQ makes statistics inapplicable; after all, statistics requires randomness.

For a complex computer model, a model run at every possible input value is impractical at best. When we implement a model run for given input, the output is deterministic, but outputs at untried inputs remain unknown. In this sense the numerical model output follows a stochastic process because of these uncertainties in the output. This is the historical reasoning behind treating model output as a realization of a stochastic process (Sacks et al., 1989). Over the years, statistical methods have been developed based on such a formulation, for the purpose of using computationally cheap statisti-

cal models to emulate the computer model outputs, and the real processes they try to estimate.

As mentioned, Sacks et al. (1989) were among the earliest authors to treat a deterministic model output as a realization of a random process. They suggested that the deterministic model output $Y(\mathbf{x})$ given covariate set $\mathbf{x}$ is a realization from a random function such as

$$Y(\mathbf{x}) = \sum_{j=1}^{k} \beta_j f_j(\mathbf{x}) + Z(\mathbf{x}). \tag{1.1}$$

The random process $\mathrm{Z}(\mathbf{x})$ is assumed to have zero mean and covariance $\sigma^2 R(\mathbf{x}, \mathbf{x}')$, where $R(\mathbf{x}, \mathbf{x}')$ is the correlation measure between computer "input sites" $\mathbf{x}$ and $\mathbf{x}'$. In such a function, $Z(\mathbf{x})$ models the random deviation from the regression function $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{f}(\mathbf{x})$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ and $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}))^{\mathrm{T}}$. They further put forth the idea that the random process $Z(\mathbf{x})$ (hence $Y(\mathbf{x})$) is governed by a Gaussian Process (GP). As the name implies, the assumption is that the random process follows a Normal distribution.

This fundamental approach has since gained prominence, and it has been studied and refined for a wide range of applications. Kennedy and O'Hagan (2001) introduced the idea of a joint-GP that combines deterministic outputs and physical data in a method called Bayesian Melding. The application is to calibrate the computer output using its correlation structure with the observed physical data. This idea of "melding" computer model output and physical observations has been adapted into more complex forms to tackle the problems of air pollution as spatial processes (Fuentes and Raftery, 2005; Liu, 2007). In these two works, the "grid-cell average" computer output is set to equal a weighted sum of latent point processes, where the weights are defined by exponential kernel functions of the distances between the locations of latent process and the grid-cell centroid. Such function "downscales" the spatial scale of computer output to match that of the observation.

Berrocal et al. (2009) and Zidek et al. (2012) ventured one step further

by implementing particular forms of hierarchical regression model whose coefficients follow GPs and are allowed to vary in both space and time, thus modelling the air-pollution as spatial-temporal processes. All the above mentioned literature implemented their model formulation using variations of a Hierarchical Bayesian Algorithm. Lindstrom et al. (2014) proposed another hierarchical regression model that contains a temporal basis function, this function is either assumed to be known or decomposed from data, and it is scaled by a spatial coefficient that follows GP.

Berrocal et al. (2012) evaluated of the temperature outputs from the Regional Climate Model (RCM), where the data are quarterly mean temperatures from 1962 to 2002 in South Central Sweden. In this paper, information from observation data are scaled, using statistical models, onto the "grid-box level" for comparison with RCM outputs. The spatial scaling is done using both the space-time downscaler model in Berrocal et al. (2009) and an upscaler model in Craigmile and Guttorp (2011). The statistical model outputs are Bayes estimates of spatially scaled quarterly mean temperatures.

Using the the methodologies of Sacks et al. (1989), Gao et al. (1996) showed that a GP model can be applied to model observational ozone concentrations. They analyzed daily ozone data from Chicago for the period 1981 to 1991, and found that a properly designed GP-based model can be a capable modeller of process-driven temporal ozone processes. The main purpose of the paper is to correct the trend over years for changing meteorology, and to assess the impact of regulatory initiatives. Dou et al. (2010) also applied GP methodologies in modelling temporal ozone patterns: the authors used a form of time-series model whose coefficients are modelled as GPs. Cooley et al. (2007) modelled the observed "extreme precipitation return" in Colorado using a Pareto-based distribution. To account for spatial non-stationarity, the Pareto-parameters are modelled as spatial Gaussian Processes. The latter two works implemented their models using variations of a Hierarchical Bayesian Algorithm.

The aforementioned literature is only a selection among a rich collection on theories and applications relating to GPs. The popularity and diversity

in the applications is a testament to GP's capability to interpolate and extrapolate systems of complex non-linear functions. Progressive refinements made to the basic framework of a GP model also point to its flexibility that allows for extensive fine-tuning and elaboration.

Unlike aforementioned literature, I do not model ozone data directly. Instead, the spatial and temporal *ozone features* are modelled as multivariate GPs. More specifically, they are modelled as GPs driven by the background processes conducive to a space-time ozone process, e.g., meteorological conditions, chemical precursor emission rates and ambient concentrations.

1.4.3 Feature-based AQM Evaluation

My literature review reveals that although there are earlier precedents, PCA-based model evaluation is not a widely and frequently adopted technique. This sentiment is also reflected in Eder et al. (2014).

Preisendorfer and Barnett (1983) is among the earliest works that proposed the idea of computer model evaluation based on data decomposition. As mentioned in Section 1.2, Preisendorfer and Barnett (1983) proposed a few statistical summaries of point-to-point data differences. The authors also mentioned a model evaluation approach where the data of *observation-model differences* are decomposed, and the leading EOF and PC are visually assessed to extract useful information about model deficiency. The paper then evaluated General Circulation Model[2] (GCM) outputs for January sea-level pressure field. However, this PCA-based model evaluation is mentioned only briefly as a possible complementary analysis to point-to-point data comparison.

Cohn and Dennis (1994) and Li et al. (1994) later proposed PCA-based AQM evaluations with more extensive implementations and discussions than the one presented in Preisendorfer and Barnett (1983). Cohn and Dennis (1994) used PCA to evaluate the capability of Regional Acid Deposition Models (RADMs). The model output for various aerosol species is evaluated against corresponding observations collected over Eastern United States at

[2]An early version from National Centre for Atmospheric Research.

high altitude (1000-1500 meters) during August-September, 1988. The data are averaged over space: each column is a time-series data of one aerosol specie. The RADM outputs are further separate into sulfur system (O_3, SO_2, H_2O_2 and SO_4^{2-}) and nitrogen system (O_3, NO, NO_2, etc.). The two groups of outputs and observations are then decomposed through PCA and their PCA results compared. Comparison metrics include percent variation explained by leading PC loadings, and angles between the PC spaces. The results indicated systematic RADMs deficiency in the nitrogen system. Through the scatter plots of O_3 outputs against NO_2 outputs, author pointed out the possibility of RADM producing fewer molecules of O_3 per NO_2 photochemical cycling.

Li et al. (1994) used PCA to evaluate the Eulerian Acid Deposition and Oxidation Model (ADOM) against observations collected from the Eulerian Model Evaluation Field Study (EMEFS). Through PCA, the modelled chemical process is decomposed into three distinct components simulating the process of chemical aging/transport, diurnal cycle and area emission. The resulting PC scores, which are in the form of a time series, are compared between ADOM and EMEFS to identify specifics of the computer model's temporal bias.

Fiore et al. (2003) evaluated AQMs' abilities to model space-time ozone processes. In this paper, two models are evaluated: Multiscale Air Quality Simulation Platform (MAQSP) and global GEOS-CHEM model at two spatial resolutions. The ozone EOFs from these AQMs are compared against corresponding EOFs of observations, both visually and through correlation statistics such as linear slope and R^2. Furthermore, from the spatial variations of the leading EOFs, the authors discussed the possible wind patterns responsible for ozone transport across the Eastern U.S. The evaluation revealed that all three models captured similar east-west spatial feature shown in the observed EOF, while both resolutions of GEOS-CHEM misplaced a midwest-northeast EOF. All three models also shown to capture the general patterns of leading temporal features from the observations.

Dennis et al. (2010) also contains an review of AQM evaluations based on data decomposition. For example, Hogrefe et al. (2000) and Porter

et al. (2010) applied spectral decomposition to longterm time-series data of CMAQ output and O_3 observations, then compared the decomposed spectral bands of varying frequencies, e.g., diurnal, seasonal and long-term fluctuations.

A more recent work is Eder et al. (2014), published during the writing of this thesis. In this paper, PCA methods are used to evaluate CMAQ outputs for SO_4^- and NH_4 against weekly observations from the Clean Air Status and Trend Network (CASTNet) over 2001-2006. Moreover, PCA is implemented on the *difference* of CMAQ outputs and observations, not individual data. The authors identified some systematic features of CMAQ-observation differences. For example, the third SO_4^- spatial feature revealed five high-elevation locations in the eastern U.S., and corresponding temporal features indicated a seasonal cycle where CMAQ under-predicted the concentrations during late summer months and over-predicted for rest of the year. Authors found that the Meso-scale Model (MM5, weather model of CMAQ) under-predicted relative humidity and over-predicted solar radiation in high-elevations. These weather model deficiencies are shown to be significant from Mann-Whiteney nonparametric tests between MM5 outputs and weather observations. In the end, author suggested that the aforementioned CMAQ deficiency with high-elevation SO_4^- modelling is caused by issues with the parameterization of clouds in CMAQ.

Another recent example of PCA-based CMAQ evaluation can be found in Marmur et al. (2009). This paper instead applied Positive Matrix Factorization (PMF) to the CMAQ output for numerous chemical species and associated observations. As with earlier literature, comparison matrices such as "percent data variation explained" and PC scores time-series were compared between CMAQ outputs and observations.

There is also a general method of climate model evaluation called optimal fingerprinting. Linear regression is used to compare climate observations (the responses) with model outputs under some external climate forcing (regressors). Usually the response and regressors are data features obtained from decomposition (Hasselmann, 1993; Allen and Tett, 1999). These data features, representing significant departures from normal climate variations,

are called "fingerprints". The extensive literature on optimal fingerprinting includes Hobbs et al. (2015). Here, the authors evaluated climate models of Antarctica ice coverage (a proxy for climate change), where above mentioned linear regression is used to analyze the seasonal feature/fingerprints of observation-model deviations.

Other works used combined analyses of PCA and cluster analysis on *meteorological data* to study measurement-model discrepancy. Beaver et al. (2010) used EOF to categorize daily physical observations and corresponding model outputs into clusters of weather patterns. They then assessed how often the outputs and observations match in this categorization. The clustering is based on hourly wind speed data: it is recorded in vector components u and v, and the measurement stations cover the bay-area of California. The hourly wind data are first concatenated into daily data, then the PCA-compressed data are clustered based on the criteria of minimum sum of squared errors. The defining wind pattern within each cluster is visually interpreted and defined. In the end, the authors recorded the number of days where observation and model output were assigned to different clusters, and analyzed the meteorological patterns during the days that are mis-categorized. The article concluded that the instances of observed mis-categorization (model inadequacy) are consistent with what they already knew from experience regarding the behaviour of models being analyzed.

Ainslie and Steyn (2007) ventured one step further. The authors implemented EOF-decomposition and cluster analysis of mesoscale wind data for regions around LFV. They then defined four types of synoptic wind patterns associated with LFV's regional ozone exceedance, where the "ozone threshold" is defined by CWS mentioned in the introductory paragraphs (page 1). Reuten et al. (2012) further applied results from Ainslie and Steyn (2007) to forecast the frequencies and types of regional ozone exceedance for the future time period 2046 to 2065. In this thesis, the four LFV wind patterns identified in Ainslie and Steyn (2007) will provide an important reference point in the ozone feature analysis and CMAQ evaluations to be presented in Chapter 3 and 5. The key results and conclusions in this paper will also be discussed in detail in Section 3.1.

1.5 Novelty of Proposed AQM Evaluation Methods

In this section, I will discuss in the most general terms, the novelties and potential contributions of my proposed AQM evaluation approaches. The goal of this thesis is then to present a coherent set of statistical analyses that demonstrate the usefulness of my proposed evaluation methods.

In the existing literature, the differences in data features (in the form of PC scores and loadings) are compared using correlation measures, RMSE or linear regression (optimal fingerprinting). To the best of my knowledge, no previous attempt has been made to model the feature differences using GP or other non-linear functions, which I propose here. As this thesis will demonstrate, ozone features have highly non-linear structures, which makes my proposed non-linear modelling a practical improvement over existing evaluations.

The aim of modelling data feature is to identify any statistical association between the observation-AQM feature differences and specific input conditions of AQM run. Using the feature difference models, one can further analyze how the feature differences change with variations in AQM inputs. I will show later in this thesis that the modelling of feature differences can reveal useful and specific insights into an AQM's capability.

My second proposed method (Section 1.3) is also a novel means of feature-based AQM evaluation. In this method, statistical ozone feature models are used to predict the AQM feature and the observation feature under the same background conditions, e.g., weather and precursor pollution. These "same-condition" features are then compared in space and time, and the significance of their feature differences are assessed. In essence, this proposed evaluation compares the stochastic properties of two air pollution processes.

The complexities of an AQM such as CMAQ dictate that one cannot run AQMs under the same real-world conditions that generated the observations. This is evident from literature regarding the CMAQ modelling of LFV air pollution, such as Reuten et al. (2012), Steyn et al. (2013) and Ainslie

et al. (2013). Therefore, the second proposed evaluation can provide useful statistical answers to the question: "can AQM produce results that are statistically similar to observations after correcting for deviations in basic background conditions?"

1.6 Thesis Structure

Chapter 2 will describe the data used in this thesis. Specifically, they are model outputs from CMAQ, WRF and SMOKE, as well as physical observations on air pollution and accompanying meteorology. I will also discuss the specific set-up of CMAQ modelling runs that produced the available data.

As discussed in Section 1.3, I will develop the necessary statistical tools for AQM evaluation. In Chapter 3, I will study the PCA-related topics that are crucial for an informative and defensible AQM evaluation. In Chapter 4, I will develop statistical models for individual ozone features. Specifically, I will estimate the exact formulation of the models, diagnose relevant statistical assumptions, and evaluate the prediction capability of these ozone feature models. Furthermore, I will analyze whether a complete space-time ozone fields can be modelled using combinations of ozone features.

Chapter 5 and 6 bring everything back to my original research motivation: the statistical evaluation of AQMs, which in this case, the CMAQ. The two general AQM evaluation approaches proposed in the beginning of Section 1.3 will be developed and implemented individually. Combined insights and modelling methodologies developed in Chapters 3 and 4 will be applied while evaluating CMAQ output against the observations.

Chapter 2

Data

Space-time ozone fields are either produced by the CMAQ modelling system or physically observed at monitoring locations. In addition to ozone, this research also uses space-time data (either from computer model or observation) of variables representing meteorological conditions, ozone precursor emission rates and surface level concentrations.

Section 2.1 introduces the sources of data and provide some background. Section 2.2 presents the way CMAQ outputs and observation data are processed for CMAQ evaluation in Chapter 5 and 6. I refer to data presented in Sections 2.1 and 2.2 as the *real* data to differentiate them from the simulated or synthetic data that are discussed in Appendix A.1 and used in B.1. Simulated ozone processes are useful for answering statistical questions that are difficult, or impossible, to answer unequivocally using the real data.

2.1 Air-quality Model Output and Physical Observations

Data undergo extensive numerical processing during statistical modelling. The details behind each data-processing procedure will be discussed at appropriate stages of the statistical analyses. Section 2.1 simply informs the reader of the original source of all my data. The computer model outputs[3] used in this thesis are part of those used in Steyn et al. (2011) and Steyn et al. (2013).

[3] All available CMAQ outputs are calculated using the `BORA` server at the University of British Columbia.

2.1.1 Data from Computer Models

CMAQ models ozone on a regular grid system, in which outputs are presented in the form of *hourly grid-cell averaged* ozone concentration in units *parts-per-billion* (ppb) (Byun and Schere, 2006). The geographical size of a grid cell, commonly referred to as the CMAQ resolution, is defined by users. The coarsest resolution available for this study is a single grid cell size of 36km $\times$ 36km with 93 cells in the east-west direction and 95 in the north-south direction. Smaller grid cell sizes available are 12km $\times$ 12km and 4km $\times$ 4km, with grids of 70×89 (E-W$\times$N-S) cells and 172×103 respectively. In this regional ozone study, I use CMAQ outputs at 4km $\times$ 4km resolution.

In addition to ozone, CMAQ outputs the concentrations of 124 photochemical compounds. Furthermore, CMAQ models the air-pollution fields at 48 different atmospheric heights. My analysis of interest is the surface level ozone.

Chemical precursor data came from the CMAQ and the SMOKE models. Each model output represents one major source of photochemical precursor: precursors already present in the atmosphere (CMAQ output) and precursors "newly" emitted into the atmosphere (SMOKE output).

The statistical analysis in this thesis will use data on the emission rates and antecedent concentrations of NOx (oxides of Nitrogen) and VOC (volatile organic compounds). NOx is the sum of NO and NO_2 data, while VOC data are created by adding the scaled values of 16 families of volatile organic compounds. The reactivity scale for each compound is calculated as the ratio between the Carbon-Bond 5 (CB5) reaction rate of that compound and the median of the 16 reaction rates (Yarwood et al., 2005).

The antecedent concentrations represent the hourly atmospheric NOx and VOC concentrations (in *ppb*) associated with CMAQ ozone modelling at each grid cell. These data are generated using lagged NOx and VOC outputs from CMAQ and spatially-weighted according to surrounding wind flow for each grid cell. The detailed method of generating antecedent concentration data will be discussed in Section 4.3.

I will also use space-time data on temperature, wind direction and speed,

and planetary boundary layer height. These meteorology data are WRF model outputs, which are post-processed by Meteorology-Chemistry Interface Processor (MCIP) into a format usable for CMAQ, and convenient for statistical analysis using `R`. Moreover, WRF output has accompanying geographical coordinate data and the topography of the modelled region, which is linked to the CMAQ output. Within the CMAQ system, each grid cell is spatially indexed by the longitude and latitude of its centre.

The WRF and SMOKE outputs are produced on the same spatial domain and grid system used by the CMAQ model. Their outputs at each location and hour is presented as the numerical result from the spatial-interpolation and temporal-averaging of each respective grid-cell's initial and boundary conditions. The details of aforementioned model variables, as well as the reasonings behind their selection will be discussed extensively in Chapter 4.3.

The setup of CMAQ modelling runs

In Chapter 1, I described a generalized picture in the way CMAQ-WRF-SMOKE modelling system is run interactively to model space-time ozone. This subsection summarizes the specific setups of CMAQ, SMOKE and WRF modelling runs that produced the data in this thesis.

The SMOKE emission model generates an emission inventory and distributes the emissions into spatial grids and time periods at varying degrees of resolutions/intervals (SMOKE v2.5, University of North Carolina, 2012). The emission inventory generated in this study is further adjusted for both the amount and source location to reflect the change in LFV's "population density and economic activity" over the years (Steyn et al., 2011, 2013). The overall annual emission rates of NOx, VOC and other pollutants are obtained from the Metro Vancouver forecast and backcast emission inventories reported by Greater Vancouver Regional District (GVRD) in 2007[4].

The SMOKE output used here is the sum from 10 types of emission source: light and heavy duty vehicles, off-road vehicles, rail-roads, aircrafts,

[4]Prepared by Metro Vancouver.

marine, other emission sources, biogenic emissions, point and area sources. The LFV mobile (vehicle) emission rates are modelled by MOBILE6.2 and MOBILE6.2C models (US Environmental Protection Agency, 2010) using backcast emission totals from the above mentioned GVRD inventory. The regional ozone modelling in this thesis is analyzed at 4km×4km spatial resolution, such detailed biogenic emission modelling is handled by MEGAN version 2.04 (Guenther et al., 2006).

WRF (v3.1, Skamarock et al. (2008)) produces the 3-dimensional meteorological fields for ozone modelling. The meteorological conditions were simulated at 48 vertical levels to model air pollution at all elevations. As with CMAQ and SMOKE, it can simulate a space-time process at varying spatial resolutions. The Kain-Fritsch convective parameterization is applied to model "unresolved cloud updraft and downdraft", and Asymmetric Convective Model (ACM, version 2) accounts for "unresolved Planetary Boundary Layer" process (Steyn et al., 2013). Data from Moderate Resolution Imaging Spectroradiometer (MODIS), either weekly average values or weekly values around the episode dates, were used to initiate the simulations of sea-surface temperatures (Steyn et al., 2013).

The CMAQ (EPA Model-3) modelling system of version 4.7.1 models the overall photochemical process. For each ozone episode, the modelling is done over a period of 96 hours with a 13-hour "spin-up" period (Steyn et al., 2011, 2013). This means that, a full 96-hour ozone episode dataset contains three full days of ozone simulation plus additional 11 hours on the last day. This spin-up period is determined based on past experiences of our CMAQ data providers. Due to the recirculation of pollutants within LFV (Seagram et al., 2013), the pollutants from the previous day remains in LFV as the initial pollutants of a new diurnal ozone process.

Furthermore, background ozone, CO and NOx concentrations were provided by Re-analysis of TROpospheric chemical composition (RETRO) monthly average outputs, which in turn are simulated jointly by general circulation model, and chemical and aerosol model called ECHAM5-MOZ: European Centre Hamburg Model-Model for Ozone and Related chemical Tracers (Steyn et al., 2013). Carbon-bond 5 (CB05) gas phase chemical mechanism

with chlorine (Yarwood et al., 2005) using the Aerosol Energetics (AE-5) aerosol module was used.

The computation time of CMAQ depends on the available computing power. In our case, it took CMAQ approximately one day to simulate one day of ozone at 4km × 4km resolution[5].

Episodes and spatial domain of our study

At 4km × 4km resolution, the CMAQ modelling region covers an area in the Pacific Northwest that includes parts of Washington state in the U.S., Alberta in the east and northern BC mountains. I focus my modelling around the Lower Fraser Valley (LFV): a region of British Columbia that encompasses Greater Vancouver Regional District (Metro Vancouver) and Fraser Valley Regional District. The Fraser Valley Regional District spans from Abbotsford to Hope in the east.

Specific for this thesis, the "full" region under analysis is a rectangular approximation to the valley floor of the LFV. The large "pins" in Figure 2.1 indicate the corners that define my rectangular LFV region. This rectangular LFV includes a small portion of the north shore mountains immediately adjacent to some urban areas, and excludes the area around Hope (to the east of Chilliwack). This modelling region is comprised of 229 CMAQ (hence WRF and SMOKE) grid cells, whereas a complete CMAQ model domain contains 17716 grid cells at 4km × 4km resolution.

Each CMAQ run is used to model a summer-time ozone episode, which typically lasts 96 hours that span over 5 days: 13 hour spin-up period on the first day, 3 full days in the middle and 11 hours on the last day. In all, model outputs for 5 episodes are used in this study, they took place in the years 1985, 1995, 1998, 2001 and 2006 (Steyn et al., 2013; Ainslie et al., 2013). Table 2.1 shows the start and end time of each episode.

[5]From Bruce Ainslie, who produced all the CMAQ outputs used in this thesis

Year	Time span
1985	July 18th, 1100PST - July 22nd, 1000PST
1995	July 16th, 1100PST to July 20th, 1000PST
1998	July 24th, 1100PST to July 28th, 1000PST
2001	August 9th, 1100PST to August 13th, 1000PST
2006	June 23rd, 1100PST to June 27th, 1000PST

Table 2.1: CMAQ modelled ozone episodes: years and the episode durations.

2.1.2 Observational Data from Air-quality Monitoring Sites

The observed data on ozone and other variables are collected at air-quality monitoring locations across the lower mainland of BC (Metro Vancouver, 2012). At each location, there is an instrument that draws in ambient air and measures the pollutant concentrations in the air sample. This procedure can be done every few seconds, and such rapid-response data collection allows for various averaging times depending on the type of data analysis (Metro Vancouver, 2013). The data currently available are based on hourly averages.

Besides air-quality data on ambient ozone and NOx concentrations, each monitoring location also collects accompanying weather data on temperature, wind direction and speed. Although weather data are recorded multiple times per hour, available data are hourly averages. The VOC concentrations are measured only at 4 of the 17 ozone monitoring stations, and they are usually available as daily values. The physical measurements of LFV planetary boundary layer height are not available for our study (Steyn et al., 2011).

I make use of observation data collected from monitoring sites located within my rectangular LFV region. The small "red pins" in Figure 2.1 show the 17 monitoring locations that recorded the data for 2001 and 2006 (Ainslie et al., 2009), and Table 2.2 shows the associated station coordinates and names. The number and locations (longitude and latitude) of available monitoring stations vary by the episode. One reason is that between 1985 to 2006, some monitoring stations were retired from service while new locations were established. Moreover, for certain years some of the locations

contain a large number of missing measurements and data collected from such locations are discarded from analysis. The 1985, 1995, 1998 , 2001 and 2006 episodes have observation data available from 11, 16, 16, 17 and 17 monitoring sites.

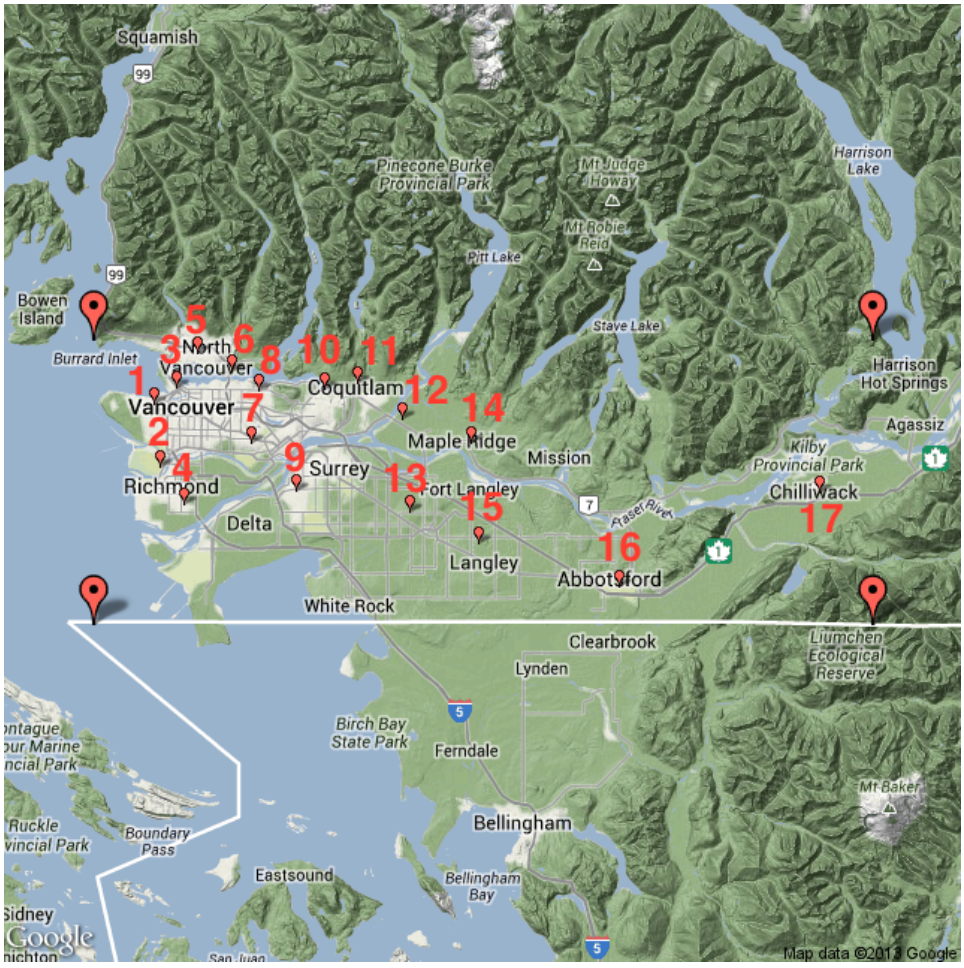


Figure 2.1: Locations of current measuring stations (small pins) and the corners of my self-defined rectangular LFV region (large pins). The station coordinates and names associated with the numbers 1 to 17 are in Table 2.2.

Number	Longitude	Latitude	Name
1	-123.16	49.26	Kitsilano
2	-123.15	49.19	YVR
3	-123.12	49.28	Robson square
4	-123.11	49.14	Richmond south
5	-123.08	49.32	Mahon park
6	-123.02	49.30	North Vancouver
7	-122.99	49.22	Burnaby south
8	-122.97	49.28	Kenshington park
9	-122.90	49.16	North Delta
10	-122.85	49.28	Rocky Point Park
11	-122.79	49.29	Coquitlam
12	-122.71	49.25	Pitt Meadows
13	-122.69	49.13	Surrey east
14	-122.58	49.22	Maple Ridge
15	-122.57	49.10	Langley central
16	-122.31	49.04	Central Abbotsford
17	-121.94	49.16	Chilliwack

Table 2.2: The Station names and coordinates of the numbers 1 to 17 in Figure 2.1: the map of the LFV monitoring network.

2.2 Data for CMAQ Evaluation in Chapters 5 and 6

This section describes the various numerical processing of CMAQ output and observation data used for CMAQ evaluations in Chapters 5 and 6.

2.2.1 Interpolated CMAQ Data

A proper implementation of my proposed CMAQ evaluation approach requires that CMAQ outputs and observations be matched on a spatial-temporal domain. Being both hourly data, CMAQ outputs and observation are matched in time. However, this is not the case with space: CMAQ outputs are air pollution data on a regular spatial grid across the entire LFV, while observations cover an irregular and sparse set of locations (Section 2.1).

The CMAQ evaluation analyses in Chapter 5 and 6 will be based on observation data from the n_{obs} monitoring sites and CMAQ-WRF-SMOKE outputs spatially interpolated onto the locations of these n_{obs} locations. Therefore, the spatial domain of statistical CMAQ evaluation is defined by the "ozone monitoring space", not the "CMAQ modelling space". As discussed in Section 2.1, the monitoring locations vary by the episode, so the CMAQ interpolation is done by the episode.

Each computer model output is spatially indexed by the longitude-latitude of the corresponding grid-cell's centre point. While the example below shows how spatial interpolation is done for CMAQ ozone output, the method is the same for other variables.

1. For each ozone monitoring station, choose 4 neighbouring CMAQ grid cells using the combined criteria: closeness in Euclidean distance and good coverage around the point-location of the monitoring site. Figure 2.2 shows for selected observation stations, the chosen 4 CMAQ neighbours that were used for interpolation.

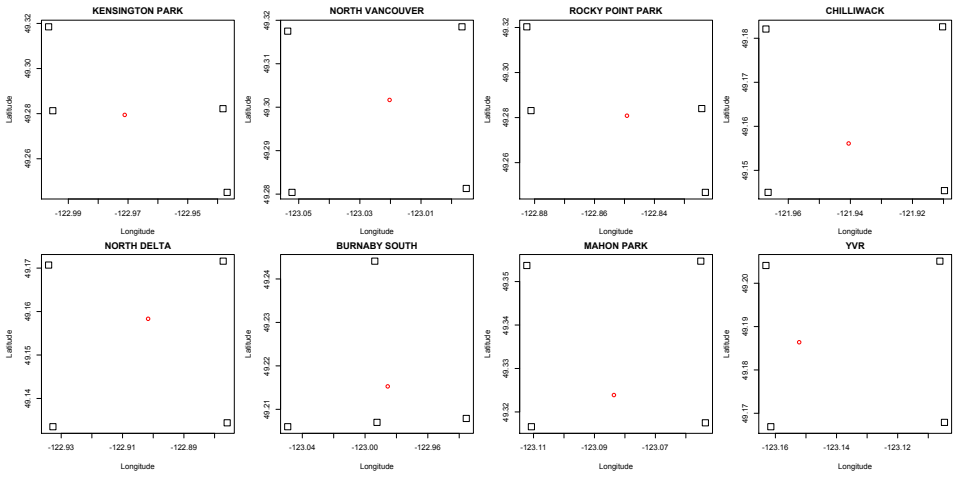


Figure 2.2: For 8 selected monitoring stations, the 4 CMAQ neighbours to be used for interpolation. The red dot is the location of a monitoring station and the squares are the centres of the 4km×4km CMAQ grid cells.

2. For each monitoring location, interpolate the 4 CMAQ outputs via Inverse Squared Weighting:

$$w_s = \frac{1}{d_s^2}\left(\sum_{s=1}^{4}\frac{1}{d_s^2}\right)^{-1}, \quad O_{int}^c = \sum_{s=1}^{4} w_s \cdot O_s^c,$$

where O_{int}^c denotes "interpolated CMAQ". O_s^c is the CMAQ output at grid cell s, d_s is the Euclidean distance between s and the monitoring site, and $s = 1, \ldots, 4$ for each monitoring site.

In the end, one obtains space-time data of interpolated CMAQ outputs, whose locations are matches to the longitude-latitude of observations.

Here, the interpolation at each observation location is based on only the four nearest CMAQ grid-cells. For such a local field, inverse squared weighting is an adequate interpolation method. For interpolations based on larger and more complex fields, one might need to adjust the power of inverse weighting according to the dataset's "coefficients of spatial variations" (Gotway et al., 1996). Alternatively, one may interpolate CMAQ outputs at the observation locations using Kriging methods (Matheron, 1963; Cressie, 1990).

2.2.2 Missing Observations and Measurement Errors

For an episode, the percentage of missing data is typically $\leq 5\%$ for all variables. In addition, data are usually missing for just an hour or occasionally, a few hours. There are also instances where the observations are completely unavailable for one or more stations, e.g., the temperature data during the 2006 episode are not available for the Robson Square and North Vancouver stations, and the percentage of missing data reached $\approx 12\%$ in total (across all 17 stations and 96 hours).

When the data are missing for 3 or fewer consecutive hours, I interpolate the missing observations by simple linear regression: "ozone concentration" is the response variable and "hour" the regressor along with an intercept term, and the regression is fitted with two available observations at either side of the missing period. Otherwise, a "proxy station" is chosen for each

location, and the missing observations at this location are filled-in with data from the "proxy station". The proxy station is selected based on the visual criteria of spatial-temporal homogeneity.

For example, at Pitt Meadows, the ozone observations are missing for 4 hours on June 26th, 2006. These missing data are filled with the same-hour observations from the Rocky Point Park. Figure 2.3 shows the spatial plot of 96-hour ozone mean produced from the 2006 CMAQ output: it helps to assess the spatial homogeneity between locations. The contour of LFV shoreline, and the locations of the Pitt Meadows and Rocky Point Park monitoring sites are also shown. The plot of mean field indicates that the 96-hour ozone means are similar between two locations. Figure 2.4 plots the hourly observations from the two locations; it helps to compare their temporal patterns. As shown, the observations from the two locations closely track each other by the hour, except that the ozone peak is higher for Pitt Meadow on the 3rd day. However, the data are missing during the pre-noon hours on the 4th day, and observations from the previous days show that during these hours the ozone levels are similar between the two locations.

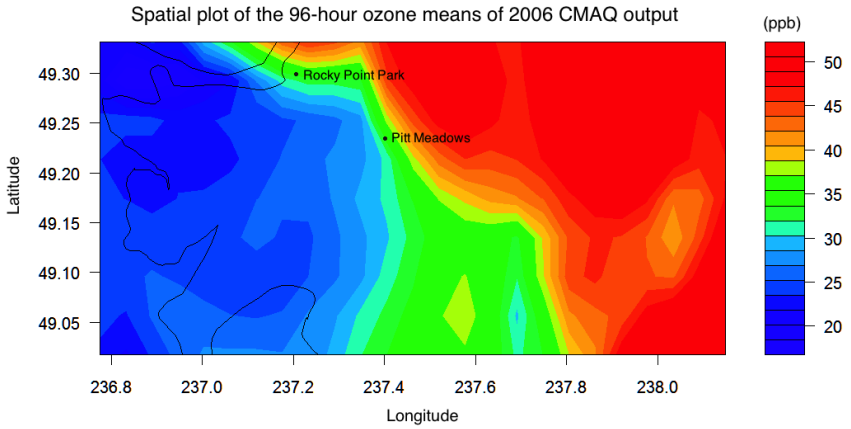


Figure 2.3: From the 2006 CMAQ ozone output: the spatial plot of the 96-hour ozone means overlaid with the LFV shoreline. Note that Longitude is expressed differently from the other plots: it is the usual longitude angle plus 360°, simply used to overlay the available map data.

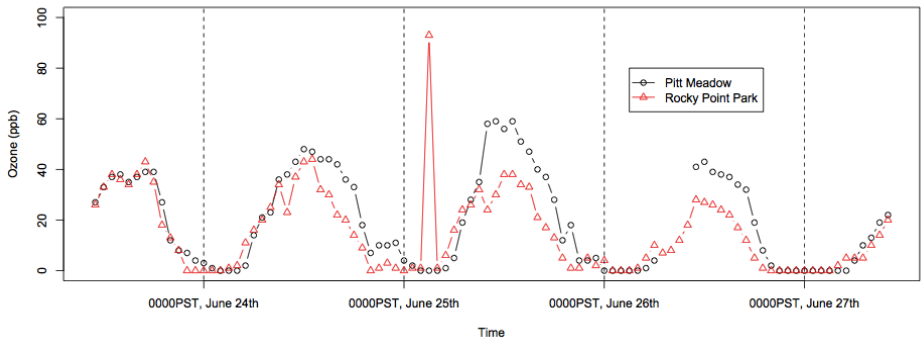


Figure 2.4: Hourly ozone observations from Pitt Meadow and Rocky Point Park during the 2006 ozone episode. The dashed line indicate the hour 0000 of each day. Notice the effect of nocturnal *down mixing* at 0400PST, the 25th, in Rocky Point Park

When a station have ozone observations that are missing or flagged as incorrect readings for consecutive 8 or more hours, the data from this station are not used for analysis. This is because observations are used as "reference data" for CMAQ evaluation, and excessive amount of interpolated/estimated inputs would introduce unwanted bias into the statistical analysis.

As Figure 2.4 shows, there is a hour long spike in the early-morning for Rocky Point Park. This is the result of what is known as Down Mixing (Salmond and McKendry, 2002). It is a nocturnal process in the boundary layer of LFV, which creates vertical mixing and downward transport of pollutants from the atmosphere above. The natural consequence is the sudden spike of ozone and precursor concentrations in particular locations around LFV, such as this example at Rocky Point Park. This phenomenon is beyond the scope of my current analysis. I chose to replace the down-mixing affected data points using the average measurements from adjacent hours at the same location.

2.3 Summary of Data Used in the Thesis

I also generated synthetic/simulated ozone data that emulates the space-time structure of the real LFV ozone field. I then implemented a few analyses using these synthetic data to complement the ozone PCA in Chapter 3. Since these synthetic data are analyzed in Appendix B.1, the details regarding the design and creation of such data is described in Appendix A.1 instead of this chapter.

In summary, the types of data used in this thesis are:

- CMAQ ozone output and associated data on meteorology, chemical precursor emission rates and antecedent concentrations (processed from WRF, SMOKE and CMAQ output).
- Physical Observations. These are data recorded at monitoring stations across LFV. Observation data include ozone and NOx concentrations, temperature, wind speed and direction.
- Above mentioned synthetic ozone data.

In Section 2.1, I mentioned that the full spatial domain of my analysis is the rectangular LFV shown in Figure 2.1. In upcoming statistical analyses in Chapter 3 to 6, I will analyze data based on either this "full" rectangular LFV or subregions of it. The decision is based on the specific goal of analysis at hand. The list below gives a quick overview of the regions analyzed:

1. In Chapter 3, I will analyze CMAQ ozone outputs across part of LFV where the elevation is below 150 meters. This region represents the triangular "valley floor" of LFV. This is a region where most of the ozone activities (chemical reactions and atmospheric transportation) occur. I will simply refer to this region as "LFV".
2. In Chapter 4, I will use CMAQ-WRF-SMOKE data across the full rectangular LFV. I refer to this region as "rectangular LFV" to differentiate it from the "LFV" defined above.

3. The CMAQ evaluation analyses in Chapters 5 and 6 will be based on the area defined by the n_{obs} monitoring locations. As discussed, the computer model outputs will be spatially interpolated onto the observation locations.

Chapter 3

Principal Component Analysis of Space-time Ozone

My research explores methods of AQM evaluation based on the the analysis and modelling of ozone features. The term *ozone features* describes the dominant spatial-temporal structures of a space-time ozone field. In this research, methods of PCA are used to decompose a space-time ozone data into spatial and temporal features. Since data features capture the key space-time variation within a dataset, one may argue conceptually that they are the most informative portion of the data for statistical analysis.

In this chapter, I use Principal Component Analysis (PCA) of CMAQ outputs to address important topics related to the statistical analysis of ozone features. Specifically, I aim to accomplish three things:

1. Determine the most appropriate PCA method for feature-based AQM evaluation. The exact PCA method will then be applied consistently in this study.

2. Analyze whether a space-time ozone field can be understood and analyzed through a small number of ozone features.

3. Interpret the ozone features: either as statistical summaries of data, or as space-time structures that explain important underlying mechanisms of the ozone process.

The following discussion explains the purposes of these analyses.

An important topic relating to the comparison of ozone features is the topic of "feature correspondence" between ozone data. In order to compare ozone features between AQM modelled ozone and physical observations, it

is crucial that we understand whether the *same* types of features are being evaluated. Otherwise, one type of ozone feature from AQM will be judged against an entirely different feature from observations, leading to wrong conclusions about the computer model's performance. One way of addressing above concern is to understand what types of ozone features dominate the ozone field being analyzed. If one is able to interpret the extracted features from both datasets, subsequent feature-to-feature evaluations can be concluded in ways that is both logically justifiable and informative.

Each spatial ozone feature is numerically represented by an Empirical Orthogonal Function (EOF) vector, and each temporal feature is represented by a Principal Component (PC) vector. Hence, PCA is used to extract EOFs from an ozone dataset, and these EOFs are the estimates (from sample dataset) of the unknown *true* EOFs of the ozone field under analysis. The following are two main PCA-related complications due to uncertainty in estimating EOFs, or sampling uncertainty. One needs to address them in order to assure feature correspondence during AQM evaluation.

- All pairs of EOF vectors are orthogonal. If the 1st EOFs of two data sets capture different, thus incomparable, types of features, then this feature discordance will carry over to higher-order features due to the aforementioned orthogonality constraint (Cohn and Dennis, 1994). In addition, if the 1st EOF of an analyzed ozone field is estimated incorrectly, perhaps due to the sparseness of sample data, then the orthogonality requirement of EOFs results in the propagation of EOF estimate errors towards higher-order features (Cohn and Dennis, 1994; Monahan et al., 2009).

- Identifiability or separability of ozone features: whether extracted ozone features are individual space-time fields separable from the rest, or whether multiple features form an inseparable couplet or multiplet. This is a data-specific statistical property defined by EOF estimate errors, and it informs us which features can be compared *individually* or *jointly*.

In this chapter, above mentioned topics will be analyzed using CMAQ ozone

outputs for the LFV during the 5 episodes in 1985, 1995, 1998, 2001 and 2006 (Section 2.1).

Through ozone PCA in this chapter, I will establish the exact PCA procedure that will be implemented in all subsequent statistical analyses. This procedure is meant to address aforementioned PCA-related complications and to maximize interpretability of ozone features.

I will further examine the number of spatial and temporal ozone features that are meaningful for statistical analyses. First, this question is addressed by analyzing the amount of data variation as well as the importance of space-time structure each ozone feature can capture. Secondly, I will implement a simulation-based approach to construct a statistical test that determines the number of meaningful ozone features. These analyses address the question of whether a complex space-time air pollution field can be understood and analyzed through simpler data features.

As discussed in Section 2.1, CMAQ outputs are produced on a regular grid with high spatial resolution of 4km×4km, while observations are irregularly placed at most, 17 locations in LFV. This means that, compared to observations it will be easier to interpret the spatial ozone features of CMAQ. Therefore in this chapter, for the purpose of learning about ozone PCA and LFV ozone features, I will use CMAQ outputs instead of observation data. However, later in Chapter 5, I will implement PCA on ozone observations to compare the CMAQ features to the observed features.

Before proceeding, I need to define a few important recurring terminologies. Let $\mathbf{O}_{t\times n}$ be an ozone dataset of dimension $t \times n$, t being the number of hours and n the number of locations. The term *spatial field of temporal ozone means* refers to the column means of $\mathbf{O}_{t\times n}$: ozone averaged across t hours, resulting in a spatial field of ozone means. Short descriptions "mean field" or "field of means" will also be used to describe the temporal ozone means. The term *hourly spatial ozone means* refers to the row means of $\mathbf{O}_{t\times n}$: a time series of ozone averaged across n locations at each hour. "Hourly LFV mean ozone" will mostly be used to describe this time series.

Similarly, *temporal ozone standard deviation* is calculated across time, i.e., the column standard deviation of $\mathbf{O}_{t\times n}$. *Spatial ozone standard devia-*

tion is then the standard deviation calculated across space.

Section 3.1 summarizes what is already known about LFV's ozone episodes. Section 3.2 reviews the PCA or EOF-decomposition methods relevant to this research, and determines the PCA procedure to be used in this thesis. In Section 3.3, I will discuss the number of ozone features useful for further analysis. In Section 3.4, I will implement PCA on the space-time CMAQ outputs from the 5 episodes (Table 2.1), where the goal is to formulate understandings and interpretations of extracted ozone features that are useful for the upcoming AQM evaluations. Section 3.5 summarizes the findings in Chapter 3.

3.1 An Overview of LFV Ozone Field during an Episode

The following discussion will use CMAQ outputs to highlight the distinct space-time structures of LFV ozone fields during an episode. Later in Section 3.4, I will determine whether the visible qualitative ozone patterns can be extracted through PCA.

Figures 3.1 and 3.2 show for selected hours during the 1985 and the 2006 episodes, the 3-dimensional spatial ozone fields outputted by CMAQ. During a summer-time ozone episode, an ozone plume forms in the morning across the west of LFV. This plume keeps building in concentration before the early afternoon peak while slowly travelling east (Ainslie and Steyn, 2007; Steyn et al., 2013). This eastward movement, driven by daytime westerly winds, carries the ozone plume inland throughout the day. As night falls, the intensity of the ozone process (its creation and destruction) decreases dramatically due to the absence of UV radiation. Hence during the night and following morning, there exists low-level background ozone across the triangular valley floor of LFV (Salmond and McKendry, 2002).

Furthermore, as the map in Figure 3.3 shows, the LFV is surrounded by mountains from the northeast and the southwest. Combined with a low boundary layer height, these surrounding mountains act as a physical bar-

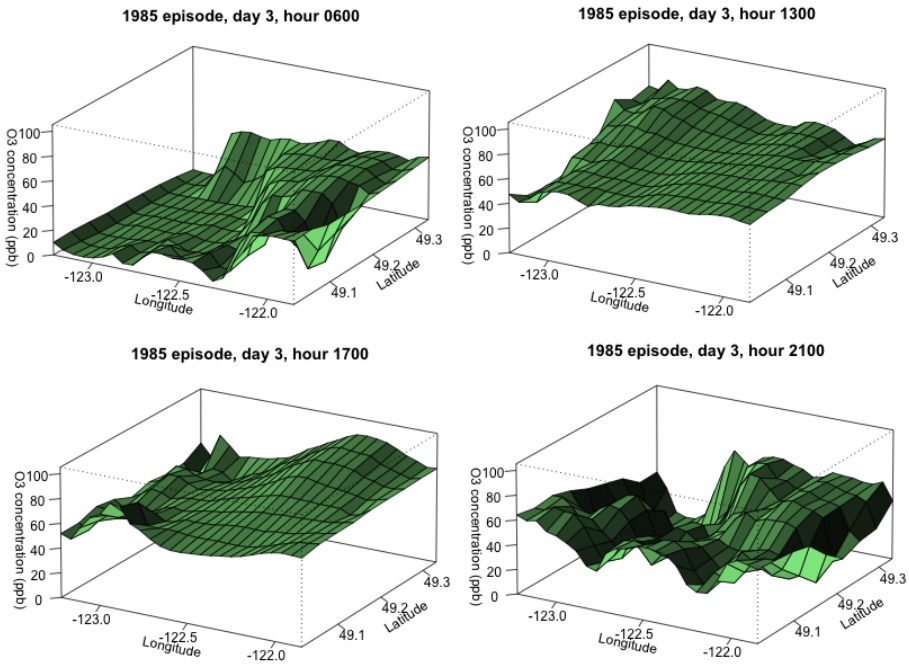


Figure 3.1: For selected hours during the 1985 ozone episode, 3-dimensional spatial plots of the hourly ozone field. The spatial domain is the "rectangular LFV" including the north shore mountains.

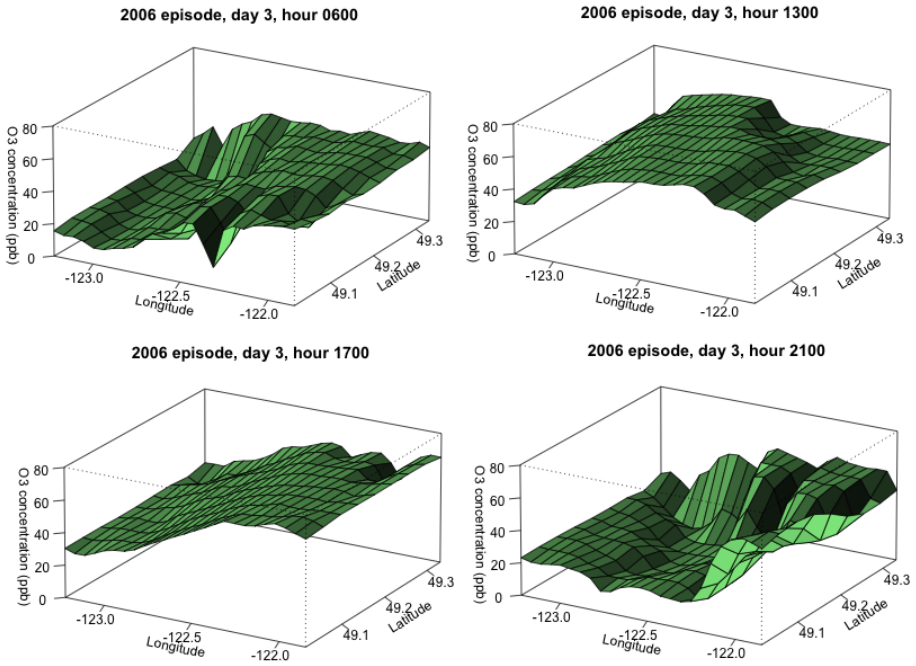


Figure 3.2: For selected hours during the 2006 ozone episode, 3-dimensional spatial plots of the hourly ozone field. The spatial domain is the "rectangular LFV" including the north shore mountains.

rier to the eastward horizontal advection of pollutants that channels them along the valley (Taylor, 1991; Steyn et al., 1997), and subsequently creates a "bottleneck effect" (Robeson and Steyn, 1990) that traps the pollution within LFV's mountainous barrier. Therefore, during the night time as ozone within LFV's valley floor reverts to background level, the surrounding mountains and parts of eastern LFV generally suffer from high ozone pollution due to the accumulation of the plume produced earlier in LFV. This is noticeable from the ozone fields at 2100PST and 0600PST in both Figures 3.1 and 3.2. The high ozone pollution along the mountains remains for the entire duration of an episode.

Figure 3.3: Locations of current measuring stations (small pins) and the corners of the complete rectangular LFV region (large pins).

Each episode is also defined by slight deviations in its space-time structure from the generalized LFV ozone structure I just described. The 1985 episode has a pronounced ozone plume around the city of Vancouver, and the plume transports eastward in a discernible wavelike pattern along the northern part of LFV (1300PST and 1700PST plots of Figure 3.1). The 2006 episode (Figure 3.2) has an ozone plume forming in the middle LFV and transports eastward as a "block" of plume that is evenly distributed from north to south. Perhaps more significant is the difference in the night time ozone fields between 1985 and 2006. For 2006, the night time spatial ozone variation is consistent with the generalized LFV night time pattern, and the

maximum ozone level is around 20 *ppb* at the valley floor (2100PST plot of Figure 3.2). For 1985, there is a heavy accumulation of ozone pollution around the southwest corner of LFV by the evening, where the maximum ozone level exceeds 60 *ppb* (2100PST plot of Figure 3.1). This is in addition to the high ozone levels along the north shore mountains.

Therefore, despite the fact that every ozone episode happens at a narrow and specific set of weather conditions and regional pollution (Steyn et al., 2013), differences do exist. One needs to consider the ever changing local emission standards that can shift the spatial patterns of air pollution over the years (Steyn et al., 2011). Furthermore, the between-episode differences in ozone structures are partly influenced by the prevailing regional wind patterns.

Each ozone episode is defined by its unique wind regime, i.e., combination of mesoscale wind direction and speed. Using wind observations at YVR, Ainslie and Steyn (2007) identified four types of possible mesoscale wind regimes during an LFV ozone episode, and characterized the dominant wind patterns of each episode. Figure 3.4 shows the hodograph of the four wind regimes from Ainslie and Steyn (2007). A point in the plot indicates the direction from which the wind blows towards the origin, the distance from the origin shows the wind speed, and the number above each point indicate the hour of the day. As shown, the daytime atmospheric circulation is usually defined by a westerly wind system. Regime IV also displays easterly winds in the evening. Table 3.1 summarizes the wind regime of the middle 3 full days of each episode (remember that the first and last days are half days).

Year	Dates	Wind Regime
1985	July 19-20-21	I-IV-IV
1995	July 17-18-19	III-III-III
1998	July 25-26-27	II-III-II
2001	August 10-11-12	II-II-II
2006	June 24-25-26	I-I-III

Table 3.1: The daily wind regime for the middle 3 full days of each episode. Figure 3.4 shows what circulation types I-IV look like.

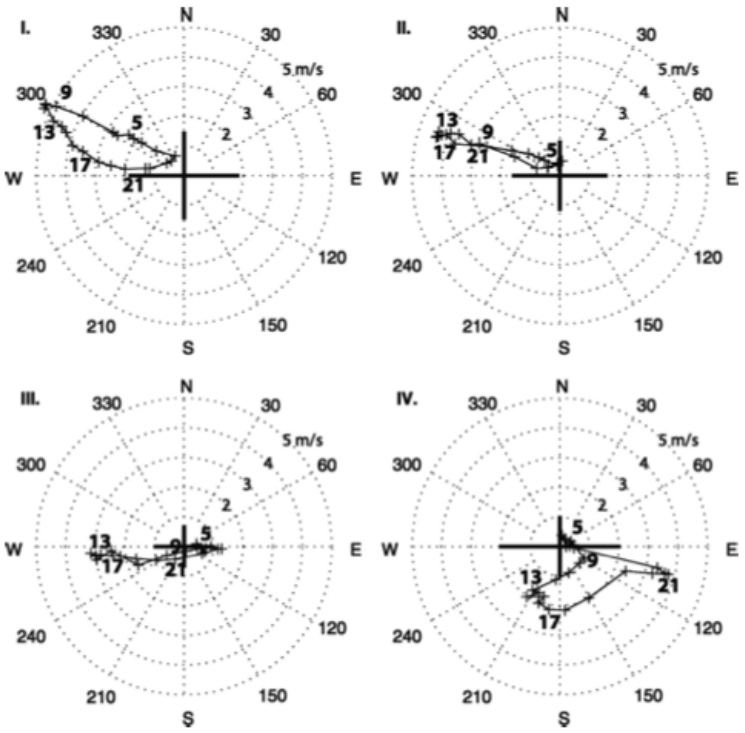


Figure 3.4: The four types of LFV wind regime during an ozone episode as described by Ainslie and Steyn (2007). The wind regimes are presented as hodograph: the number on top of a point indicates the hour of the day, the point's position indicates the direction in which the wind blow towards the origin and the distance from origin indicates the speed.

In Section 3.4, the LFV ozone structure and space-time patterns described in this section will be used as physical references for interpreting the ozone features. Furthermore, PCA will be done episode-by-episode to identify any changes (man-made or weather-driven) in LFV's ozone structure over the years. PCA will be also be done on ozone fields under separate wind regime types. This helps to analyze possible influence of wind regime on the dominant pattern of LFV ozone advection.

3.2 PCA Methods and Related Topics

This section defines the PCA notations and methods used throughout the thesis. I will also (1) review the mathematical properties and identities of PCA that are useful for subsequent statistical analyses, and (2) discuss the PCA-related complications (briefly described in the beginning of this chapter) that require attention during feature-based AQM evaluations in Chapters 5 and 6.

3.2.1 Definitions of EOFs and PCs

Let $\mathbf{O}$ be a $t \times n$ matrix of ozone data, where t is the number of time points and n the number of locations (longitude and latitude concatenated). In a general sense, a column of $\mathbf{O}$ can be viewed as containing t observations on one of n variables, and a row of $\mathbf{O}$ is one set of observations on n variables.

PCA is typically implemented on centered or standardized data. In centered data, each ozone value in $\mathbf{O}_{t\times n}$ is "centered" by subtracting its location-specific mean over the t hours. Denote the column centered data as $\tilde{\mathbf{O}}_{t\times n}$, where the j-th column of $\tilde{\mathbf{O}}$ is

$$\tilde{\mathbf{O}}_{t\times n}[,j] = \mathbf{O}[,j] - \frac{\sum_{i=1}^{t} \mathbf{O}[i,j]}{t}.$$

Hence $(1/n)\tilde{\mathbf{O}}^{\mathrm{T}}\,\tilde{\mathbf{O}}$ is an $n\times n$ sample covariance matrix since the mean of each column (corresponding to one ozone location) is 0. Standardized data are subsequently obtained by dividing each column of $\tilde{\mathbf{O}}$ by its column standard

deviation. Column standardization maps the origins of data vectors onto a common origin in data space, and the scales are unitless and comparable.

However, I believe data centering/standardization is unnecessary given the purpose of my analyses. In this research, decompositions are performed on space-time data of single photochemical or meteorological quantities, so all data elements have the same units and scale. The space-time data produced by computer models are well-behaved, and except for rare occasions of equipment malfunctioning, observed data are well behaved as well. Hence influential data outliers are not a pressing concern. More importantly, given my goal of feature-based AQM evaluation and space-time ozone modelling, it is reasonable to keep the mean structure intact during PCA. This point will be elaborated at the end of this section. First, I will present the PCA procedures and properties that are relevant to this research. The discussions in this chapter are focused on the data matrix $\mathbf{O}_{t\times n}$, but the presented algebraical properties are also applicable to $\tilde{\mathbf{O}}_{t\times n}$ (and any data in general).

In PCA, the matrix $\mathbf{O}^{\mathrm{T}}\mathbf{O}$ undergoes eigen-decomposition $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{\mathrm{T}}$, giving the matrix $\mathbf{E}_{n\times n}$ whose columns are n-length eigenvectors of $\mathbf{O}^{\mathrm{T}}\mathbf{O}$, and a diagonal matrix of n eigenvalues $\mathbf{\Lambda}$. Here, each of the n eigenvectors is referred to as an EOF, and PCA eigen-decomposition is interchangeably referred to as EOF decomposition. There is an eigenvalue corresponding to each eigenvector. Denote each eigenvalue as λ_j, $j = 1, \ldots, n$, we have $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$. Multiplying $\mathbf{O}_{t\times n}$ by $\mathbf{E}_{n\times n}$ gives a matrix of PCs $\mathbf{P} = \mathbf{OE}$, which has the same dimension as the original data $\mathbf{O}_{t\times n}$. The combined analyses of EOF decomposition and PC calculations are referred to here as PCA.

$\mathbf{P}$ is an orthogonal basis for $\mathbf{O}$, and each column of $\mathbf{P}$ consists of t weighted row-sums of $\mathbf{O}$. In other words, each value in a PC is the location-weighted sum of the ozone values at the corresponding time point. Therefore, the values in an EOF can be seen as the spatial weights, or the "importance", of each location over the time period t captured by $\mathbf{O}$. Here, each column of $\mathbf{E}$, denoted as $\mathbf{E}_j$, $j = 1, \ldots, n$, is a normalized eigenvector. Hence the EOF values are unitless. Each column of $\mathbf{P}$ is denoted as $\mathbf{P}_j$, and the PC values have the ozone unit *ppb*.

As discussed above, each element in an $\mathbf{E}_j$ represents a certain type of spatial weight of the corresponding location. To state it explicitly, each $\mathbf{E}_j$ contains n "spatial variables" that can be defined on the spatial domain of $\mathbf{O}_{t\times n}$: each element of $\mathbf{E}_j$ can be located on the spatial domain by the location of its corresponding column in $\mathbf{O}_{t\times n}$. This leads to the fact that all $\mathbf{E}_j$'s are in essence, spatial processes that can be conveniently plotted over the spatial domain of the ozone field. The same can be said for the $\mathbf{P}_j$'s: each of the t elements in a $\mathbf{P}_j$ is the weighted row-sum of $\mathbf{O}_{t\times n}$ at hour i, where $i = 1, \ldots, t$. Hence each $\mathbf{P}_j$ can be plotted as a time-series of length t. The nature of the $\mathbf{E}_j$ and $\mathbf{P}_j$ will be illustrated by the ozone feature analysis in following sections.

3.2.2 Mathematics of PCA

The data in this thesis can have $t > n$ (observations and interpolated CMAQ output) or $t < n$ (CMAQ output for LFV). Although the following discussion on the mathematical properties of PCA is presented for the the case of $t > n$, equivalent results hold for $t < n$.

An important property of $\mathbf{E}$ is its orthogonality, i.e., any pair of columns is orthogonal. Moreover, I will also show the orthogonality of the columns is also a property of the principal component matrix $\mathbf{P}_{t\times n}$. Taking into account the orthogonality of $\mathbf{E}$ and $\mathbf{P}$, one may derive and arrive at some useful definitions and relationships.

Interpretation of Eigenvalues

Starting with the definition of eigenvectors and eigenvalues, we have:

$$(\mathbf{O}^{\mathrm{T}}\,\mathbf{O})\mathbf{E} = \mathbf{E}\mathbf{\Lambda}, \tag{3.1}$$

where $\mathbf{\Lambda}$ is an $n \times n$ diagonal matrix of eigenvalues. Left-multipling each side by $\mathbf{E}^{\mathrm{T}}$, we arrive at $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{\Lambda}$:

$$\begin{aligned}\mathbf{E}^{\mathrm{T}}\mathbf{O}^{\mathrm{T}}\mathbf{O}\mathbf{E} = \mathbf{E}^{\mathrm{T}}\mathbf{E}\mathbf{\Lambda} &\Rightarrow \mathbf{E}^{\mathrm{T}}\mathbf{O}^{\mathrm{T}}\mathbf{P} = \mathbf{E}^{\mathrm{T}}\mathbf{E}\mathbf{\Lambda} \\ &\Rightarrow (\mathbf{O}\mathbf{E})^{\mathrm{T}}\mathbf{P} = \mathbf{\Lambda} \Rightarrow \mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{\Lambda}.\end{aligned}$$

Hence $\mathbf{P}$ is column-orthogonal, and the eigenvalues correspond to the second moments of the variables represented by the columns of $\mathbf{P}$. Furthermore, right-multiplying (3.1) by $\mathbf{E}^{\mathrm{T}}$ gives $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{\mathrm{T}}$. It can be easily shown that the trace (diagonal sum) of $\mathbf{O}^{\mathrm{T}}\mathbf{O}$ is equal to the trace of $\mathbf{\Lambda}$:

$$\mathrm{tr}(\mathbf{O}^{\mathrm{T}}\mathbf{O}) = \mathrm{tr}(\mathbf{E}\mathbf{\Lambda}\mathbf{E}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{E}(\mathbf{\Lambda}\mathbf{E}^{\mathrm{T}})) = \mathrm{tr}((\mathbf{\Lambda}\mathbf{E}^{\mathrm{T}})\mathbf{E}) = \mathrm{tr}(\mathbf{\Lambda}).$$

Note that the diagonal elements of $\mathbf{O}^{\mathrm{T}}\mathbf{O}$ correspond to the second moments of the variables in $\mathbf{O}$. One may conclude that the sums of the second moments of $\mathbf{O}$ and $\mathbf{P}$ are equal. Using this important relationship, we can estimate the *proportion of data variation* explained by each EOF-PC pair:

$$\text{proportion of data variation accounted for by } \mathbf{E}_j \text{ and } \mathbf{P}_j = \frac{\lambda_j}{\sum_{j=1}^{n}\lambda_j},$$

where λ_j is the j-th diagonal element of $\mathbf{\Lambda}$.

Since $\mathbf{P}$ is column-orthogonal, any cross-product (uncorrected for the means) between the variables in $\mathbf{P}$ is 0. Therefore, $\mathbf{E}$ maps $\mathbf{O}$, a column-wise correlated matrix (correlated variables) into a new data matrix $\mathbf{P}$ whose variables are orthogonal.

In this thesis, I rank the $\mathbf{E}_j$'s and $\mathbf{P}_j$'s, $j = 1, \ldots, n$, according to the amount of data variation the j-th EOF-PC pair explains. Thus, the "1st EOF" and "1st PC" together account for the most data variation (equivalent to having the largest λ_j), and so forth for the successively higher order EOFs and PCs. Here, the "order " of an EOF-PC pair corresponds to their rank j: compared to $\mathbf{E}_1$ and $\mathbf{P}_1$, $\mathbf{E}_j$'s and $\mathbf{P}_j$'s of $j \geq 2$ are referred to as "higher order" data features.

Data Reconstruction via EOFs and PCs

The data matrix $\mathbf{O}$ can be constructed as follow:

$$\begin{aligned} \mathbf{P} &= \mathbf{OE} \Rightarrow \text{ multiply each side by } \mathbf{E}^{\mathrm{T}} \Rightarrow \\ \mathbf{PE}^{\mathrm{T}} &= \mathbf{OEE}^{\mathrm{T}} \Rightarrow \mathbf{O} = \mathbf{PE}^{\mathrm{T}}. \end{aligned}$$

The equality $\mathbf{O} = \mathbf{PE}^{\mathrm{T}}$ can be explicitly expanded into a sum of n EOF-PC terms:

$$\begin{aligned} \mathbf{P}_{t\times n}\mathbf{E}^{\mathrm{T}}_{n\times n} &= \begin{bmatrix} P_{11}E_{11} & P_{11}E_{21} & \dots & P_{11}E_{n1} \\ P_{21}E_{11} & P_{21}E_{21} & \dots & P_{21}E_{n1} \\ \vdots & \vdots & \vdots & \vdots \\ P_{t1}E_{11} & P_{t1}E_{21} & \dots & P_{t1}E_{n1} \end{bmatrix} + \\ &\quad \begin{bmatrix} P_{12}E_{12} & \dots & P_{12}E_{n2} \\ P_{22}E_{12} & \dots & P_{22}E_{n2} \\ \vdots & \vdots & \vdots \\ P_{t2}E_{12} & \dots & P_{t2}E_{n2} \end{bmatrix} + \dots + \begin{bmatrix} P_{1n}E_{1n} & \dots & P_{1n}E_{nn} \\ P_{2n}E_{1n} & \dots & P_{2n}E_{nn} \\ \vdots & \vdots & \vdots \\ P_{tn}E_{1n} & \dots & P_{tn}E_{nn} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{P}_1 & \dots & \mathbf{P}_n \end{bmatrix} \begin{bmatrix} \mathbf{E}_1^{\mathrm{T}} \\ \dots \\ \mathbf{E}_n^{\mathrm{T}} \end{bmatrix} = \mathbf{P}_1\mathbf{E}_1^{\mathrm{T}} + \mathbf{P}_2\mathbf{E}_2^{\mathrm{T}} \dots + \mathbf{P}_n\mathbf{E}_n^{\mathrm{T}} \\ &= \sum_{j=1}^{n} \mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}. \end{aligned} \tag{3.2}$$

$\mathbf{P}_j$ and $\mathbf{E}_j$ are vectors of length t and n respectively.

The $t \times n$ matrix $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ is the j-th order space-time component of data $\mathbf{O}_{t\times n}$ that captures the space-time interaction between $\mathbf{E}_j$ and $\mathbf{P}_j$. In other words, $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ is the j-th spatial-temporal ozone feature. As equation (3.2) shows, a complete ozone data can be recovered or constructed by summing n spatial-temporal ozone features.

The PCA of CMAQ outputs in the following sections will show that the eigenvalues decrease rapidly as the order of the EOF/PC increases. The implication here is that the amount of data variation explained by higher-

order space-time ozone features decreases rapidly. Indeed, the traditional statistical purpose of PCA is to reduce the dimensionality of a large dataset, using the fewest possible components to explain the most possible data variation (Hardle and Simar, 2012). If the first p features capture most of the variation in $\mathbf{O}$, instead of (3.2), we write

$$\mathbf{O} \approx \sum_{j=1}^{p} \mathbf{P}_j \mathbf{E}_j^{\mathrm{T}}, \;\; p \ll n. \tag{3.3}$$

The values of p will be discussed in Section 3.3.

Relationship Between PCA and SVD

PCA is related to Singular Value Decomposition (SVD). In general terms, SVD decomposes data as $\mathbf{O} = \mathbf{U}\mathbf{M}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U}^{\mathrm{T}} = \mathbf{U}^{-1}$ and $\mathbf{V}^{\mathrm{T}} = \mathbf{V}^{-1}$. When $\mathbf{O}$ is a non-square matrix, $\mathbf{M}$ is a rectangular-diagonal matrix of nonnegative real numbers. It follows that:

$$\mathbf{O}^{\mathrm{T}}\mathbf{O} = (\mathbf{U}\mathbf{M}\mathbf{V}^{\mathrm{T}})^{\mathrm{T}}\mathbf{U}\mathbf{M}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\mathbf{M}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{M}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{V}^{\mathrm{T}}.$$

In addition, $\mathbf{V}$ is an eigenvector matrix of $\mathbf{O}^{\mathrm{T}}\mathbf{O}$ (i.e., $\mathbf{V} = \mathbf{E}$), and $\mathbf{U}$ is an eigenvector matrix of $\mathbf{O}\mathbf{O}^{\mathrm{T}}$. Utilizing the aforementioned equality $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^{\mathrm{T}}$, we have

$$\mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^{\mathrm{T}} = \mathbf{V}\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{V}^{\mathrm{T}}.$$

In other words, $\mathbf{E} \equiv \mathbf{V}$ and $\boldsymbol{\Lambda} \equiv \mathbf{M}^{\mathrm{T}}\mathbf{M}$.

When $t < n$, there will be $t = \min(t, n)$ $\mathbf{E}_j$'s and λ_j's. The upcoming PCA in this chapter and the ozone feature modelling in Chapter 4 will be implemented using CMAQ outputs with $t < n$. However, to avoid confusion in notation, I will still denote the dimension of $\mathbf{E}$ and $\boldsymbol{\Lambda}$ by $n \times n$. Written as such, λ_j values of orders $j > t$ will be 0, and $\mathbf{E}_j$'s of $j > t$ will be structured such that associated $\mathbf{P}_j$ is an approximately zero-vector. In other words, data components $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ of orders $j > t$ will recover no additional data variation.

3.2.3 Relevant PCA-Related Topics

Feature-based AQM evaluation approach requires that following PCA-related complications to be addressed: (1) the ozone features may be incomparable between AQM and observations, (2) propagation of EOF estimate errors from $\mathbf{E}_1$ towards higher-order features, and (3) the ozone features may be inseparable or arbitrarily ordered. In addition, it is important to discuss PCA methods that may enhance the interpretability of ozone features.

Complications from Orthogonality Constraint and Data Column-Centering

If $\mathbf{E}_1$ from the AQM modelled ozone captures a different type of spatial ozone feature from the observed $\mathbf{E}_1$, then these $\mathbf{E}_1$'s are incomparable. Due to the orthogonality requirement of $\mathbf{E}_j$, this problem of feature discordance will be carried over to higher-order features, making the practice of feature-based AQM evaluation problematic. This problem is mentioned in the context of Acid Deposition Model evaluation in Cohn and Dennis (1994).

Moreover, the $\mathbf{E}_j$ calculated from the PCA of $\mathbf{O}_{t\times n}$ are estimates of the *true* EOF of the underlying ozone process. This is because $\mathbf{O}_{t\times n}$ can be regarded as sample data, i,e, one realization of the process. Suppose the estimated 1st-order feature $\mathbf{E}_1$ is incorrect or different from the true EOF, then the EOF orthogonality will cause $j \geq 2$ $\mathbf{E}_j$'s to be incorrect. This is the problem of "error propagation" between $\mathbf{E}_j$'s, and the implication here is that AQM outputs will be evaluated against observations based on incomparable sets of features.

The PCA of original un-centered data $\mathbf{O}_{t\times n}$ helps to maximize feature-correspondence during CMAQ/AQM evaluation. As PCA from subsequent sections and chapters will show, for a LFV ozone field, the *spatial-temporal mean* dominates the ozone variation and it is reliably captured by $\mathbf{E}_1$ and $\mathbf{P}_1$. Therefore, by keeping the mean structure of $\mathbf{O}_{t\times n}$ intact, i.e., no column-centering, we may use the actual data of spatial and temporal means as reference points to assess (1) whether $\mathbf{E}_1$ and $\mathbf{P}_1$ are comparable between AQM and observations, and (2) how "close to correct" the extracted $\mathbf{E}_1$

and $\mathbf{P}_1$ are. Without the mean structure in $\mathbf{O}_{t\times n}$, the 1st-order features may capture dynamic modes of ozone variations that cannot be properly interpreted or assessed for their EOF estimate errors.

Furthermore, except for the feature representing ozone means, PCA of column-centered data $\tilde{\mathbf{O}}_{t\times n}$ gave same set of individual ozone features as those from the PCA of original data (Appendix B.2). The differences lie in the *orders* of said features. Therefore, for our particular LFV ozone data, no additional ozone features or dynamic patterns are uncovered by subtracting out the mean structure.

There are other practical reasons for not centering data. For feature-based AQM evaluation, the space-time mean structure is the most fundamental and important feature that can be compared. Column-centering prior to PCA will process out this key feature from evaluation. Furthermore, in Chapter 4, I will use equation (3.3) to model space-time ozone process using individual ozone features. PCA of $\tilde{\mathbf{O}}_{t\times n}$ means that ozone features can only be constructed to model an ozone process without the mean structure. Earlier in this section, it was also pointed out that all PCA in this research is done on individual data with uniform within-data units and scale. Thus, the usual statistical reasoning behind data centering/standardization does not apply.

Considering the aforementioned results, the ozone PCA in this thesis will be implemented on the original un-centered data $\mathbf{O}_{t\times n}$.

"Degeneracy" of EOFs and Eigenspectrum

North et al. (1982) explained the concept of "effective degeneracy" of EOFs in terms of the sampling error of eigenvalues. As the case with EOFs, eigenvalues λ_j's calculated by the PCA of $\mathbf{O}_{t\times n}$ are estimates of some true eigenvalues of the underlying physical process. If the error associated with a certain λ_j is close to its distance to an adjacent λ_{j+1}, then $\mathbf{E}_j$ and $\mathbf{E}_{j+1}$ form a "degenerate multiplet" where their estimated orders and captured features become arbitrary.

The implication for this research is that "degenerate" ozone features are

ordered and separated arbitrarily, making the same-order feature comparison of AQM and observations difficult to analyze. For instance, if $\mathbf{E}_1$ and $\mathbf{E}_2$ from AQM are inseparable, then the equivalent features from observations may be captured by EOFs of different orders, i.e., the $\mathbf{E}_1$ from AQM and observations may capture different, thus incomparable ozone features. Furthermore, if $\mathbf{E}_1$ and $\mathbf{E}_2$ (from either AQM or observation) form a degenerate couplet, then the ozone features they capture may be "mixed" or "contaminated" into each other (Storch and Zwiers, 1999; Bjornsson and Venegas, 1997). One should note that the last problem is not always the case, as I will show in Section 3.4 with PCA of LFV ozone data.

A generally applied rule of thumb to assess the separation between adjacent eigenvalues is summarized by North et al. (1982) based on asymptotic results. It states that a "confidence interval" of an eigenvalue λ_j is

$$\hat{\lambda}_j \cdot \left(1 \pm \sqrt{\frac{2}{t}}\right), \tag{3.4}$$

where $\hat{\lambda}_j$ is the eigenvalue estimate, and t is the number of hours in $\mathbf{O}_{t\times n}$ or the number of data observations. Due to correlations between observations, one may use "effective sample size", denoted as n^e, $n^e \leq t$, in place of t (Hannachi et al., 2007). Estimates of n^e based on time-series autocorrelation measures have been introduced by Thiebaux and Zwiers (1984) and Preisendorfer (1988), among others. There does not seem to be a definitive approach to estimate n^e, and some existing literature (North et al., 1982; Hannachi et al., 2007) simply applied the sample size n. One example of an estimating equation from Thiebaux and Zwiers (1984), also referenced in EOF review paper Hannachi et al. (2007), has the form (using this thesis's notations):

$$n^e = t\left(1 + 2\sum_{k=1}^{t-1}\frac{1-k/t}{\rho(k)}\right)^{-1},$$

where $\rho(k)$ is the autocorrelation function of order k and t is number of hours in $\mathbf{O}_{t\times n}$.

Another means of estimating the sampling error of an eigenvalue is by

Monte Carlo (MC) based sampling (Bjornsson and Venegas, 1997). One sampling approach, applied in Section 3.4, randomly samples locations from predetermined LFV subregions to form subsamples of space-time ozone data, then PCA is applied to the subsample. Repeated realizations of subsample PCA yield a MC sampling distribution of $\mathbf{E}_j$'s, $j = 1, \ldots, n$, and the eigenspectrum can be constructed. However, the sampling analysis in Section 3.4 is done for purposes other than estimating eigenspectrum.

Rotation of EOFs

By design, PCA extracts data features $\mathbf{E}_j$ that are orthogonal to each other, and this constraint may introduce difficulties in interpreting data features. One reason is that, for a real world physical process the features are not independent, hence the orthogonality of features should not be expected (Monahan et al., 2009). Richman (1986) provided an extensive discussion on the complications facing PCA of spatial processes, two of which are related to our study: "domain shape dependence" and "subdomain stability" of PCA.

Domain shape dependence suggests that the shapes of the extracted $\mathbf{E}_j$'s are influenced by the topography of the studied region, and important underlying physical processes will not be properly captured. Subdomain stability is related to the problem when data from a portion of the domain is used in PCA to make conclusions about dominant modes of a physical process. For example, if an $\mathbf{E}_j$ decomposed from a complete LFV ozone data is different from a same-order $\mathbf{E}_j$ from a sub-regions data from the lower valley, which $\mathbf{E}_j$ captures the true underlying feature?

In our case, one may account for the problem of domain shape dependence by implementing PCA on a part of LFV with similar topography, e.g., locations with low elevation. Such ozone PCA will be implemented in Section 3.4, where the region of interest will be the part of the rectangular LFV (Figure 3.3) where elevations are below 150 meters. In Section 3.4, I will also use a method of spatial sampling to analyze LFV ozone's subdomain stability.

There are also purely mathematical approaches to alleviate the aforementioned concerns of regular PCA, and rotation of EOFs is "perhaps the most widely used" method for such purpose (Hannachi et al., 2007). This procedure rotates the estimated EOFs by maximizing the spatial weights within the $\mathbf{E}_j$'s toward specific regions, potentially highlighting underlying spatial variation/structure and aid EOF interpretation. There are various approaches to the rotation of EOFs, but the general idea is the same. Suppose an EOF rotation matrix $\mathbf{M}^r$ of dimension $p \times p$, where $p \ll \min(n, t)$ is the number of *dominant* data features. Then the rotated EOF $\mathbf{E}^r$ of dimension $n \times p$ is

$$\mathbf{E}^r_{n\times p} = \mathbf{E}[, 1 : p]\mathbf{M}^r_{p\times p}.$$

The matrix $\mathbf{M}^r$ is found by solving the maximization problem MAX $[f(\mathbf{E}[, 1 : p]\mathbf{M}^r_{p\times p})]$ where $f(\cdot)$ represents a function defining a specific rotation method.

A review of the literature suggests that VARIMAX rotation is the most commonly used method, e.g., see the recent work by Eder et al. (2014). Hannachi et al. (2007) also mentioned VARIMAX as being the "most well-known and used" rotation method. In VARIMAX rotation, the maximization of $f(\mathbf{E}[, 1 : p]\mathbf{M}^r_{p\times p})$ is implemented under the constraint $\mathbf{M}^r(\mathbf{M}^r)^{\mathrm{T}} = (\mathbf{M}^r)^{\mathrm{T}}\,\mathbf{M}r = \mathbf{I}_{p\times p}$, i.e., the rotation matrix is orthogonal. The function $f(\mathbf{E}[, 1 : p]\mathbf{M}^r_{p\times p}) = f(\mathbf{E}^r_{n\times p})$ has the form

$$f(\mathbf{E}^r_{n\times p}) = \sum_{k=1}^{p}\left[n\sum_{j=1}^{n}\mathbf{E}^r[j,k]^4 - \left(\sum_{j=1}^{n}\mathbf{E}[j,k]^2\right)^2\right].$$

A numerical feature of VARIMAX is that elements or individual weights in the rotated $\mathbf{E}_j$'s are shifted towards either 0 or ± 1, thus revealing a more focused and simpler spatial pattern for interpretation.

Appendix B.2 shows the comparison plots between normal $\mathbf{E}_j$'s and $p = 4$ VARIMAX rotated $\mathbf{E}_j$'s, where the $\mathbf{O}_{t\times n}$ data is the CMAQ output for the 2006 episode. The results indicate that the utility of EOF rotations is not obvious for the LFV ozone data: the pattern shifting and focusing of $\mathbf{E}_j$'s spatial weights do not result in easier-to-interpret ozone features. This

is also true for higher-order features with closely spaced eigenvalues, i.e., features that form degenerate multiplets. In the end, I found no reason to apply EOF-rotation given our data, and the upcoming feature-based ozone analyses will proceed without further considering EOF-rotation.

Poorly conditioned covariance matrix

Ledoit and Wolf (2004) pointed out that large-dimensional sample covariance matrices are often non-invertible and ill-conditioned, and decomposing such a matrix will give a more "dispersed" sets of eigenvalues than the underlying truth. The authors proposed an optimal covariance matrix estimator that is the linear combination of the sample covariance and scaled identity matrices, where the scales are estimable from the sample and proved to be asymptotically consistent. The idea is to "shrink" the sample covariance towards the identity matrix to be well-conditioned.

Eigen-decomposition does not require a matrix to be full-rank. One can still decompose $\mathbf{O}_{t\times n}$ into $\mathbf{E}_j$'s and $\mathbf{P}_j$'s when the sample covariance matrix is ill-conditioned. The main problem is that the sample $\mathbf{E}_j$ may be poorly-estimated and not allow meaningful further analysis. However, as I will show in Section 3.4, this is not the case here. Moreover, it can be argued that highly dispersed λ_j are preferable in our application. First, it means that a very small number of ozone features can capture most of the data structure. Secondly, by North's rule-of-thumb, well-separated λ_j alleviate the problem of feature degeneracy, thus allowing for individual analysis of the leading features.

Thus the method of sample covariance correction, although a viable alternative, is not attempted.

3.3 The Number of Useful Ozone Features

Before delving into the analysis of ozone features, it is useful to first define criteria for the number p of ozone features that are *useful* for further analysis. A useful ozone feature should at the very least, recover a non-trivial

amount of space-time ozone data variation. Ideally, it should also capture interpretable ozone features that enhance our big picture understanding of an ozone process, i.e., interpretability makes subsequent feature-based ozone modelling (Chapter 4) and CMAQ evaluation (Chapter 5 and 6) more informative.

The number of "useful" EOFs is often determined by the number required to explain a large portion of data variation, so that these EOFs are sufficient to represent the original data. What constitutes "a large portion of variation" is rather an arbitrary decision. Typically the chosen EOFs should combine to explain $\geq 80\%$ of data variation (Higdon et al., 2008; Hannachi et al., 2007).

3.3.1 Recovering Data Variations

Figures 3.5 and 3.6 shows the hourly RMSEs of $\mathbf{O}_{t\times n}$ reconstruction for all five episodes. Here, $\mathbf{O}_{t\times n}$ is the CMAQ output for different episodes over the entire 96 hours with spatial domain containing the locations at elevation$\leq$ 150 meters. The plots show the RMSE by hour between $\mathbf{O}_{t\times n}$ and the reconstructions from $\sum_{j=1}^{p} \mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ done for $p = 1, \ldots, 4$. Note that there is no modelling of $\mathbf{E}_j$'s and $\mathbf{P}_j$'s involved in this result. I simply recovered the CMAQ outputs using increasing numbers of space-time components, used these reconstructions as ozone estimates and calculated RMSEs at each hour.

As shown, the first two space-time features recover most of the daytime ozone variation, with notable exceptions during afternoons of 1998, where the addition of the 3rd ozone components improves the RMSE noticeably. Similar improvements are also observed to lesser extent during last afternoons of 2001 and 2006. Otherwise, the addition of the 3rd feature recovers variations between late evening and morning hours. Subsequent ozone features capture mostly ozone variations during nocturnal hours of diminished ozone activity. This result supports the findings in the following section that successive higher order features capture increasingly *localized* ozone variations in space and time. As I will elaborate in the next section, the description "localized variation" means that the recovered space-time

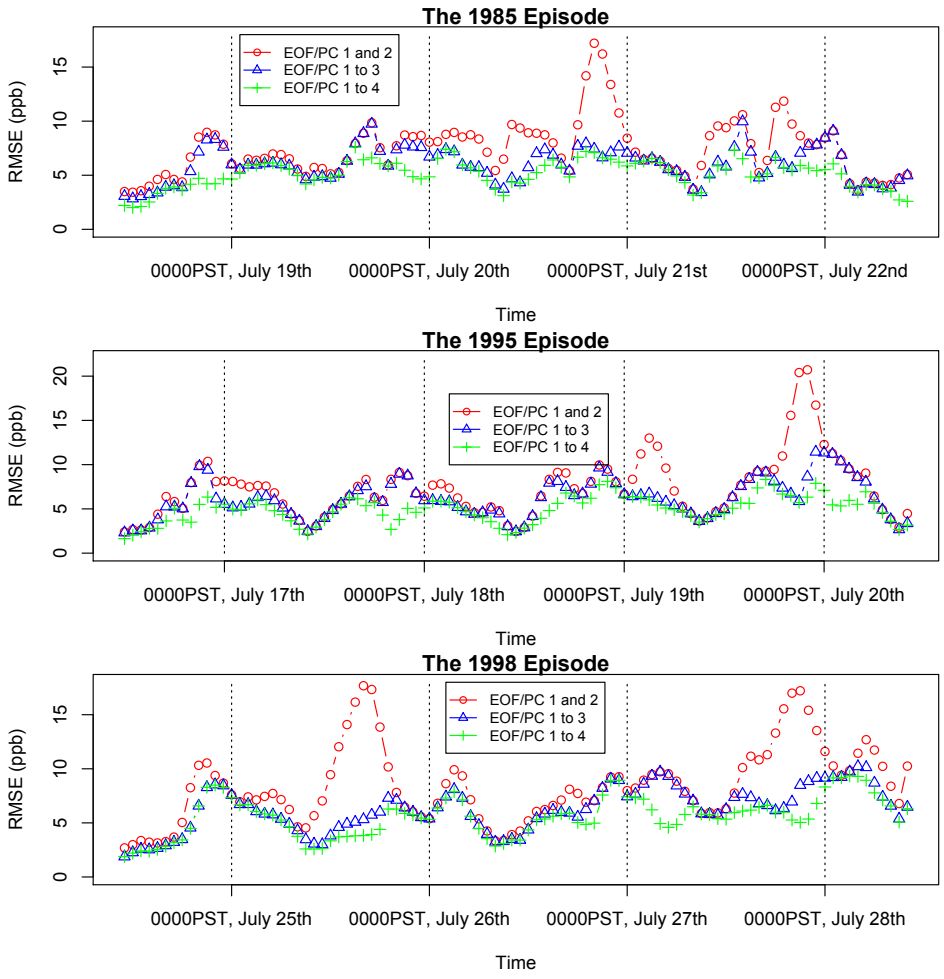


Figure 3.5: Hourly RMSE (units *ppb*) of the $\mathbf{O}_{t\times n}$ *reconstruction* ($\sum_{j=1}^{p}\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$) using an increasing number of data components to $p = 4$. The $\mathbf{O}_{t\times n}$ are the 1985, 1995 and 1998 CMAQ outputs across the LFV at elevation$\leq$ 150m.

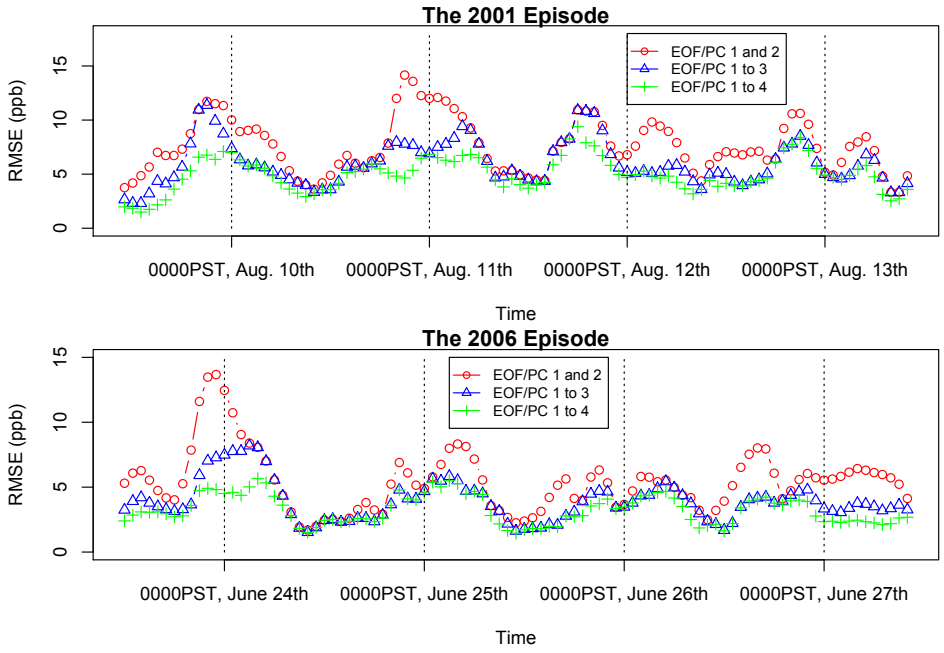


Figure 3.6: Hourly RMSE (units *ppb*) of the $\mathbf{O}_{t\times n}$ *reconstruction* ($\sum_{j=1}^{p} \mathbf{P}_j \mathbf{E}_j^{\mathrm{T}}$) using an increasing number of data components to $p = 4$. The $\mathbf{O}_{t\times n}$ are the 2001 and 2006 CMAQ outputs across the LFV at elevation$\leq$ 150m.

variations are confined to a narrow window of time period and geographical region. Furthermore, these features capture space-time variations that are generally episode specific, and they are neither influential nor structured enough to allow for definite interpretation.

Figures 3.7a and 3.7b show for select hours during day 3 of the 2006 episode, the spatial ozone fields of the CMAQ output (original data) and the corresponding approximation with $p = 4$: $\sum_{j=1}^{4} \mathbf{P}_j \mathbf{E}_j^{\mathrm{T}}$. The hourly spatial plots show that the sum of leading $p = 4$ ozone features appear to recover the defining space-time structure of their originating ozone data.

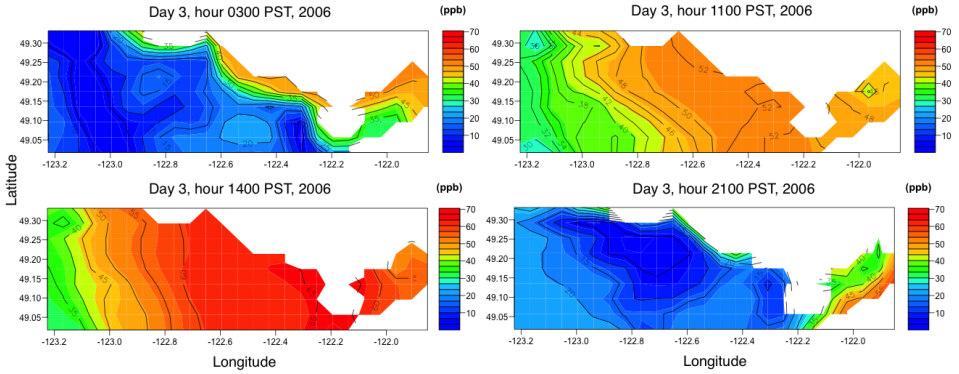
Table 3.2 shows the $\mathbf{O}_{t\times n}$ reconstruction RMSEs averaged across all location and hours for all five episodes, and Table 3.3 shows the proportion of data variation explained by leading ozone features. The main results are consistent across episodes: (1) $\mathbf{E}_1$ and $\mathbf{P}_1$ jointly recover $\geq 90\%$ of data variation and the proportion of recovered variation quickly decreases to $\approx 0\%$ at order $j = 5$ (Table 3.3), and (2) the improvement in RMSE decreases to ≤ 0.6 *ppb* from $p = 5$ onward (Table 3.2).

	p: number of features used for reconstruction							
Episode	1	2	3	4	5	6	7	8
1985	10.4	7.57	6.01	5.15	4.78	4.34	4.04	3.93
1995	10.8	7.82	6.32	5.17	4.72	4.28	3.87	3.63
1998	13.2	8.94	6.37	5.69	5.25	4.71	4.25	4.04
2001	11.1	7.84	6.23	5.16	4.72	4.35	3.92	3.77
2006	7.66	5.80	4.07	3.27	2.83	2.62	2.39	2.29

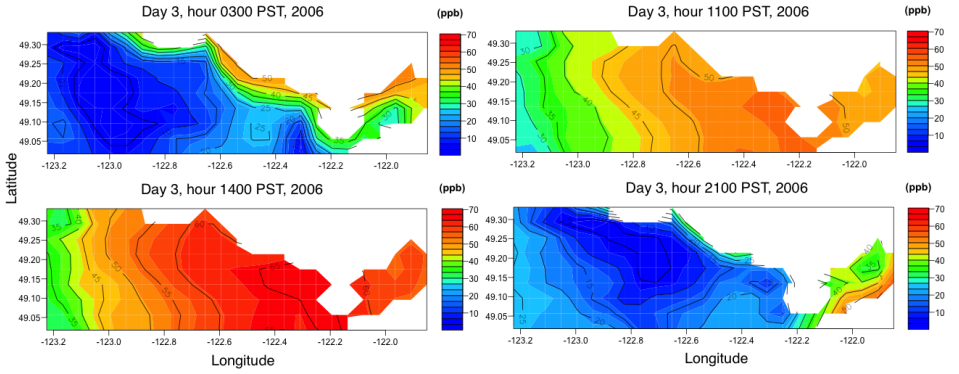
Table 3.2: For the CMAQ outputs of 5 episodes: the $\mathbf{O}_{t\times n}$ reconstruction RMSEs in units *ppb* (averaged across all location and hours) at $p = 1, \ldots, 8$. The decompositions are for the entire episode $h = 96$ hours.

Furthermore, ozone feature analysis in Chapters 4 to 6 will model ozone features as spatial or temporal processes driven by background meteorology, chemical precursor emissions and antecedent concentrations. It will be shown that statistical models begin to lose their predictive capability with high-order ozone features $j \geq 3$, suggesting diminishing associations between these features and process-driving background conditions.

In summary, for real CMAQ outputs of most episodes, $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$ and $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$



(a) Original ozone fields.



(b) Ozone fields recovered with leading $p = 4$ ozone features.

Figure 3.7: For selected hours, the spatial field of the 2006 CMAQ output and corresponding feature-based data reconstruction using $p = 4$: $\sum_{j=1}^{4} \mathbf{P}_j \mathbf{E}_j^{\mathrm{T}}$. The colour scales are held consistent for all plots.

	EOF/PC order j				
Episode	1	2	3	4	5
1985	0.91	0.04	0.02	0.01	0.00
1995	0.91	0.04	0.01	0.01	0.00
1998	0.93	0.04	0.02	0.01	0.00
2001	0.90	0.05	0.02	0.01	0.00
2006	0.95	0.03	0.02	0.01	0.00

Table 3.3: For the 5 ozone episodes: the proportion of data variation explained by the leading 5 EOFs/PCs. The decompositions are done for the entire episode $t = 96$ hours.

are sufficient for the analysis of daytime ozone process in terms of recovering variations in the data. Features of orders $j = 3, 4$ recover the remainder of daytime ozone variations and most of the variations outside of daytime ozone peak hours. The ozone features of orders $j \geq 5$ do not recover significant amounts of space-time variation.

I also implemented a simulation-based analysis (Appendix B.1) where the ozone feature sum $\sum_{j=1}^{p} \mathbf{P}_j \mathbf{E}_j^{\mathrm{T}}$ from one synthetic dataset is used to predict another realization of a space-time ozone field, the procedure is repeated a large number of times for various values of p. I then found that the predictive accuracy starts to deteriorate at order $p = 3$. This indicates a lack of structure, i.e., domination of data noise within the ozone features of orders $j \geq 3$. Since the synthetic ozone field emulates the structures of the daytime LFV ozone field, the simulation-based statistical test implies that there is no practical reason for using more than $j = 2$ ozone features to model a complete daytime process. This result are consistent with the aforementioned results using real CMAQ outputs.

The just described simulation-based approach combines the elements of the PC selection methods proposed in the Chapter 5 of Preisendorfer (1988), which are categorized into three general selection rules. In "Dominant Variance" selection rule, synthetic data are generated from assumed Gaussian processes, these data are then decomposed to obtain a sample of eigenvalues. An EOF/PC pair from the real data is retained if their associated eigenvalue is above the 95th percentile of the sample eigenvalue (called Rule N). The

"Time History" selection rule uses statistical tests of spectral whiteness of serial correlation to assess the nosiness of PCs, which determines the PCs to keep. Lastly, in the "Space-map" rule, $\mathbf{E}_j$ decomposed from data are compared to a set of well-understood modes of variation, such as those from a General Circulation Model. Any $\mathbf{E}_j$ from the real data that has a close match (based on tests of conical direction angles) to a dynamic mode is retained.

3.3.2 Order of Ozone Feature Degeneracy

The number of useful ozone features may also be determined through the spectra of PCA eigenvalues. From the eigenspectra, one may judge the order j at which $\mathbf{E}_j$'s begin to lose their separability, thus making individual feature analysis questionable.

Figure 3.8 shows for the 1985 and 2006 episodes, the eigenspectra composed using North's rule-of-thumb described in Section 3.2. Here, the 1st eigenvalue is not shown because its high value makes the higher-order eigenspectrum difficult to visualize. The PCA is based on different datasets: for 1985, the dataset is from its last two full days which are dominated by type IV wind regime, and for 2006, the dataset is the first two full days that are driven by type I regime.

It was mentioned in Section 3.2 that various estimates of effective sample size n^e exist, most of which are based on autocorrelation measures calculated form the data. I tried methods described in Thiebaux and Zwiers (1984) and Preisendorfer (1988), and found all estimated n^e resulted in eigenspecturms that give the same overall conclusion regarding the separability of leading features. Both plots are made using $n^e = 18$, which is lower than all estimated n^e. Hence, based on equation (3.4) the width of the confidence band can only be narrower, i.e., the order of EOF degeneracy can only get *higher* than those shown.

The 1985 eigenspectrum shows that for ozone fields dominated by the type IV regime, although eigenvalues λ_2 and λ_3 are significantly higher than the rest, they cannot be separated themselves. For the 2006 episode, the

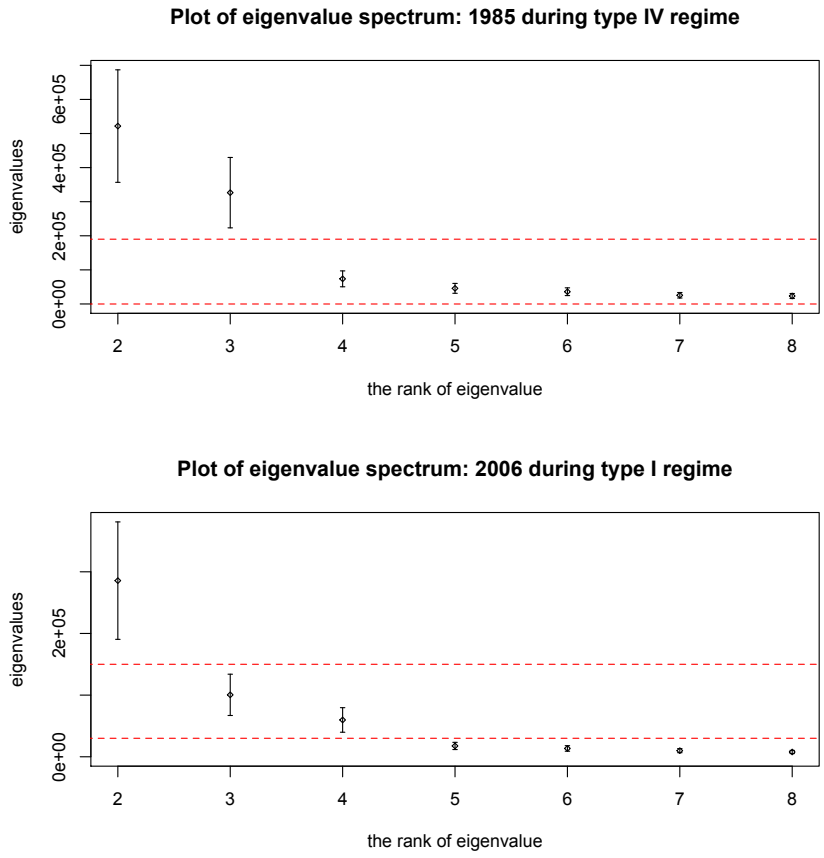


Figure 3.8: The eigenspectra of λ_j, $j = 2, \ldots, 10$, decomposed from the the 1985 (top) CMAQ output under type IV regime and 2006 (bottom) outputs under type I regime. The spectrum is based rule-of-thumb of North et al. (1982). The dashed-lines indicate the orders of ozone features that can be separated from the rest, whether individually or as degenerate sets.

first 2 ozone features are separable from the rest, while the 3rd and 4th-order ozone features form a couplet. This latter result is the case for all episodes *not* dominated by type IV wind regime.

These results from the PCA eigenspectrum are crucial when analyzing individual ozone features, and Figure 3.8 will be revisited in the next section. The evidence from eigenspectra suggests that for ozone process driven by wind regime types I, II and III, the first two ozone features can be studied independently. For episodes dominated by the type IV regime, the 2nd and 3rd-order ozone features should be interpreted jointly, or analyzed with consideration of the other.

Various analyses presented in this section and Appendix B.1 are designed to help answer the question: how many ozone features are useful? The answer is that ozone features of order $j = 1, 2, 3, 4$ should be the focus. However, given the specific contexts and foci of upcoming statistical analyses, the number of useful ozone features will be less. The reasoning behind the use and analysis of any particular feature will be discussed at the appropriate places in following chapters.

3.4 Ozone Features of LFV Ozone Episodes

In this section, PCA will be implemented on CMAQ ozone outputs to identify and understand what types of ozone features define the LFV ozone field during an episode. Such an exercise also determines if interpretable ozone features can be obtained through the PCA of space-time ozone data. Here, I am not attempting to interpret physical modes of variation. By "interpretable", I am referring to an ozone feature that either represents statistical summaries of data, or captures recognizable LFV ozone structures and behaviour, e.g., general patterns of diurnal ozone advection across LFV. Furthermore, the ozone feature analysis is done by accounting for the non-separability or "degeneracy" of multiple ozone features. Lastly, the end of this section contains a study of LFV's PCA subdomain stability.

PCA is implemented on CMAQ outputs by episode, and CMAQ outputs where days are separated into wind regime types. It should be noted that I

do *not* combine datasets from different episodes under the same wind regime. PCA of each wind regime type is instead analyzed episode-by-episode. PCA of specific regime within the same episode helps to analyze the effect of regional wind pattern on dominant ozone features. Same episode PCA such as this controls the effect of changing emission source distribution over the years (Section 2.2). PCA based on wind regime is also a means of CMAQ evaluation without comparison with observations. It evaluates how well CMAQ can capture the effect of regional wind flow on its modelling of LFV's ozone process.

The PCA results to be shown are implemented on ozone fields at elevations ≤ 150 meters. Such datasets capture the ozone field across a roughly triangular region covering the lower valley of the "rectangular" LFV (Figure 3.3), with Chilliwack on the eastern edge. I will simply refer to this lower valley region as "LFV". LFV is where the physical ozone monitoring stations are located, and it is the region of interest for CMAQ model evaluation.

In this study, I also implemented PCA on data of dimension $n \times t$. The resulting ozone features (not shown) are equivalent to the upcoming results, where the spatial and temporal features are captured by $\mathbf{P}_j$ and $\mathbf{E}_j$ instead.

3.4.1 Common Ozone Features of All Episodes

Figure 3.9 shows the spatial plots of temporal ozone means (the mean field) summarized from the 96-hour CMAQ outputs. The same figure also shows the spatial plots of the 1st-order features $\mathbf{E}_1$ extracted from the same CMAQ outputs. Figure 3.9 shows that the spatial patterns of mean fields are similar between episodes, and the pattern of each mean field is accurately and reliably captured by the corresponding $\mathbf{E}_1$. Eastern LFV has distinctly higher temporal means than the west. This result indicates that the eastward ozone advection (in combination with low boundary layer heights, Section 3.1) causes the eastern LFV to experience continuous high level of ozone pollution, thus higher temporal means. The same advection pattern also gives the western LFV, the main source of ozone formation, a diurnal cycle of increase-peak-decrease. Thus, western LFV locations have smaller temporal

means.

Figure 3.10 shows the spatial plots of ozone standard deviations (calculated across time) and the 2nd-order spatial features $\mathbf{E}_2$ of each episode. As shown, $\mathbf{E}_2$ captures the spatial pattern of ozone standard deviation, and its values range from negative to positive. The spatial variation of ozone standard deviation can be explained by aforementioned behaviour of LFV's ozone process: pronounced ozone fluctuation in the west and continuous high pollution in the east. Hence, once summarized over time, the western LFV locations have higher ozone standard deviation than the eastern locations.

As I will elaborate in Section 3.4.3, $\mathbf{E}_2$ can be interpreted as the *spatial ozone contrast* between the area of ozone plume formation (western LFV) and the area where most of the ozone move to (eastern LFV). More importantly, the interactive space-time feature $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ will be shown to capture a specific pattern of eastward ozone advection, as well as the magnitudes of ozone formation and destruction across LFV.

Figure 3.11 shows the hourly spatial ozone means (averaged across space) and $\mathbf{P}_1$ decomposed from the 5 ozone episodes. As in Figures 3.9 and 3.10, the PCA is based on the entire 96-hour period of each episode. Hence these $\mathbf{P}_1$ are the temporal ozone features associated with the spatial features $\mathbf{E}_1$ in Figure 3.9. As shown, the $\mathbf{P}_1$ captures the temporal pattern of hourly LFV mean ozone, and the temporal patterns of 1st-order ozone features between-episodes are near identical. One exception is the last afternoon of 1985, where we see a bi-modal ozone peaks not evident in other episodes, and this pattern is captured by the 1985 $\mathbf{P}_1$.

Lastly, Figure 3.12 shows the temporal patterns of hourly LFV ozone standard deviations and $\mathbf{P}_2$ from the 5 ozone episodes. These time series indicate that the standard deviations of ozone across space varies temporally in a less consistent way from one episode to another. It also shows that each 2nd-order temporal ozone feature $\mathbf{P}_2$ captures a pattern of *temporal ozone contrasts* that cycles diurnally. Temporal trend of $\mathbf{P}_2$ further corresponds to a smoothed and inverse curve of hourly LFV standard deviation over the 96 hours.

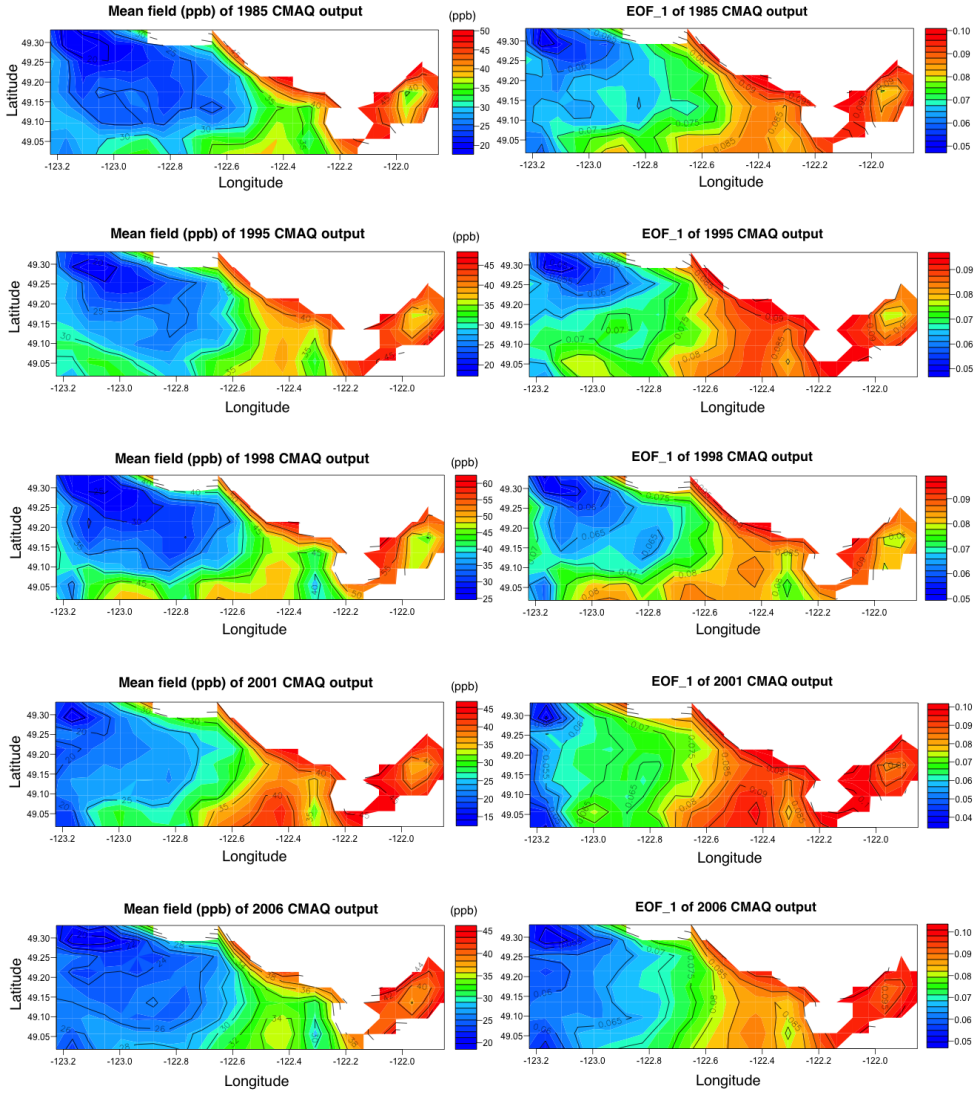


Figure 3.9: Plots of the mean fields and $\mathbf{E}_1$'s of the 5 ozone episodes. The PCA is implemented on the ozone episodes in their entirety (96 hours).

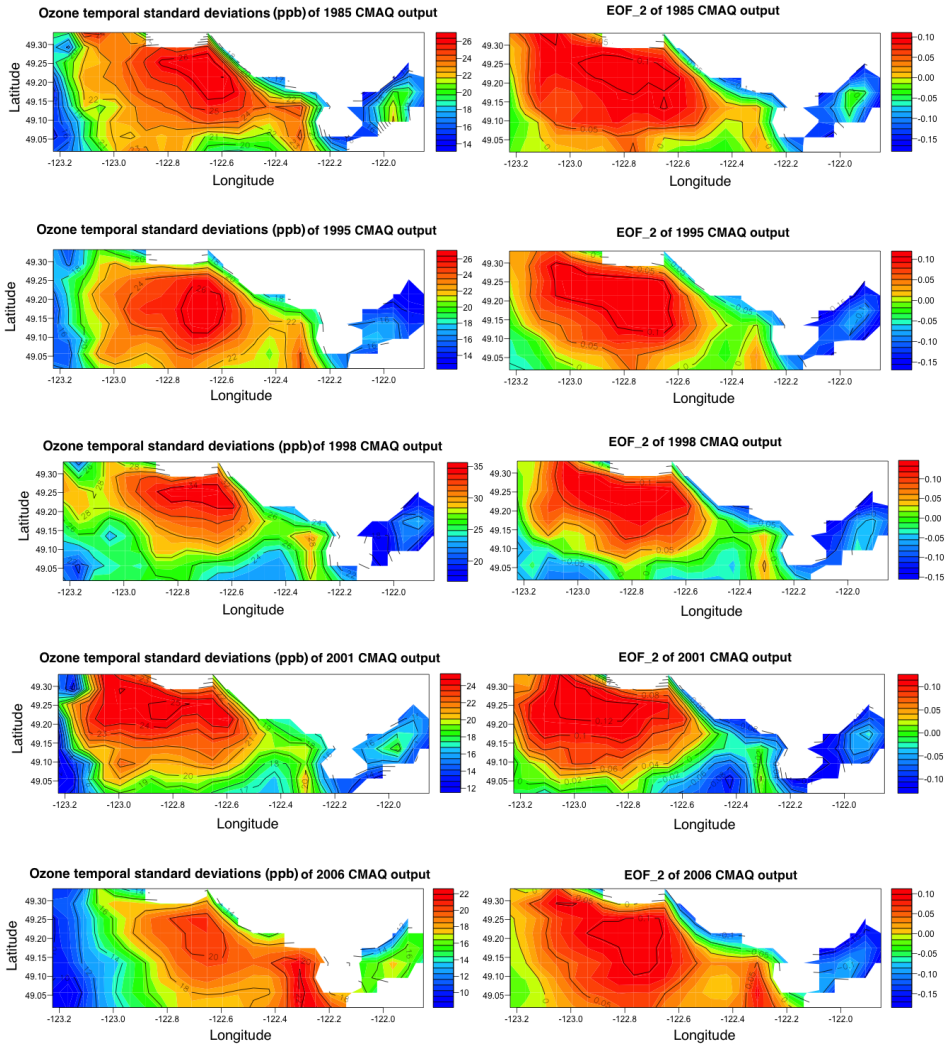


Figure 3.10: Plots of spatial field of temporal ozone standard deviations (calculated across time) and $\mathbf{E}_2$'s of the 5 ozone episodes. The PCA is implemented on the ozone episodes in their entirety (96 hours).

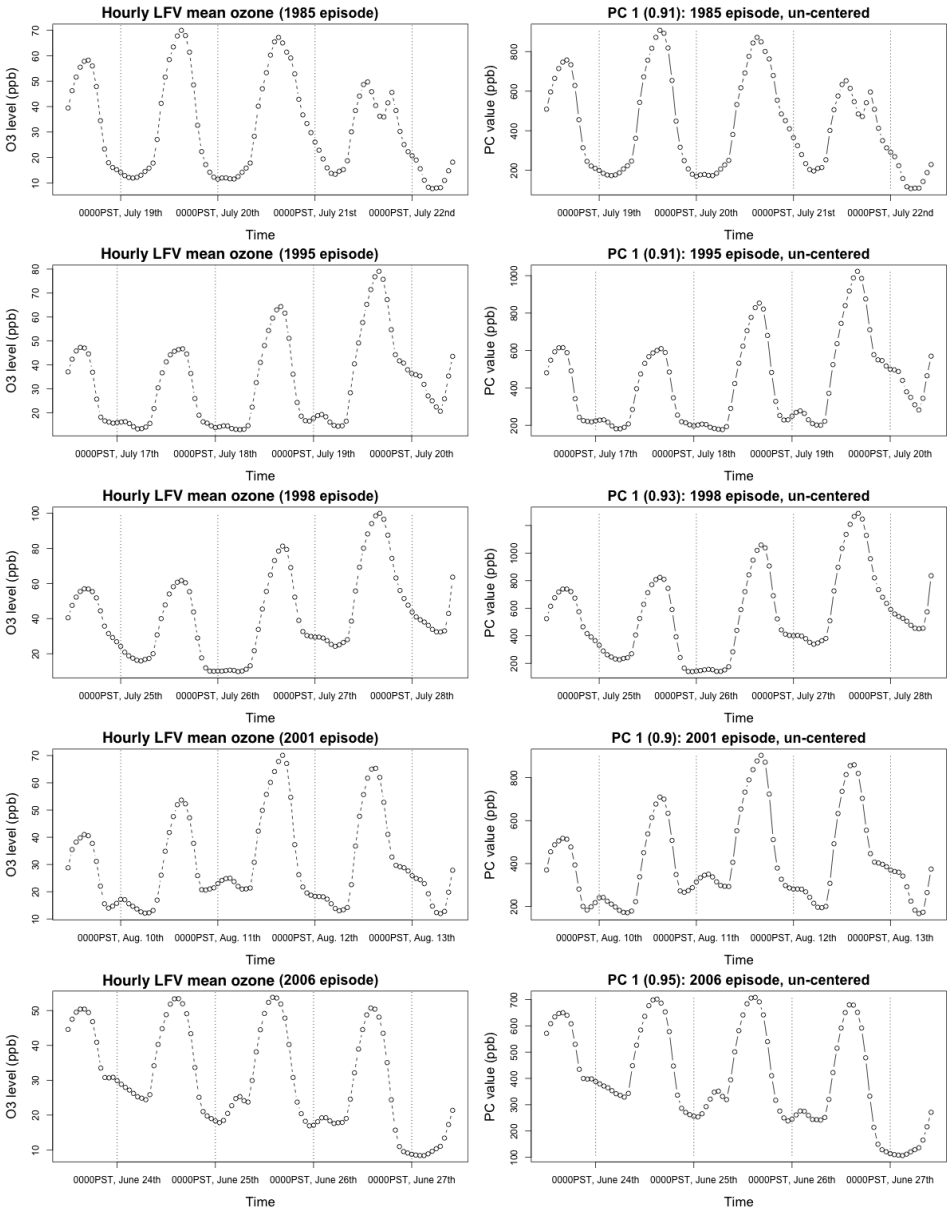


Figure 3.11: Time series plots of hourly spatial (LFV) ozone means and $\mathbf{P}_1$'s of the 5 ozone episodes. The number in each PC plot heading is the proportion of data variation explained. All plotted data have units *ppb*.

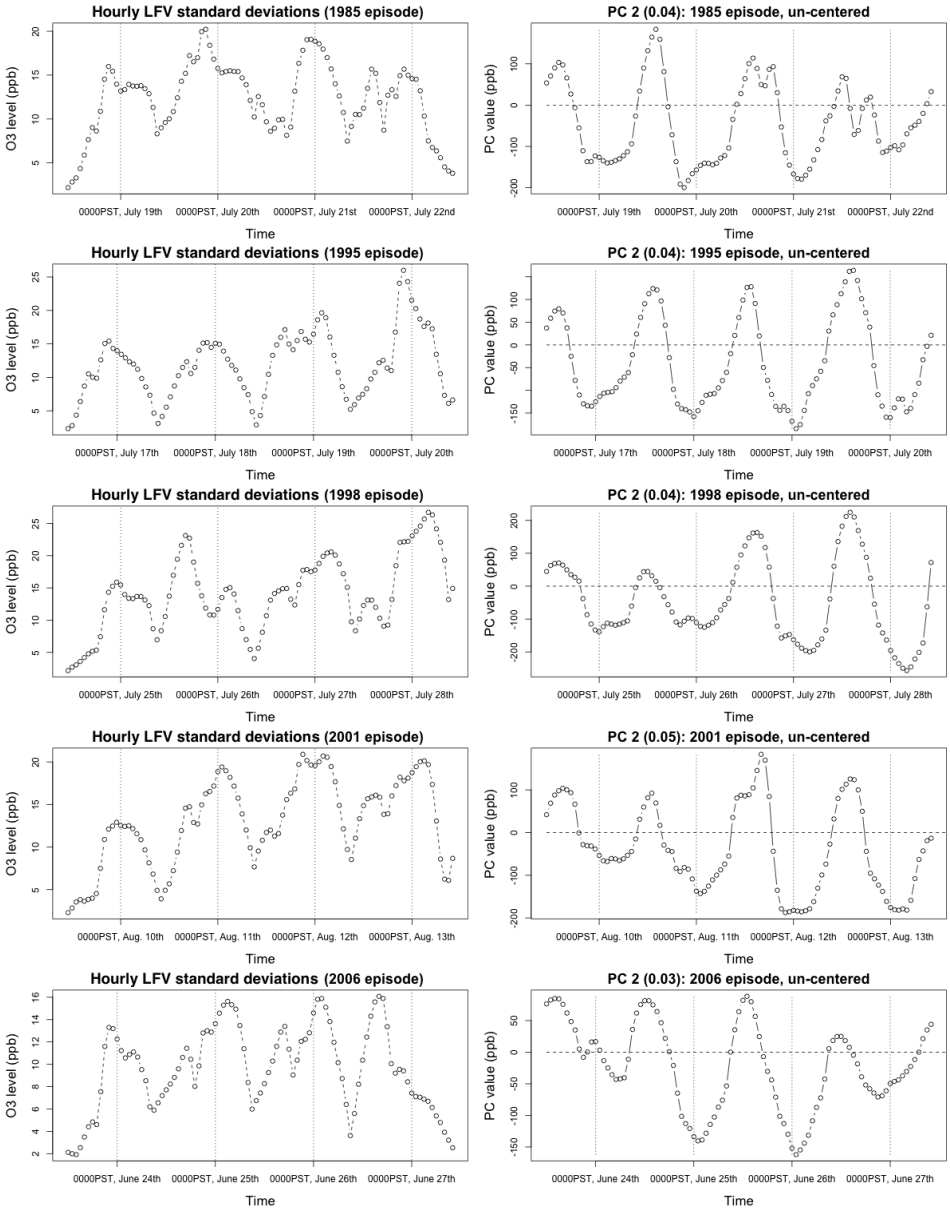


Figure 3.12: Time series plots of hourly LFV ozone standard deviations (calculated across space) and $\mathbf{P}_2$'s of the 5 ozone episodes. The number in each PC plot heading is the proportion of data variation explained. All plotted data have units *ppb*.

Figures 3.9 to 3.12 illustrate an important point that LFV ozone is consistently dominated by the *same* systematic ozone structures during every episode, and they can be reliably captured by the first 2 EOFs and PCs. These results highlight the stable recurring nature of dominant LFV ozone features, despite the episodic variations in emissions and wind regimes. In the remainder of this section, I will interpret ozone features in terms of their space-time interactions $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$, i.e,, interpreting the spatial-temporal ozone features.

3.4.2 $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$: Structure of Space-Time Ozone Mean

Figures 3.13 gives the *dynamic spatial plots* of space-time feature $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$ of the 2006 episode. Each space-time ozone feature $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ is a data matrix of dimension $t \times n$, where each row i, $i = 1, \ldots, t$, relates to the spatial ozone feature of that particular hour i. The presented dynamic spatial plots are for selected hours from the 3rd day of 2006.

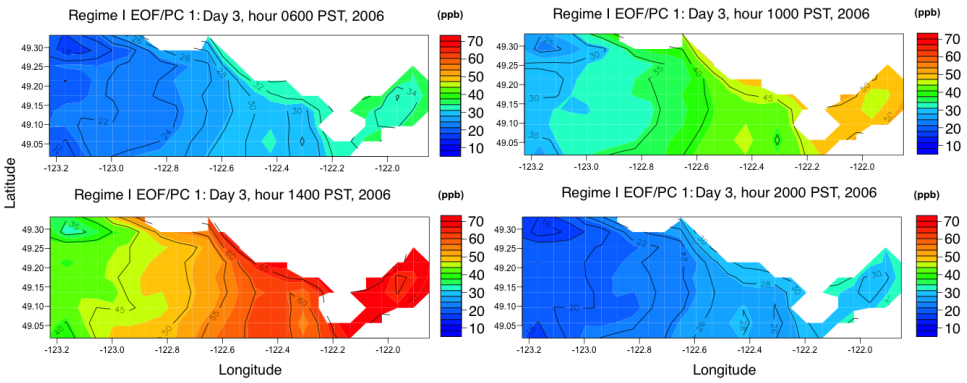


Figure 3.13: From the 2006 ozone episode under the type I regime: spatial plots of $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$ (units *ppb*) at selected times shown in plot headers.

When multiplied, $\mathbf{P}_1$ and $\mathbf{E}_1$ capture the space-time interaction between the spatial and temporal ozone means; a spatial-temporal feature that represents the underlying mean structure of the data $\mathbf{O}_{t\times n}$. As the episode progresses (hour changes and different rows of $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$ selected), $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$'s hourly pattern of ozone variation remain as defined by $\mathbf{E}_1$, only the spatial values

are scaled by $\mathbf{P}_1$ at each hour. $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$'s from other episodes capture very similar dynamic structure illustrated in Figure 3.13. The aforementioned general pattern of the 1st-order ozone feature (as well as ozone means) remains consistent when ozone PCA is done based on wind regime type (not shown).

Lastly, without subtracting the column means out of $\mathbf{O}_{t\times n}$, the space-time mean structure $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$ dominates LFV's regional ozone variation. It account for $> 90\%$ of data variation of any ozone episode or ozone data under any wind regimes. The exact proportions were shown in the header of each $\mathbf{P}_1$ plots and summarized in Table 3.3 from Section 3.3.

3.4.3 $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$: Dominant Patterns of Ozone Advection

In the following discussion, the joint spatial-temporal ozone features of orders $j \geq 2$ are extracted from ozone data under individual wind regime types. I should note again that I do not combine dataset from different episodes under the same regime. PCA of each regime type is done episode-by-episode: each decomposed dataset is the part of an episode under one specific regime. Similar analysis can still be made based on PCA of entire episodes, but ozone feature discussions under the context of background wind regime provide a clearer picture of the way CMAQ models ozone advection across LFV.

Figure 3.14 shows the dynamic spatial plot of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ for the 2006 CMAQ output under type I wind regime (first 2 full days). The result reveals that the second ozone feature captures the dynamic evolution of *spatial ozone contrast* between the west and the east of LFV. More specifically, $\mathbf{E}_2$ captures the spatial contrast between the area of ozone plume formation (western LFV) and the area that is the destination of ozone advection. The term "contrast" is also used in Jin et al. (2011) to highlight two ozone regions with contrasting signs.

During the afternoon ozone-peak hours, the spatial contrast is positive in the west and negative in the east. The spatial contrast then reverses sign from 2000PST onward, and this diurnal evolution of ozone contrast is repeated throughout the episode. This dynamic alternation of contrast sign

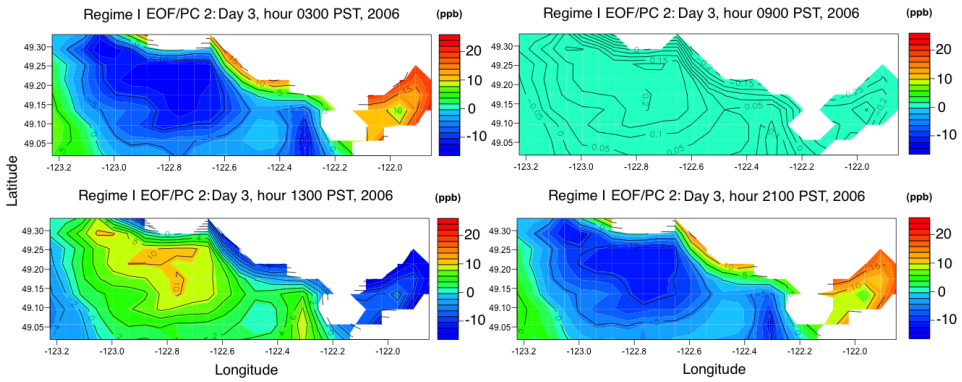


Figure 3.14: From the 2006 ozone episode under the type I regime: spatial plots of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ (units *ppb*) at selected hours.

is the result of $\mathbf{E}_2$ and $\mathbf{P}_2$ representing corresponding spatial and temporal ozone contrasts (Figures 3.10 and 3.12). As discussed, $\mathbf{E}_2$ has positive contrast in the west and negative contrast in the east. Regardless of the episode, $\mathbf{P}_2$ is a dipole wave that alternates between (+) during daytime and (−) during night time to early morning. Thus, the interaction of $\mathbf{E}_2$ and $\mathbf{P}_2$ results in the type of dynamic spatial contrast shown.

Interpretation of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$: the dominance of westerly wind flow

The first ozone feature $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$ contains only positive values, and it captures the underlying structure of the space-time ozone mean. Each element in $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$ represents the base ozone concentration at a particular location and time. Higher order ($j = 2, 3, \ldots$) ozone features have values ranging from negative to positive; they are ozone correction or adjustment terms that in successive order of $j = 2, 3, \ldots$, subtract or add ozone values that are specific to each location and time.

During the afternoon peak hours (1300PST, Figure 3.14), contrast is positive around western LFV, indicating the formation of ozone plume in the west. By evening (2100PST, Figure 3.14), the $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ values changes to negative in the west, indicating that ozone plume is transported out of this region between 1300PST and 2100PST. During the same hours, the eastern

LFV (Chilliwack to be specific) has contrast change from (−) to (+): the ozone from the west are transported to the east. Therefore, this $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ captures a processes of west-to-east ozone advection driven by a general westerly wind system. Its order of decomposition further indicates that this westerly wind is the most dominant flow regime of LFV ozone.

The magnitudes of positive and negative ozone contrasts further reveal the numerical amount of ozone formation or destruction based on the ozone means $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$. For example, Figure 3.14 shows that at 1300PST, the center of ozone formation (western LFV) creates about 10 *ppb* of ozone in addition to the underlying ozone means. At 2100PST, the contrast is generally at -10 *ppb*, indicating that ozone is lost by this amount due to both advection and local photochemical reactions. At 2100PST in Chilliwack (eastern tip of the map), the positive contrasts show that this area gained around 10-15 *ppb* of ozone due to both the transport of pollution system from the west and local ozone creation.

The hourly values of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ are near 0 *ppb* during the "transitional" hours when the positive contrast switches from west to east. This implies that during these hours the pattern spatial ozone resembles the underlying mean.

$\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ of Episodes dominated by Type I, II and III Wind Regime

All 2nd-order ozone features $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ from episodes dominated by type I, II and III wind regimes capture the same general structure of dynamic east-west ozone contrast (Figure 3.14). As I will now show, although differences in spatial patterns do exist between $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$'s, they are subtle. In following discussions I will use "type I ozone feature" as short for "the feature of ozone fields dominated by type I wind regime", and so forth.

The 2006 episode is dominated by wind regime type I on the first 2 full days and type III on the 3rd day, hence comparison of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ between wind regime types I and III is done through the PCA of the two subsets of 2006 CMAQ output. Selected dynamic spatial plots of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ under regime types I and III are shown in Figure 3.15. When compared to the feature under

type III regime, the type I feature has positive contrast covering a smaller area of LFV during the hour of daily ozone peak (1300 PST). At night time the negative contrast is also larger in magnitude for the type I feature.

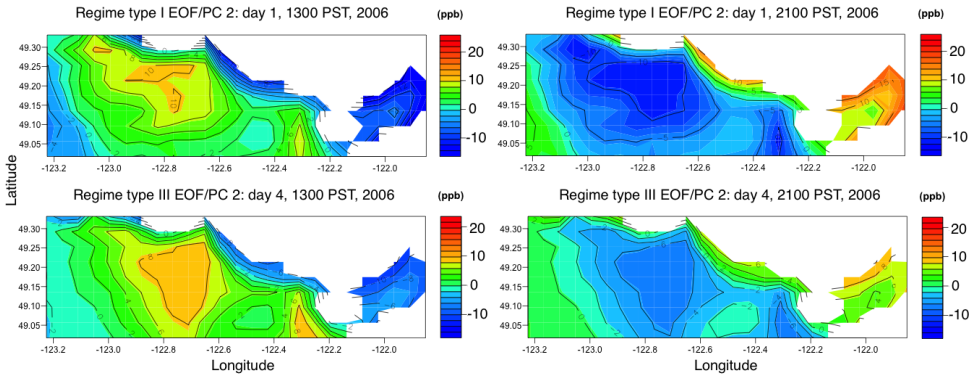


Figure 3.15: From the 2006 episode under type I (top) and III (bottom) wind regime: $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ (units *ppb*) from the same selected hours.

Figure 3.16 shows the $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ from the 2001 CMAQ output, which is dominated by a type II wind regime. As shown in Figure 3.4, the directional flow under the type II regime stays close to 290° throughout the day, so it is defined by a stable northwesterly flow regime. The PCA results show that the $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ of 2001 episode (all type II) and 1998 episode (type II for two days) also captured the same form of dynamic east-west ozone contrasts as seen in type I and III features. One unique pattern for the year 2001: the eastern ozone contrast extends beyond Chilliwack to include a large portion of Abbotsford, whereas the area of eastern contrast from other wind regime types are focused solely around Chilliwack.

In summary, any variations between the space-time patterns of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ under regime types I, II and III are slight. and they do not affect the overall conclusion that $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ under these wind regimes represent the same form of space-time ozone contrast. Interpreted as atmospheric process, preceding analyses reveals that the eastward ozone advection, driven by a westerly wind system, is the most dominant advection mechanism of LFV ozone process under the type I, II and III wind regimes.

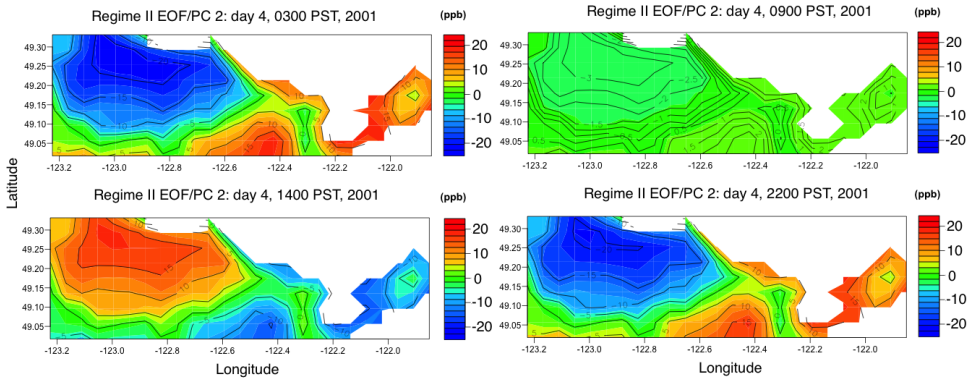


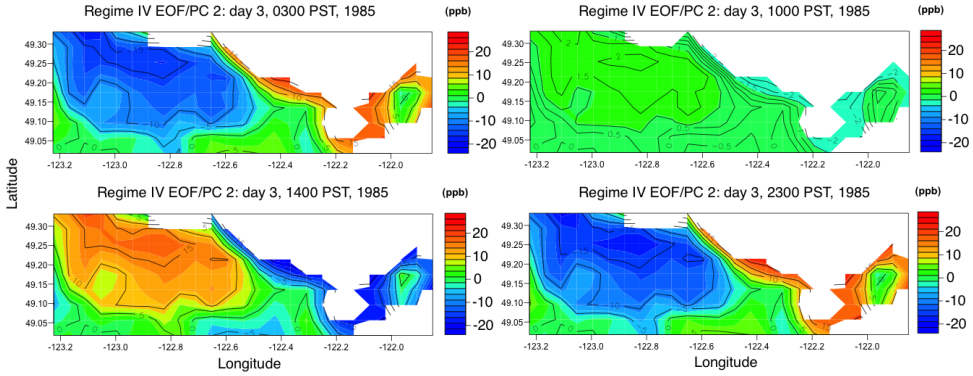
Figure 3.16: From the 2001 episode under type II wind regime: $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ (units *ppb*) from selected hours.

Ozone features of the 1985 episode: type IV wind regime

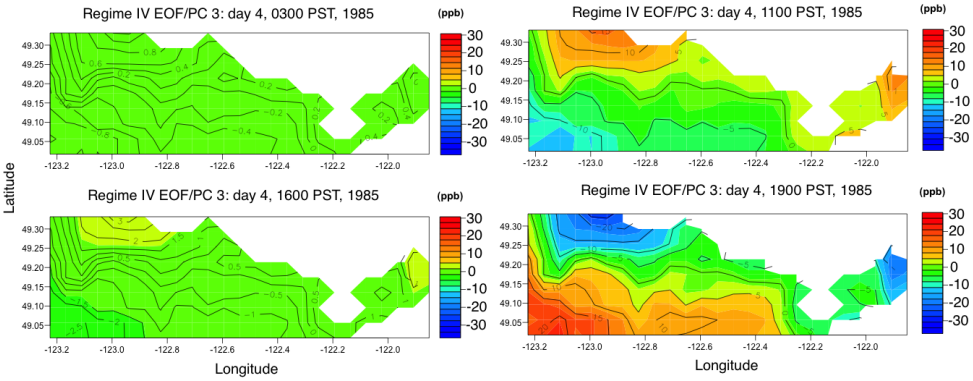
From Figure 3.4, one can see that the type IV regime is dominated by easterly winds for most of the day, and westerly wind is only observed during a few afternoon hours. Figure 3.1 showed the 3-dimensional ozone fields from selected hours of the 1985 episode under the type IV wind regime. One main feature of this episode is the heavy accumulation of ozone in the southwest LFV during night time, a feature of which is not noticed during the 2006 episode (Figure 3.2).

The following results are based on the PCA of CMAQ ozone output for the 3rd and 4th full days of 1985: the days dominated by type IV wind regime. Figure 3.17a shows $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ from selected hours. As shown, the 2nd-order ozone feature still captured a form of dynamic east-west ozone contrast. However, it differs from the preceding results (features of type I to III regimes) in that the east-west alternation of spatial contrast takes place around 2300PST to midnight, not evening. Furthermore, this feature did not capture the aforementioned night-time ozone pollution around southwest LFV.

Figure 3.17b shows the dynamic spatial plots of $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ at selected hours. This feature captures a diurnal evolution of spatial ozone contrast between the area around Vancouver's city core (northwest LFV) and the southwest



(a) Hourly plots of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ (in units *ppb*).



(b) Hourly plots of $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ (in units *ppb*).

Figure 3.17: From the 1985 episode under type IV regime: dynamic spatial plots of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ and $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ (in units *ppb*).

of LFV. More specifically, the contrast in northwest LFV transitions from positive in the morning to negative in the evening, while the opposite is true for southwestern LFV. This pattern of dynamic contrast shows that $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ captures the advection of ozone plume from northwest LFV to the southwest. Moreover, $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ captures spatial contrasts between morning and evening, which is different from the 2nd-order feature which contrasts afternoon peak hours and late night.

Figure 3.18 shows the dynamic spatial plot of joint ozone feature $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}} + \mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$. During the afternoon, instead of an eastward advection, the ozone plume circulates from northwest LFV towards the southwest. This is evident from the movement of positive ozone contrast from the northwest to the southwest between 1300PST and 2000PST. The positive contrast then transitions to the eastern tip of LFV at the conclusion of a diurnal cycle. Therefore, $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ and $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ *jointly* capture the continuous space-time mechanism of LFV ozone driven by the unique flow pattern of the type IV regime. Here, the space-time structure is defined by an advection pattern of northwest→southewest→east. This complex dynamic is expressed via east-west spatial contrast captured by $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ and north-south spatial contrast captured by $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$.

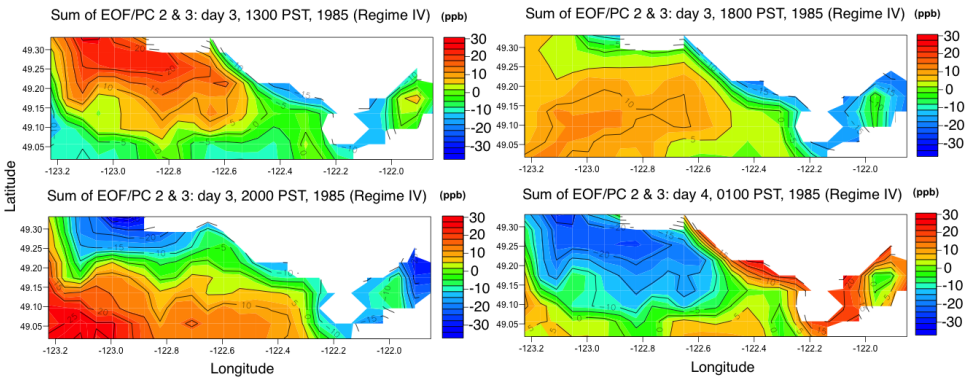


Figure 3.18: From the 1985 episode under type IV regime: dynamic spatial plot of joint ozone features $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}} + \mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ (units *ppb*).

As shown in the eigenspectrum analysis (Figure 3.8) of Section 3.3, for

the two days of 1985 under type IV regime, the ozone features of orders $j = 2, 3$ form a degenerate couplet. The main implications are: (1) the extracted patterns of these features are possibly mixed into each other, making individual interpretations difficult, and/or (2) the orders of feature extraction are arbitrary. The results in this section do not support the first implication. As shown in Figures 3.10 and 3.17a, the $\mathbf{E}_2$ from 1985 captured the same general contrast pattern as other episodes. This means that $\mathbf{E}_2$ from 1985 still managed to clearly capture the defining east-west contrast of LFV, and that there is no significant evidence of EOF-estimate error or mixing of ozone features between $\mathbf{E}_2$ and $\mathbf{E}_3$. Therefore, the above mentioned feature degeneracy indicates that for an ozone field under the type IV regime, $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ and $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ capture well defined individual features that are equally important and should be analyzed jointly.

Summary discussion of Section 3.4.3

The general structure of $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ remained consistent across episodes dominated by wind regime types I, II and III, all of which are defined by westerly wind flow. The conclusion here is that $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ captures the space-time process of eastward ozone advection driven by a westerly wind system, and it is the most important advection pattern of LFV ozone. The ozone field under type IV regime has more complex structure. The PCA results show that the ozone advection towards southwest LFV represents an equally important dynamic behaviour as the eastward advection.

The hodograph in Figure 3.4 (from Ainslie and Steyn (2007)) showed that the type IV regime is dominated by easterly flow for most of the day. However, the ozone features from the 1985 episode revealed a more complex pattern of ozone advection that cannot be clearly explained by this prior knowledge. The wind regime hodograph was made from average of observed wind speed and direction at YVR. Hence, preceding ozone feature analyses showed that the type IV wind regime should be defined by a regional-scale flow pattern that cannot be properly captured by point-based information at YVR. Steyn et al. (2011) also raised the similar point that the wind data

from YVR sometimes may not capture the complexity of LFV's regional wind. However, Ainslie and Steyn (2007) did correctly categorize the 1985 episode as driven by its unique wind regime, which was shown (in this section) to create dynamic ozone features that are different from the features under regimes types I to III.

3.4.4 Higher-order Ozone Features

Figure 3.19 shows the dynamic spatial plots of $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ and $\mathbf{P}_4\mathbf{E}_4^{\mathrm{T}}$ from the 2006 episode under type I regime. As shown, the higher-order features $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ and $\mathbf{P}_4\mathbf{E}_4^{\mathrm{T}}$ are "active" mainly between late evening and early morning. The term "active" indicates that an hourly $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ spatial field contains moderately non-zero values, i.e., the feature $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ recovers non-trivial ozone variation from the data. Therefore, the 3rd and 4th-order data features may be viewed as *nocturnal ozone features*, where they recover ozone variations during hours with minimal ozone formation and destruction. Ozone features of orders $j \geq 3$ under regime types II and III, as well as type IV features of orders $j \geq 4$ also capture nocturnal ozone features.

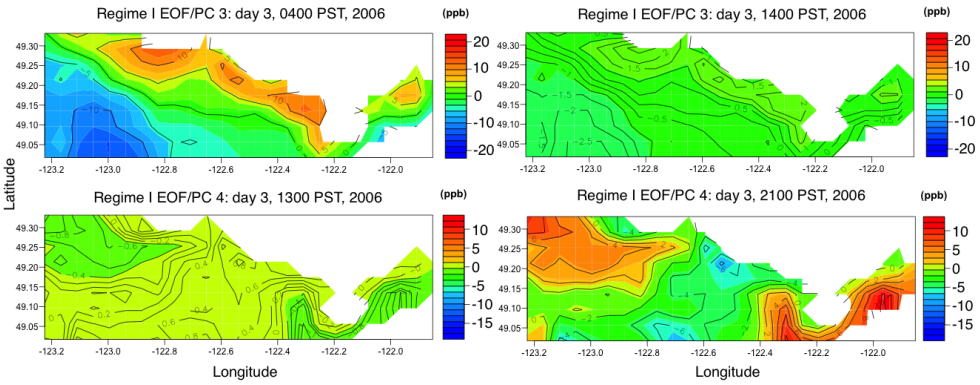


Figure 3.19: From the 2006 episode under type I regime: $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ and $\mathbf{P}_4\mathbf{E}_4^{\mathrm{T}}$ (units *ppb*) from selected hours.

Overall, the higher-order ozone features do not generate easily interpretable spatial or temporal patterns, although they certainly capture well-defined spatial structures, especially $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$. This lack of interpretability can

be attributed mainly to the fact that wind regimes I to III do not deviate significantly from a general westerly flow system. This wind system creates a relatively simple ozone processes where the ozone plume forms in the west and accumulates in the east due to eastward advection. As discussed earlier in this section, these features can be sufficiently captured by $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$ and $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$. Therefore, the higher-order features are left to recover location and time-specific ozone corrections or adjustments. For ozone fields dominated by the type IV regime, the leading 3 ozone features capture a slightly more complex ozone structure, hence ozone features of orders $j \geq 4$ are left to recover non-systematic data variations.

In summary, $j \geq 3$ or $j \geq 4$ ordered ozone features recover episode-specific and localized ozone variations in space and time. This assertion is also supported by the data reconstruction RMSEs showed in Section 3.3 (Table 3.2, Figures 3.5 and 3.6). These higher-order features can be understood as the representations of particular deviations from the general LFV ozone structure.

3.4.5 Sampling Stability of Ozone Features

As discussed by Richman (1986), the "correctness" of EOFs estimated from a space-time data can be affected by the quality of spatial sampling of that data. This is an important point to consider when one uses the EOF from a subsample of space-time data to estimate the unknown EOF of the whole domain. The $\mathbf{E}_j$ and associated $\mathbf{P}_j\mathbf{E}_j$ shown earlier in this chapter should not suffer from sampling errors. This is because the decomposed data are high-resolution CMAQ output in 4km-by-4km grids, and any defining features of CMAQ modelled ozone should be sufficiently captured.

However, when CMAQ modelled ozone features are evaluated against the observed features, one needs to consider the subdomain sampling stability of LFV ozone. As discussed in Chapter 2, the spatial domain of evaluated CMAQ outputs are matched to observations by spatially interpolating CMAQ outputs onto the locations of the n_{obs} irregularly placed monitoring sites. This means that I will be implementing PCA on a sub-sample of the

"full LFV" ozone fields.

Therefore, it is useful to analyze whether the sub-domain PCA gives the same ozone features we have seen in this chapter. One way of doing so is to (1) randomly sample n_{obs} number of locations from the space-time LFV data, (2) decompose the sample data into ozone features, and (3) repeat these two steps to obtain a sample of ozone features. To assess the sampling stability of PCA, one may compare the patterns of sample $\mathbf{P}_j$, $j = 1, \ldots, n_{obs}$, to the $\mathbf{P}_j$ decomposed from the full LFV data.

The preceding results have shown that $\mathbf{P}_j\mathbf{E}_j$ capture space-time features that represent the interactions between $\mathbf{E}_j$ and $\mathbf{P}_j$. If the type of temporal feature captured by subsample $\mathbf{P}_j$ remains stable, then the associated subspace $\mathbf{E}_j$ may be interpreted similarly to the j-th spatial feature of the full field. For instance, if sub-domain $\mathbf{P}_2$ captures day-night contrast as before (Section 3.4.3), then $\mathbf{E}_2$ still captures the path of diurnal ozone advection. As a result, the ozone feature interpretations formulated in the preceding sections are applicable to later CMAQ evaluations.

PCA sample plots shown in Figures 3.20, 3.21 and 3.22 are designed to analyze the above mentioned PCA sampling stability of LFV ozone. In each $\mathbf{P}_j$ sample plot, the *red-coloured* time series is the $\mathbf{P}_j$ of full LFV data during the middle 3 days of the 2006 CMAQ output. Here, the "full" LFV is the same triangular lower valley region we analyzed thus far in this chapter. Each *grey-coloured* curve is the $\mathbf{P}_j$ decomposed from one LFV subset with $n_{obs} = 17$ randomly sampled locations, and this subset covers the same time period as the full data. The $\mathbf{P}_j$'s of order $j = 1, 2, 3$ are each sampled 50 times to create Figures 3.20, 3.21 and 3.22.

As discussed in Section 3.2, the PCs in this thesis are calculated as $\mathbf{P}_{t\times n} = \mathbf{O}_{t\times n}\mathbf{E}_{n\times n}$: each $\mathbf{P}_j$ contains hourly weighted sums of $\mathbf{O}_{t\times n}$ where $\mathbf{E}_j$ are the spatial weights. This explains the difference in scale between $\mathbf{P}_1$'s from the full LFV data and its subsamples. As shown, the patterns of sampled $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ remained stable throughout the episode. The exceptions being $\mathbf{P}_2$ and $\mathbf{P}_3$ during a few nocturnal hours. The same PCA sampling analyses of different ozone episodes gave the same results. In the end, there is no reason to believe that PCA sampling error or feature

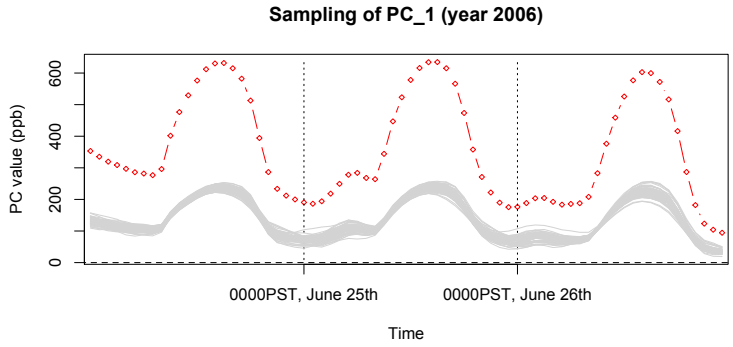


Figure 3.20: Sampling stability of PCA for $\mathbf{P}_1$. The full dataset is the 2006 CMAQ output within the LFV region with elevation$\leq$ 150 meters (the "full LFV" analyzed thus far in this chapter). The red curve is $\mathbf{P}_1$ of full LFV and each grey curve is $\mathbf{P}_1$ decomposed from one sub-data with $n = 17$ randomly sampled LFV locations.

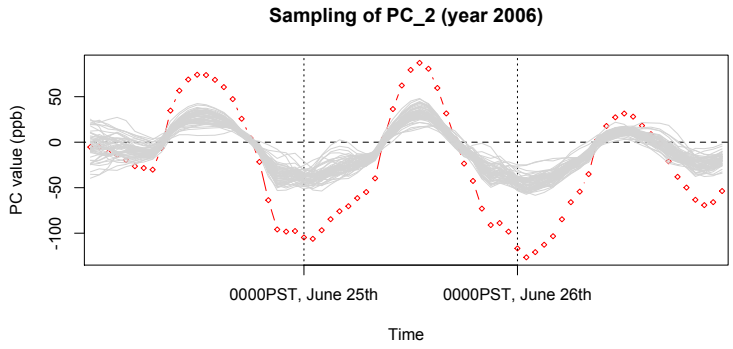


Figure 3.21: Sampling stability of PCA for $\mathbf{P}_2$. The full dataset is the 2006 CMAQ output within the LFV region with elevation$\leq$ 150 meters (the "full LFV" analyzed thus far in this chapter). The red curve is $\mathbf{P}_2$ of full LFV and each grey curve is $\mathbf{P}_2$ decomposed from one sub-data with $n = 17$ randomly sampled LFV locations.

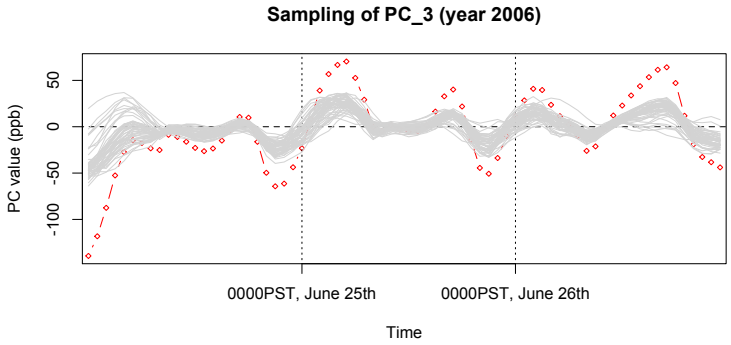


Figure 3.22: Sampling stability of PCA for $\mathbf{P}_3$. The full dataset is the 2006 CMAQ output within the LFV region with elevation$\leq$ 150 meters (the "full LFV" analyzed thus far in this chapter). The red curve is $\mathbf{P}_3$ of full LFV and each grey curve is $\mathbf{P}_3$ decomposed from one sub-data with $n = 17$ randomly sampled LFV locations.

instability is an overriding concern for LFV ozone field.

3.5 Chapter Conclusion

This chapter answered PCA-related topics that are crucial for feature-based AQM evaluation and ozone modelling in the following chapters. In chapter introduction, I raised the point of *feature correspondence* between AQM output and observations. This allows for direct comparison of ozone features and subsequent statistical modelling of feature differences.

First, it was determined that the ozone PCA in this thesis will be done on original ozone data without the use of column-centering/standardization (Section 3.2). The worry here is that if the 1st-order feature $\mathbf{E}_1$'s are incomparable, then the orthogonality of EOF will make all higher-order feature comparisons questionable. The analyses in this chapter have shown that $\mathbf{E}_1$ reliably captures the pattern of the spatial field of temporal ozone means, and this mean field can be calculated directly from the data and plotted. This physical reference can then be used to assess the "correctness" of the extracted $\mathbf{E}_1$ and determine whether the $\mathbf{E}_1$'s from AQM and observations

represent comparable features. This decision is further supported by the results where the PCA of anomaly data did not reveal additional features of LFV regional ozone (Appendix B.2).

Another means of ensuring defensible and informative feature comparison is to formulate an understanding of ozone features being evaluated. Here are the key features of the LFV ozone modelled by CMAQ:

- All episode have the same general structure of space-time ozone means, and this feature is reliably captured by $\mathbf{P}_1\mathbf{E}_1^{\mathrm{T}}$. This result reveals the highly consistent nature of LFV ozone process during an episode. Over the period of 1985 to 2006, the changes in emission standard and differences in weather condition did not significantly alter the space-time distribution of LFV ozone.

- For episodes driven by wind regime types I, II and III, the 2nd-order feature $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ consistently captured the same form dynamic east-west ozone contrast. This dynamic contrast revealed the most dominant pattern of ozone advection across LFV: it is the eastward horizontal transport of ozone plume, which is driven by a westerly wind system. Wind regime types I to III are defined mostly by a westerly wind system with little deviations in direction.

  The ozone feature under type IV wind regime is different in that $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ and $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ jointly capture a more complex advection pattern, where ozone plume moves from the northwest to the southwest during the afternoon, and then advects eastward by the end of a diurnal cycle. This advection feature is only present from days dominated by the type IV wind regime.

- Higher order ozone features are less interpretable, and they primarily recover localized space-time ozone variations rather than systematic space-time features. Here, the "higher order" means $j \geq 3$ for episodes under regimes types I, II and III, and $j \geq 4$ for episode under regime type IV.

These results also fill an existing knowledge gap regarding the regional-scale features of LFV ozone.

Furthermore, existing works on ozone PCA (Section 1.4) rely on visual analyses of $\mathbf{E}_j$ to interpret possible modes of variation. This is useful when the pollution field covers a large spatial domain where one can identify the underlying climate systems, e.g., North Atlantic Oscillation. However, such large scale process is not readily identifiable when analyzing a smaller, but complex regional pollution field such as the LFV ozone. Instead of analyzing $\mathbf{E}_j$ on its own, it was found that the analysis of space-time interaction feature $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ provides a more informative and big picture understanding of the underlying dynamic process. In summary, a $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ at $j \geq 2$ is a dynamic ozone contrast that captures certain spade-time advection process. Three pieces of information can be interpreted from $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ plots: (1) the direction of ozone advection, (2) the time period (within an episode) during which the advection took place, and (3) the order j, i.e., the importance of this advection to the overall ozone process.

Study in this chapter also showed that one may analyze a space-time ozone field using a few leading ozone features; a high-dimensional data can be analyzed through simpler data components $\mathbf{E}_j$ and $\mathbf{P}_j$. As discussed, these few leading features capture the systematic structure and behaviour of LFV ozone. Therefore, through feature-based AQM evaluation, one may draw up a "big picture" of how AQM modelled ozone differ from the real-world physical field. As mentioned, the actual evaluation will be implemented in Chapters 5 and 6. In the next chapter, I will proposed a framework for modelling individual ozone features, as well as the complete space-time ozone process defined by these features.

Chapter 4

A Statistical Model of Space-Time Ozone Features

In the previous chapter, I presented the PCA methods used to extract ozone features $\mathbf{E}_j$'s and $\mathbf{P}_j$'s and discussed topics related to ozone feature analysis. In this chapter, I propose methods of modelling individual ozone features as random processes driven by variables capturing atmospheric conditions, ozone precursor emission rates and antecedent concentrations. I will also analyze a method of modelling a complete space-time ozone process through its features. The statistical ozone feature models are intended to apply to both AQM ozone and physical observations. The purpose of this chapter is to estimate the details of ozone feature models and assess their modelling capability through goodness-of-fit analyses and exercises in space-time ozone forecasting.

I am placing significant effort on identifying the Gaussian Process covariates that can be used to model the process behind each ozone feature. This model developing exercise is driven by my prior experience that spatial and temporal variables such as longitude, latitude and "hour of the day" are not sufficient in modelling ozone fields. The complexities of the AQM modelling system (Chapters 1 and 2) also means that it does not have a clearly definable input structure typical of a computer model. Hence, careful analysis is need to formulate a set of AQM input conditions useful for statistical modelling.

All modelling analyses in this chapter will be done using CMAQ-WRF-SMOKE outputs (Section 2.1, Table 2.1). Due to the richness of information the computer models can provide, their outputs are used instead of obser-

vation data to estimate the ozone feature models. The computer models produce data for a comprehensive list of variables representing regional meteorology and particulate pollution (more details in Section 4.3), whereas observations provide data on a few basic weather and air pollution measurements. Moreover, computer model outputs are produced on a dense and regular spatial grid covering a large geographic domain and they, unlike the observations, do not suffer from missing or erroneous data.

The statistical ozone feature model developed in this chapter will then be applied in Chapters 5 and 6 to implement two distinct types of feature-based AQM evaluation. As discussed in the introductory chapter, for one evaluation, I will model the difference in ozone features between AQM and observations as functions of AQM input conditions. In another evaluation, I will estimate statistical ozone feature models for both AQM ozone and physical observation, then compare the statistical properties of two ozone processes under the same condition.

Besides serving as a means to my main objective of AQM evaluation, the methodologies developed in this chapter are a novel and efficient approach for modelling of space-time air pollution processes (not limited to ozone). The bulk of existing statistical air pollution models deal only with either a spatial process or a time series (temporal process). In a spatial model, the random process is the spatial air pollution field whose values are usually a temporal summary, i.e., summer time means (Fuentes and Raftery, 2005; Liu, 2007). In a temporal model, the random process is expressed by a time series of air pollution whose values are either based on a point location, or a spatial average (Gao et al., 1996). Spatial-temporal pollution modelling has received more attention in recent years, much of it is based on Gaussian Processes and Kriging (Berrocal et al., 2009; Conti and O'Hagan, 2010; Zidek et al., 2012). Usually, Gaussian Process based models are designed to handle data in their original form - what I call the "raw data".

Before preceding, a reminder on the terminology. "Spatial ozone means" describe the ozone averaged across space, and "temporal means" describe averages across time. Hence, a "spatial field of temporal ozone means", or "mean field" and "field of means" for short, describes a spatial field of mean

values obtained by averaging ozone across time. Similarly, "time series of spatial means" or "hourly LFV mean ozone" describe a times series obtained by averaging ozone across space for each hour.

Similarly, "spatial ozone standard deviation" and "temporal ozone standard deviation" are summarized respectively, across space and time. The same way of describing spatial and temporal summaries are also used for other variables like temperature, NOx and VOC emission rates, etc.

Section 4.1 presents the general model formulations for ozone features $\mathbf{E}_j$'s and $\mathbf{P}_j$'s. Section 4.2, summarizes the kriging-based methods for predicting unobserved spatial/temporal processes and approach to estimate GP model parameters. Section 4.3 introduces the covariates used to model spatial-temporal ozone features and methods of covariate selection. Section 4.4 outlines the framework of implementing *feature-based ozone modelling* using methods described in Sections 4.1 to 4.3. Data analyses are on Sections 4.4-4.7, where the ozone feature models will be estimated and their modelling capability will be assessed. The data used in this chapter are 2006 CMAQ output over the entire "rectangular" LFV domain (more details in Section 4.4).

4.1 Ozone Features and Gaussian Process Models

Be it a CMAQ output or a physical measurement, ozone values are influenced by a number of background meteorological processes, chemical emissions and reactions. As discussed in Section 1.4, without physically measuring or making an estimate using CMAQ, the ozone value given a set of background conditions is unknown, hence it is reasonable to treat this unknown ozone value as a random variable. Since the EOFs and PCs are extracted from data treated as random ozone values, both the EOFs and the PCs are by logic, random vectors (multivariate data).

Denote the extracted EOFs as $\mathbf{E}_j$, $j = 1, \ldots, p$, where p is the number of EOFs used for modelling. Similarly, $\mathbf{P}_j$ $(j = 1, \ldots, p)$ denotes the PC vectors. As before, I define the dimension of space-time ozone data as $t \times n$, t and n are respectively the number of time points and locations. As such,

each EOF vector is multivariate random spatial data of size n, and each PC vector is multivariate random temporal data of length t.

I showed in the last chapter that $\mathbf{E}_j$'s and $\mathbf{P}_j$'s capture strong spatial and temporal structures, this implies the presence of spatial and temporal correlations, i.e, the spatial and temporal features are not dominated by white noise. Hence the ozone feature models should be derived from multivariate distributions that explicitly account for internal correlations. In this study, Gaussian Process models are used. As its name suggests, this model is built on the idea that a vector of random variables can be modelled by a Multivariate Normal (MVN) distribution (Sacks et al., 1989; Kennedy and O'Hagan, 2001). In the introductory chapter, I discussed its well-documented proficiencies in modelling both computer model outputs and physical observations, and the initial reasoning behind my application of GP models for this research. The theoretical and practical appropriateness of GP models will be examined during data analyses.

4.1.1 Gaussian Process Model for an EOF

Let the $n \times k$ matrix $\mathbf{X}_{\mathbf{E}\_j}$ denotes the model covariate set for the $n \times 1$ response vector $\mathbf{E}_j$, where k is the number of covariates. These covariates will be selected in Section 4.5.

I propose a GP-based model for $\mathbf{E}_j$:

$$\mathbf{E}_j|\mathbf{X}_{\mathbf{E}\_j} = \mathbf{F}(\mathbf{X}_{\mathbf{E}\_j})\boldsymbol{\beta}_{\mathbf{E}\_j} + \mathbf{Z}_{\mathbf{E}\_j}. \tag{4.1}$$

The regression function $\mathbf{F}(\mathbf{X}_{\mathbf{E}\_j})\boldsymbol{\beta}_{\mathbf{E}\_j}$ is the mean of the GP model, where $\boldsymbol{\beta}_{\mathbf{E}\_j}$ is the regression coefficient vector. I define the regression design matrix as $\mathbf{F}(\mathbf{X}_{\mathbf{E}\_j}) = \{\mathbf{f}(\mathbf{x}_{j,1}), \ldots, \mathbf{f}(\mathbf{x}_{j,n})\}^{\mathrm{T}}$, where $\mathbf{f}(\mathbf{x}_{j,1}), \ldots, \mathbf{f}(\mathbf{x}_{j,n})$ are column vectors of functions of the covariates, and $\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}}$ are rows of $\mathbf{X}_{\mathbf{E}\_j}$. $\mathbf{Z}_{\mathbf{E}\_j}$

is a zero-mean Gaussian Process:

$$\begin{aligned}
\mathbf{Z}_{\mathbf{E}_j} &\sim \text{MVN}(\mathbf{0},\ \sigma^2_{\mathbf{E}_j}\mathbf{R}_{\mathbf{E}_j}),\ \text{where} \\
\mathbf{R}_{\mathbf{E}_j} &= \begin{bmatrix}
1 & R(E_{j,1},E_{j,2}) & R(E_{j,1},E_{j,3}) & \dots & R(E_{j,1},E_{j,n}) \\
R(E_{j,2},E_{j,1}) & 1 & R(E_{j,2},E_{j,3}) & \dots & R(E_{j,2},E_{j,n}) \\
R(E_{j,3},E_{j,1}) & R(E_{j,3},E_{j,3}) & 1 & \dots & R(E_{j,3},E_{j,n}) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
R(E_{j,n},E_{j,1}) & R(E_{j,n},E_{j,2}) & R(E_{j,n},E_{j,3}) & \dots & 1
\end{bmatrix}.
\end{aligned}$$

The correlations between $\mathbf{E}_j$ elements E_j and E'_j are quantified by a function of covariate distances: $R(E_j, E'_j) = f_R(\mathbf{x}_j - \mathbf{x}'_j)$, where $\mathbf{x}_j$ and $\mathbf{x}'_j$ are the respective covariate sets for E_j and E'_j. Multiplied by the constant variance $\sigma^2_{\mathbf{E}\_j}$, one obtains the model covariance.

The mean represents the *fixed* component in (4.1), and models the simple association between $\mathbf{E}_j$ and $\mathbf{X}_{\mathbf{E}\_j}$ as a regression function. The *random* component $\mathbf{Z}_{\mathbf{E}\_j}$ models the stochastic behaviour of $\mathbf{E}_j$ based on $\mathbf{E}_j$'s correlation structure (a function of $\mathbf{X}_{\mathbf{E}\_j}$). In my experience, it is often reasonable to simply specify the fixed regression term to be a constant scalar mean (though I do introduce more general regression terms during model implementation later in the chapter). The true focus of GP model lie with the modelling of the stochastic process $\mathbf{Z}_{\mathbf{E}\_j}$.

The distribution of $\mathbf{E}_j$ is given by:

$$\mathbf{E}_j|\mathbf{X}_{\mathbf{E}\_j} = \begin{bmatrix} E_{j,1}|\mathbf{x}_{j,1} \\ \vdots \\ E_{j,n}|\mathbf{x}_{j,n} \end{bmatrix} \sim \text{MVN}\left\{ \begin{bmatrix} f^{\mathrm{T}}(\mathbf{x}_{j,1})\boldsymbol{\beta}_{\mathbf{E}\_j} \\ \vdots \\ f^{\mathrm{T}}(\mathbf{x}_{j,n})\boldsymbol{\beta}_{\mathbf{E}\_j} \end{bmatrix}, \sigma^2_{\mathbf{E}\_j}\mathbf{R}_{\mathbf{E}\_j} \right\}.$$

$E_{j,1}, \dots, E_{j,n}$ are the n elements of $\mathbf{E}_j$, $f^{\mathrm{T}}(\mathbf{x}_{j,1})\boldsymbol{\beta}_{\mathbf{E}\_j}, \dots, f^{\mathrm{T}}(\mathbf{x}_{j,n})\boldsymbol{\beta}_{\mathbf{E}\_j}$ and $\mathbf{x}_{j,1}, \dots, \mathbf{x}_{j,n}$ are means and covariate sets of each random variable in $\mathbf{E}_j$.

4.1.2 Gaussian Process Model for a PC

If $X_{\mathbf{P}\_j}$ is the covariate matrix of $\mathbf{P}_j$, a Principal Component vector is modelled similarly by

$$\mathbf{P}_j|\mathbf{X}_{\mathbf{P}\_j} = \mathbf{F}(\mathbf{X}_{\mathbf{P}\_j})\boldsymbol{\beta}_{\mathbf{P}\_j} + \mathbf{Z}_{\mathbf{P}\_j}, \tag{4.2}$$

where $\mathbf{Z}_{\mathbf{P}\_j}$ is a zero-mean Gaussian Process:

$$\begin{aligned}
\mathbf{Z}_{\mathbf{P}_j} &\sim \text{MVN}(\mathbf{0},\ \sigma^2_{\mathbf{P}_j}\mathbf{R}_{\mathbf{P}_j}),\ \text{where} \\
\mathbf{R}_{\mathbf{P}_j} &= \begin{bmatrix}
1 & R(P_{j,1},P_{j,2}) & R(P_{j,1},P_{j,3}) & \dots & R(P_{j,1},P_{j,n}) \\
R(P_{j,2},P_{j,1}) & 1 & R(P_{j,2},P_{j,3}) & \dots & R(P_{j,2},P_{j,n}) \\
R(P_{j,3},P_{j,1}) & R(P_{j,3},P_{j,2}) & 1 & \dots & R(P_{j,3},P_{j,n}) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
R(P_{j,n},P_{j,1}) & R(P_{j,n},P_{j,2}) & R(P_{j,n},P_{j,3}) & \dots & 1
\end{bmatrix}.
\end{aligned}$$

4.1.3 Modelling a Complete Space-time Ozone Process

To model the ozone process in its original space-time field and value scale (*ppb*), I apply the constructive relationship between space-time ozone and its defining ozone features:

$$\mathbf{O}_{t\times n}|\mathbf{X}_{\mathbf{E}},\mathbf{X}_{\mathbf{P}} \approx \sum_{j=1}^{p}(\mathbf{P}_j|\mathbf{X}_{\mathbf{P}\_j})(\mathbf{E}_j|\mathbf{X}_{\mathbf{E}\_j})^{\mathrm{T}},\ p \ll \min(t,n). \tag{4.3}$$

In essence, I propose to model the spatial and temporal ozone features as GPs, then the complete space-time ozone process as the joint sum of outer products of its defining features. This is a departure from the traditional statistical practice of directly modelling given data. I name this approach the *feature-based ozone model* to emphasize this central idea.

4.1.4 Use of GPs for Modelling Ozone Features

I showed in Chapter 3 that the structures of individual spatial and temporal ozone features are highly non-linear, implying that the usual linear

regression models are not sufficient for modelling these features. Bloomfield et al. (1996) and Thompson et al. (2001) provided good examples of the deficiencies of linear regression when modelling non-linear air pollution processes.

Given the GP assumption of individual ozone features, the resultant space-time ozone process has a complicated distribution. It is common to transform space-time data, e.g., a square-root transformation, before applying the GP model (Le and Zidek, 2006). So it is reasonable to place the assumption of non-normality on a complex process such as space-time ozone. Examples also exist where the GP assumption is imposed on the original data, e.g., Gao et al. (1996) and Fuentes and Raftery (2005) (in the context of SO_2 modelling). However, the research in this thesis is focused on the analysis and statistical modelling of ozone features rather than the original ozone field. Given this particular focus, it is sensible to place any statistical assumption on the ozone features, i.e., the random process being analyzed directly.

Furthermore, by modelling each $\mathbf{E}_j$ (and each $\mathbf{P}_j$) as a GP, I am assuming that a particular ozone feature of a particular episode is treated as one realization of GP. The purpose of GP model is to predict the outcome of this particular realization for unobserved values of the covariates. This analysis will be done in Chapters 5 and 6 to evaluate CMAQ.

In practice, the use of a GP is mainly for convenience. It is also a decision informed by both my past experience and existing literature (Section 1.4), where GP-based models were shown to be proficient in emulation complex non-linear processes. Analysis later in this chapter will further validate the appropriateness, thereby the usefulness of GPs for modelling the ozone features.

Lastly, I should note that the general concept of "modelling the data components as GPs" is a recent one. Higdon et al. (2008) built a statistical emulator of a computerized "implosion simulator". This is a high-dimensional output computer model in that it produces multiple outputs for every model run. The authors applied SVD to decompose a data matrix of model outputs into loading components (in PCA terminology) that

are regarded as invariant basis vectors. The model output data is a matrix of multiple model runs, where each row contains the multi-dimensional output from one model setting. The high-dimension model output is then statistically modelled as a scaled-sum of said invariant basis vectors, and each "scale" is modelled as a GP. Kleiber et al. (2014) implemented this approach to emulate the output of a "geomagnetic storm simulator". The "scales" in Higdon et al. (2008) are expanded into vectors of Principal Components and modelled as GPs, the loading components are still regarded as invariant. These works are both motivated by the objective of "computer model calibration" while data-decomposition is applied to reduce the dimensionality of the model output in order to increase the efficiency of their calibration algorithms, which are both based on Bayesian methodologies.

One might view the feature-based ozone modelling framework as an elaboration of the above mentioned modelling ideas, but driven by a different focus and intended application. My focus is to develop individual ozone feature models as means to feature-based CMAQ evaluation. Whereas the common focus of Higdon et al. (2008) and Kleiber et al. (2014) is to reduce computation load during computer model calibration, attention was not placed in detailed modelling of data components. The computer models in Higdon et al. (2008) and Kleiber et al. (2014) are simple enough to have a definitive set of model covariates, which is not the case in my air pollution modelling. Furthermore, there is no invariant basis in my modelling framework; all data features $\mathbf{E}_j$'s and $\mathbf{P}_j$'s are modelled, whereas the aforementioned references both regarded their equivalent of the $\mathbf{E}$'s as vectors of an invariant basis.

4.2 Background on Gaussian Process Models

In this section, I first present a non-parametric (without assumptions about a specific probability distribution) prediction method and discuss its connection with Gaussian Processes. I will then present the general method of estimating Gaussian Process models. For ease of discussion, I use the generic notation of Y as a random response and X as process covariates.

4.2.1 Best Linear Unbiased Predictor (BLUP)

Let $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be realizations of a random vector $\mathbf{Y}_{n\times 1}$. Each y_i, $i = 1, \ldots, n$, is realized given a set of k covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^{\mathrm{T}}$. In more general notation, the GP model for $\mathbf{Y}(\mathbf{x})$ has the formulation

$$\mathbf{Y}(\mathbf{x}) = \mathbf{F}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{x}). \tag{4.4}$$

$\mathbf{F} = (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_n))^{\mathrm{T}}$ is a row matrix of n covariate functions, where the i-th covariate function is $\mathbf{f}(\mathbf{x}_i) = (\mathbf{f}_1(\mathbf{x}_i), \ldots, \mathbf{f}_k(\mathbf{x}_i))^{\mathrm{T}}$, making $\mathbf{F}$ an $n \times k$ design matrix. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ is a regression coefficient vector. $Z(\mathbf{x})$ is a zero mean random process with covariance matrix $\sigma^2\mathbf{R}$. The elements of $\mathbf{R}$ quantify the correlation between random variables in $\mathbf{Z}(\mathbf{x})$, and subsequently $\mathbf{Y}(\mathbf{x})$ ($\mathbf{Z}(\mathbf{x})$ is the random component of $\mathbf{Y}(\mathbf{x})$). Also as mentioned, correlation between any ij-th pair $Y_i(\mathbf{x}_i)$ and $Y_j(\mathbf{x}_j)$ is a function of some distance measure between relevant covariate pairs $(\mathbf{x}_i, \mathbf{x}_j)$.

One way to predict the response at an "unobserved" covariate setting $\mathbf{x}_0$ is called *universal kriging* (Matheron, 1963; Cressie, 1990), named after the South African mining engineer Danie G. Krige. Sacks et al. (1989) adapted the same mathematics to problems in computer experiments with higher-dimensional covariates. Given n "observed covariate inputs" $\mathbf{X}_{\mathrm{S}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, we have corresponding output/data $\mathbf{y}_{\mathrm{S}} = (y_1(\mathbf{x}_1), \ldots, y_n(\mathbf{x}_n))^{\mathrm{T}}$, and the universal kriging predictor has the form $\hat{y}(\mathbf{x}_0) = \mathbf{w}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{y}_{\mathrm{S}}$ with $\mathbf{w}(\mathbf{x}_0)$ being an $n \times 1$ vector. It is essentially a weighted average of the data. From the frequentist perspective, $\mathbf{y}_{\mathrm{S}}$ and $y(\mathbf{x}_0)$ are realizations of equation (4.4). The Best Linear Unbiased Predictor (BLUP) is the $\mathbf{w}(\mathbf{x}_0)$ that minimizes Mean Squared Error (MSE)

$$\mathrm{MSE}\left[\hat{y}(\mathbf{x}_0)\right] = \mathrm{E}\left[\mathbf{w}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{Y}_{\mathrm{S}} - Y(\mathbf{x}_0)\right]^2 \tag{4.5}$$

subject to the unbiasedness constraint $\mathrm{E}\left[\mathbf{w}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{Y}_{\mathrm{S}}\right] = \mathrm{E}\left[Y(\mathbf{x}_0)\right]$, i.e., $\mathbf{w}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{F}\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{f}(\mathbf{x}_0)$ for all $\boldsymbol{\beta}$.

When one takes the usual Bayesian approach, $\hat{y}(\mathbf{x}_0)$ is the posterior mean $\mathrm{E}\left[Y(\mathbf{x}_0)|\mathbf{y}_{\mathrm{S}}\right]$. Currin et al. (1991) suggested modelling $\mathbf{Z}(\mathbf{x})$ as a Gaussian

Bayesian prior on the unknown function and Handcock and Stein (1993) applied Bayesian methodologies in estimating model parameters.

In addition to $\mathbf{R}$, a matrix of correlations between Y's at "observed" settings $\mathbf{x}$, we have the correlation between Y at an "unobserved" $\mathbf{x}_0$ and the Y's at $\mathbf{x}$ (Sacks et al., 1989). This correlation is represented by the $n \times 1$ vector $\mathbf{r}(\mathbf{x}_0) = [R(\mathbf{x}_1, \mathbf{x}_0), ..., R(\mathbf{x}_n, \mathbf{x}_0)]^{\mathrm{T}}$.

Without accounting for the aforementioned unbiasedness constraint, the MSE in equation (4.5) can be expanded to arrive at the expression:

$$\begin{aligned} \mathrm{MSE}\left(\hat{y}(\mathbf{x}_0)\right) = [(\mathbf{w}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{F} - \mathbf{f}^{\mathrm{T}}(\mathbf{x}_0))\boldsymbol{\beta}]^2 + \\ \sigma^2[1 + \mathbf{w}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{R}\mathbf{w}(\mathbf{x}_0) - 2\mathbf{w}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{r}(\mathbf{x}_0)], \end{aligned} \tag{4.6}$$

where the first term in (4.6) is the squared prediction bias. Incorporating the unbiasedness constraint $(\mathbf{w}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{F} - \mathbf{f}^{\mathrm{T}}(\mathbf{x}_0))\boldsymbol{\beta} = 0$, one is left to minimize the second term in (4.6). This term is generally referred to as the "unbiased MSE equation". To obtain the $\mathbf{w}(\mathbf{x}_0)$ that minimizes the term $1 + \mathbf{w}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{R}\mathbf{w}(\mathbf{x}) - 2\mathbf{w}^{\mathrm{T}}(\mathbf{x})\mathbf{r}(\mathbf{x}_0)$ under said constraint, one would add the $k \times 1$ Lagrangian term $\boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{x})[\mathbf{F}^{\mathrm{T}}\mathbf{w}(\mathbf{x}_0) - \mathbf{f}(\mathbf{x}_0)]$ to the unbiased MSE equation, differentiate and find $\mathbf{w}(\mathbf{x}_0)$ defining the BLUP from the equation

$$\begin{pmatrix} 0 & \mathbf{F}^{\mathrm{T}} \\ \mathbf{F} & \mathbf{R} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\mathbf{x}) \\ \mathbf{w}(\mathbf{x}_0) \end{pmatrix} = \begin{pmatrix} \mathbf{f}(\mathbf{x}_0) \\ \mathbf{r}(\mathbf{x}_0) \end{pmatrix}.$$

Making use of the following result regarding the inversion of a partitioned matrix:

$$\begin{pmatrix} 0 & \mathbf{F}^{\mathrm{T}} \\ \mathbf{F} & \mathbf{R} \end{pmatrix} = \begin{pmatrix} \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{F}\mathbf{K}^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1} & \mathbf{R}^{-1}\mathbf{F}\mathbf{K}^{-1} \\ \mathbf{K}^{-1}\mathbf{F}\mathbf{R}^{-1} & -\mathbf{K}^{-1} \end{pmatrix},$$

where $\mathbf{K} = \mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{F}$, one may solve for $\mathbf{w}(\mathbf{x}_0)$. This results in the predictor

$$\hat{y}(\mathbf{x}_0) = \mathbf{f}^{\mathrm{T}}(\mathbf{x}_0)\hat{\boldsymbol{\beta}} + \mathbf{r}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{R}^{-1}(\mathbf{y}_{\mathrm{S}} - \mathbf{F}\hat{\boldsymbol{\beta}}), \tag{4.7}$$

with $\hat{\boldsymbol{\beta}}$ being the generalized least-square estimate.

Substituting the optimized $\mathbf{w}(\mathbf{x}_0)$ into the unbiased MSE equation, one obtains the prediction standard error

$$\begin{aligned}\mathrm{SE}\left(\hat{y}(\mathbf{x}_0)\right) = \sigma[1 - \mathbf{r}^{\mathrm{T}}(\mathbf{x}_0)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}_0) + \\ (\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}_0))^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}_0))]^{\frac{1}{2}}\end{aligned} \tag{4.8}$$

The preceding derivations are based on a single-point prediction at $\mathbf{x}_0$. For a multivariate prediction (simultaneous predictions of multiple responses), the prediction equation for individual response is still (4.7). However, due to the correlation between unobserved responses, we now have a prediction covariance matrix in place of a single prediction standard error (or variance) (Bastos and O'Hagan, 2009). Let $\mathbf{Y}(\mathbf{X}_0)$ be a vector of m "unobserved" responses, where $\mathbf{X}_0$ is the $m \times k$ matrix of the covariates. Further denote $\mathbf{F}_0$ as the corresponding $m \times k$ design matrix of $\mathbf{X}_0$. The prediction covariance matrix is

$$\begin{aligned}\boldsymbol{\Sigma}(\hat{\mathbf{y}}(\mathbf{X}_0)) = \sigma^2[\mathbf{R}_m - \mathbf{r}_m^{\mathrm{T}}(\mathbf{x}_0)\mathbf{R}^{-1}\mathbf{r}_m(\mathbf{x}_0) + \\ (\mathbf{F}_0 - \mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}_m(\mathbf{x}_0))^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{F}_0 - \mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}_m(\mathbf{x}_0))],\end{aligned} \tag{4.9}$$

where $\boldsymbol{\Sigma}(\hat{\mathbf{y}}(\mathbf{X}_0))$ has dimension $m \times m$. $\mathbf{R}_m$ is the $m \times m$ correlation matrix of $\mathbf{Y}(\mathbf{X}_0)$. The correlation matrix $\mathbf{r}_m(\mathbf{x}_0)$ has dimension $n \times m$, each of its columns is a n-length vector of correlations between an element in $\mathbf{Y}(\mathbf{X}_0)$ and $\mathbf{Y}_{\mathrm{S}}$. As one can see, multivariate-prediction does not change the components involving training data, it simply increases the dimensions of terms that are functions of $\mathbf{X}_0$. Moreover, the diagonal elements of $\mathbf{R}_m$ are 1, hence the individual prediction standard errors are still calculated as (4.8).

4.2.2 Connecting the BLUP to a Gaussian Distribution

Suppose the random response $\mathbf{Y}(\mathbf{x})$ follows a multivariate normal (Gaussian) distribution, and let $\mathbf{Y}_{\mathrm{J}} = (\mathbf{Y}_{\mathrm{S}}, Y(\mathbf{x}_0))^{\mathrm{T}}$ denote the $n \times 1$ vector of the training responses and the response to be predicted. Then $\mathbf{Y}_{\mathrm{J}}$ is also MVN

with density:

$$f_{\mathbf{Y}_\mathrm{J}}(\mathbf{y}_\mathrm{J}|\boldsymbol{\beta},\sigma,\mathbf{R}_{n+1}) = \frac{1}{(2\pi\sigma^2)^{n/2}\det(\mathbf{R}_{n+1})}\times \exp\left(-\frac{(\mathbf{y}_\mathrm{J}-\mathbf{F}_\mathrm{J}\boldsymbol{\beta})^\mathrm{T}\mathbf{R}_{n+1}^{-1}(\mathbf{y}_\mathrm{J}-\mathbf{F}_\mathrm{J}\boldsymbol{\beta})}{2\sigma^2}\right),$$

$$\text{where } \mathbf{F}_\mathrm{J} = (\mathbf{F}^\mathrm{T}, \mathbf{f}(\mathbf{x}_0))^\mathrm{T},$$

$$\text{and } \mathbf{R}_{n+1} = \begin{pmatrix} \mathbf{R} & \mathbf{r}(\mathbf{x}_0) \\ \mathbf{r}^\mathrm{T}(\mathbf{x}_0) & 1 \end{pmatrix}.$$

The multivariate distribution of $\mathbf{Y}_\mathrm{S}$ can be written out in the same fashion.

The conditional distribution of $Y(\mathbf{x}_0)$ given $\mathbf{Y}_\mathrm{S}$ is

$$f_{Y(\mathbf{x}_0)|\mathbf{Y}_\mathrm{S}}(y(\mathbf{x}_0)|\mathbf{y}_\mathrm{S}) = \frac{f_{\mathbf{Y}_\mathrm{J}}(\mathbf{y}_\mathrm{J}|\boldsymbol{\beta},\sigma,\mathbf{R}_{n+1})}{f_{\mathbf{Y}_\mathrm{S}}(\mathbf{y}_\mathrm{S}|\boldsymbol{\beta},\sigma,\mathbf{R})}. \tag{4.10}$$

It can be shown through rather tedious matrix algebra, that the conditional distribution (4.10) follows a normal (Gaussian) distribution with mean and variance:

$$\begin{aligned} \mathrm{E}(Y(\mathbf{x}_0)|\mathbf{Y}_\mathrm{S}) &= \mathbf{f}^\mathrm{T}(\mathbf{x}_0)\boldsymbol{\beta} + \mathbf{r}^\mathrm{T}(\mathbf{x}_0)\mathbf{R}^{-1}(y_\mathrm{S}-\mathbf{F}\boldsymbol{\beta}) \\ \mathrm{Var}(Y(\mathbf{x}_0)|\mathbf{Y}_\mathrm{S}) &= \sigma^2[1-\mathbf{r}^\mathrm{T}(\mathbf{x}_0)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}_0)]. \end{aligned}$$

One may notice that the conditional mean has exactly the same expression as the BLUP (4.7), while the conditional variance is the squared BLUP standard error (4.8) without the 3rd term inside the bracket. This is because the above expressions for $Y(\mathbf{x}_0)|\mathbf{Y}_\mathrm{S}$ is derived under the assumption that $\boldsymbol{\beta}$ is known, whereas the BLUP is derived using the Generalized Least Square (GLS) estimator $\hat{\boldsymbol{\beta}}$. The 3rd bracketed term in (4.8) simply represent the extra error resulting from not knowing $\boldsymbol{\beta}$.

4.2.3 Estimating GP Parameters: Fitting the GP Models

A decision needs to be made regarding the form of the correlation function $R(\mathbf{x},\mathbf{x}')$. The requirement here is that the resultant covariance matrix

$\sigma^2 R(\mathbf{x}, \mathbf{x}')$ is positive-definite (Nychka et al., 2002; Fuentes and Raftery, 2005). One possible specification of $R(\mathbf{x}, \mathbf{x}')$ is the power-exponential correlation function, which has form

$$R(\mathbf{x}, \mathbf{x}') = \exp(-\sum_{j=1}^{k} \theta_j |x_j - x'_j|^{\alpha_j}), \quad \theta_j > 0 \text{ and } 1 \leq \alpha_j \leq 2, \tag{4.11}$$

where x_j and x'_j are one of k covariates of $\mathbf{x} = (x_1, .., x_k)^{\mathrm{T}}$ and $\mathbf{x}' = (x'_1, ..., x'_k)^{\mathrm{T}}$ respectively. Such a correlation function is utilized repeatedly in the literature: Sacks et al. (1989), Gao et al. (1996) and Kennedy and O'Hagan (2001) for example.

This formulation is attractive in its ease of interpretation. A small normalized distance between $\mathbf{x}$ and $\mathbf{x}'$ gives high correlation that tends to 1 as the distance between $\mathbf{x}$ and $\mathbf{x}'$ goes to 0, and conversely, the correlation approaches 0 when the distance becomes large. A small value of θ_j implies that Y as a function of x_j is relatively insensitive to the fluctuation of x_j. In other words, for any fixed distance $x_j - x'_j$, covariate x_j's numerical influence in (4.11) tends to 0 when θ_j is small, where a zero-value inside the exponential function gives a perfect correlation of 1 between $Y(\mathbf{x})$ and $Y(\mathbf{x}')$ for dimension j. Thus θ_j can be viewed as a measure of correlation strength or "activity" of associated covariate x_j. The parameter α_j controls the smoothness of the correlation function, where $\alpha_j = 2$ gives a smooth surface (infinitely differentiable). Furthermore, such a correlation function makes the GP model scale-invariant to the covariate inputs. Notice in (4.11), when a covariate x_j is scaled as $\tilde{s}x_j$ ($\tilde{s}$ being the scaling factor), the resultant correlation parameters are simply scaled up to α_j-th power $\theta_j/\tilde{s}^{\alpha_j}$ to keep $R(\mathbf{x}, \mathbf{x}')$ constant. Discussions of the interpretation of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ can be found in Welch et al. (1992) and Jones et al. (1998).

Assuming $Y(\mathbf{x})$ follows a Gaussian process (GP), one may write the

likelihood based on $\boldsymbol{\beta}$, σ^2 and the correlation parameters in $R(\mathbf{x}, \mathbf{x}')$ as:

$$L(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \boldsymbol{\alpha} | \mathbf{y}_{\mathrm{S}}, \mathbf{F}) = \frac{1}{(2\pi\sigma^2)^{n/2}|\mathbf{R}|^{1/2}} \exp\left[-\frac{(\mathbf{y}_{\mathrm{S}} - \mathbf{F}\boldsymbol{\beta})^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y}_{\mathrm{S}} - \mathbf{F}\boldsymbol{\beta})}{2\sigma^2}\right]. \tag{4.12}$$

Given the correlation parameters, the Maximum Likelihood Estimator (MLE) of variance σ^2 has expression

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y}_{\mathrm{S}} - \mathbf{F}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y}_{\mathrm{S}} - \mathbf{F}\hat{\boldsymbol{\beta}}),$$

and the Generalized Least Squared estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{F})^{-1} \mathbf{F}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{y}_{\mathrm{S}}.$$

Placing these two expressions back into the likelihood function, one is left with a *profile likelihood* that is a function of the correlation parameters (Sacks et al., 1989; Jones et al., 1998). This profile likelihood is then maximized over the correlation parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$ and $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$.

In this thesis, all GP model parameter estimates are MLEs. I will use the common statistical expression "fit the model" to describe the procedure of estimating GP model parameters given a dataset.

Strictly speaking, the BLUP results in Section 4.2.1 assume that the covariance parameters $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are known. In practice however, they are usually estimated by the methods described here, or by fitting a certain variogram model, which is often used in geo-statistics instead of the correlation function. The variogram is

$$\begin{aligned} 2\gamma(\mathbf{x}, \mathbf{x}') &= \mathrm{Var}\left[Y(\mathbf{x}) - Y(\mathbf{x}')\right] \\ &= \sigma^2 R(\mathbf{x}, \mathbf{x}) + \sigma^2 R(\mathbf{x}', \mathbf{x}') - 2\sigma^2 R(\mathbf{x}, \mathbf{x}') = 2\sigma^2[1 - R(\mathbf{x}, \mathbf{x}')] \end{aligned}$$

and $\gamma(\mathbf{x}, \mathbf{x}')$ is called the semi-variogram. Le and Zidek (2006) described popular variogram models, such as the exponential and spherical variograms. Cressie (1990) also derived the predictive function based on the semi-variogram

under the weighting condition $\sum w(\mathbf{x}) = 1$ (i.e., Ordinary Kriging).

4.3 Variable and Covariate Selection

In this thesis, the term *model variable* refers to: temperature, wind speed, planetary boundary layer height, NOx and VOC emission rate and ambient concentration. The term *model covariate* refers a function of a variable used in the statistical model. For instance, "temperature" is a variable, while the form of temperature values inputted into the model, such as 24-hour mean temperature, is a covariate. The term *design matrix* denotes a matrix of model covariates.

This section discusses one of the most important aspects of model development: selecting model covariates. Naturally, an ozone model with a well-chosen set of covariates should properly model the spatial-temporal features of an ozone process. This desired attribute is expressed numerically through a combination of high likelihood value and low prediction error. Likelihood based statistical analyses are theoretical assessments of model quality, while prediction-error based analyses are practical measurements of how well a model performs.

Proper variable and covariate selection also make the GP models more parsimonious, i.e., devoid of unnecessary model covariates. This is particularly useful for my application. When modelling physical observations, the accompanying variable measurements are not always available. Hence from a practical standpoint, a model containing fewer variables/covariates is easier to implement.

From the following discussion, readers will notice that my variable and covariate selection process is a balancing act between statistical analysis and scientific reasoning.

4.3.1 Model Variables

In this subsection, I discuss and analyze the usefulness of various variables for the ozone model. Below is a list of scientifically relevant variables:

- **Location and time variables:** I use longitude, latitude, elevation and hour of the day. As an alternative to longitude and latitude, one may also use unitless values that index point locations on a 2-D $x - y$ Cartesian grid. Considering the aforementioned scale-invariant property of the GP model, the scale of a location variable is arbitrary. I chose longitude, latitude and elevation to allow identification of real-life locations in modelling, an important feature for the upcoming CMAQ model evaluation against observations.

- **Meteorological variables:** These are wind speed and direction, temperature, pressure, humidity and boundary-layer height. Boundary-layer height is the depth of the atmospheric layer in which the surface-bound photochemical processes are contained (Stull, 1988). Taylor (1991) and Salmond and McKendry (2002) are examples of works discussing the associations between LFV's boundary layer height and its ozone process.

- **Chemical precursor information:** Important precursors to ozone are NOx (oxides of Nitrogen) and VOC (volatile organic compounds). I use two types of precursor variables: the incoming rate of new precursor molecules (rate of emissions) in units of *mole per second*, and concentrations of precursors (in *ppb*) already present in the atmosphere, which is referred to here as *antecedent precursors*. As discussed in Section 2.1, NOx is the sum of NO and NO_2 data, while VOC data are created by adding the scaled values of 16 families of volatile organic compounds.

As mentioned in Chapter 2, the meteorological variables are outputs from WRF (with MCIP post-processing), and precursor emission rates are outputs from SMOKE. These variables are created over the same spatial domain and time period as CMAQ. The antecedent precursor concentrations are processed from CMAQ outputs; the processing methods are discussed in a later subsection titled "Antecedent Precursor Concentrations". As for GP models of observed ozone features, the model variable data are the cor-

responding (in space and time) observations. Related details of observation-based modelling are presented in Chapter 6, the discussion here is focused on the statistical modelling of the CMAQ ozone features.

Selecting Meteorological Variables

On the matter of variable selection (not covariate selection), the decisions are based on available scientific advice and references. In previous chapters, I discussed the defining space-time behaviours of a LFV ozone process and its relationship with the dominant wind regime. The following paragraphs further complete the picture on the meteorological conditions that are conducive to an ozone episode, they provide the reasoning and justification behind my science-based variable selection approach.

The start of an ozone episode requires meteorological conditions such as high temperatures, clear sky and low wind speed (Robeson and Steyn, 1990; Taylor, 1991; Ainslie and Steyn, 2007). All these conditions can be triggered by a meso-scale high pressure system. High pressure near ground creates a low pressure system in the atmospheric layer above, the cold (thus dense) air from above moves downward and warms due to adiabatic heating. This results in a clear sky that allows unobstructed UV radiation, directly driving the photochemical process. A high pressure system also results in low wind speed: fast enough for localized chemical mixing, but not fast enough to transport the ongoing photochemical process out of the LFV quickly. In short, pressure is negatively correlated to wind speed and positively correlated to temperature (Taylor, 1992; Ainslie and Steyn, 2007).

However, an ozone model based solely on pressure without temperature or wind speed is hard to justify. Although high pressure is indirectly essential to the formation of an ozone episode in the most generalized way, temperature and wind speed are the meteorological forces that drive the more detailed photochemical and atmospheric transportation process; temperature and wind are simply more relevant and useful for the geographical scale of my ozone modelling (Beaver et al., 2010; Jin et al., 2011; Reuten et al., 2012). Furthermore, my statistical analysis has shown that the addi-

tion of pressure along with temperature and wind as model variables tends to degrade the goodness-of-fit and forecasting capability of my feature-based ozone model.[6] This is perhaps partially the result of covariate confounding introduced by the addition of pressure. Therefore, pressure is not used for my ozone modelling.

The intensity of ultra-violet radiation, measured by UV index, is another important variable. As discussed, an ozone episode takes place during days with clear sky, thus unobstructed UV radiation. This variable is not used in modelling for two reasons. First, it is collinear with temperature: during summer days, intense UV radiation results in higher regional temperature. Secondly, the UV radiation is near uniformly distributed in space (at least within LFV). Thus, the practical utility of including UV index in statistical ozone modelling is questionable.

In the end, I decided to incorporate 3 meteorological variables: temperature (in units *Kelvin*), wind speed (in *meter/second*) and boundary layer height (in *meter*). Boundary layer height is particularly important given the topography of LFV. The mountains surrounding the LFV act as a physical "barrier" to the horizontal advection of pollutants that channels them along the valley (Steyn et al., 1997), and a shallow boundary layer would trap the pollutants within the LFV's barrier (Robeson and Steyn, 1990; Taylor, 1991). An ozone episode may be initiated by this accumulation of air pollutants in conjunction with the meteorology conducive to photochemical reactions: high temperature, light wind and strong UV (Boubel et al., 1994, Chapter 17).

The use of wind direction data and generating data for antecedent precursor concentrations

Although wind direction is an integral part of CMAQ that dictates the direction of ozone plume transport (as I have shown in chapter 3), its usefulness in ozone feature modelling is questionable. This is because its values are

[6] The types of "statistical analyses" mentioned here are presented from Sections 4.5 to 4.7.

measured as the angle that orients clockwise from the north. Given an angle, the direction of the vector (from the origin) is the direction in which the wind blows *from*, e.g., a wind direction value of 270° represents a westerly wind (blowing directly from the west). As I will discuss in more detail, the covariate data, which are space-time in nature, will be numerically summarized before being incorporated into my model. Sets of wind direction data with vastly different temporal or spatial profiles may end up being similar in value when summarized. This lack of identifiability would make the modelling effect of wind direction difficult to interpret, thus wind direction is not incorporated directly as a model variable. However, the influence of wind direction (hence wind regime) on spatial-temporal ozone patterns are instead expressed through the creation of a new variable: *antecedent chemical precursor concentrations*.

For grid cell s at hour h, CMAQ produces concentrations for O_3, NOx, and VOC (among others). CMAQ integrates a system of partial-differential equations given the initial and boundary conditions for the grid cell. For each s, concentration at h is integrated over the time period between $h-1$ and h, where h in indexed in "hours". So output at hour 1300 is an averaged concentration during $1201-1300$. Within this hour, CMAQ produces air pollution estimates in smaller time-intervals. I denote this small-interval time variable as τ, where $\Delta\tau \ll \Delta h$ and Δ represents incremental time scale of τ and h. Therefore, from an input-output perspective of a computer model, precursor concentrations from $\tau - \Delta\tau$ are the "inputs" to ozone output at τ: they are corresponding antecedent precursor information. As is the case with ozone output, hourly CMAQ outputs for NOx and VOC's are the $\Delta\tau$-interval concentrations integrated over an hour.

NOx and VOC concentrations at h are used as antecedent precursor data associated with O_3 at h. This is because atmospheric photochemical reactions happen at a rate more appropriately indexed by τ, and lagged concentrations at $h-1$ (previous hour) are too far back in time considering both the reaction rates and mesoscale wind speed. This point can be illustrated through simple mathematics: under a light wind of 3 ms^{-1}, which is typical during an ozone episode within LFV (Section 2.2), the per-hour-distance of

ozone-plume advection is $3\,ms^{-1} \cdot 3600\,s = 10800\,m = 10.8\,km$. This is much larger than the CMAQ grid-cell size of 4km×4km.

Furthermore, due to the presence of atmospheric circulation, each grid cell s has two sources of antecedent precursor chemicals: its own grid cell *and* all neighbouring grid cells (Kalenderski and Steyn, 2011). In a regular grid system, each s has 8 immediate neighbours.

At time h, the neighbouring precursor concentrations are part of the initial/boundary conditions for s, and the wind direction of grid cell s at time h determines how neighbouring cell concentrations affect its ozone production process. I weight lagged neighbouring precursor concentrations using a scheme called Arcsin Weighting. The entire range of wind direction ($0^\circ - 360^\circ$) is partitioned into four intervals: $> 315^\circ$ and $\leq 45^\circ$, between 45° and 135°, between 135° and 225°, and between 225° and 315°. The wind direction value of s at time h will fall into one of the intervals. Figure 4.1 illustrates how the arcsin weighting is done when the wind direction is within the range 225° and 315°. Within any interval, only three neighbours are used for weighting: the lagged concentration of "directly upwind neighbour" (`U` in Figure 4.1) is multiplied by $1/(1 + 2/\sqrt{2})$, while the two neighbours adjacent to the upwind one (`A` in Figure 4.1) are each multiplied by $(1/\sqrt{2})/(1 + 2/\sqrt{2})$. The antecedent concentrations at time h for each s is the sum of the concentration from s and arcsin weighted concentrations from its neighbours.

In total, there are 4 types of chemical precursor variables: the NOx and VOC emission rates, and antecedent (or lagged) NOx and VOC concentrations.

4.3.2 Selection of Model Covariates

Remember that a space-time ozone field is decomposed into spatial and temporal components $\mathbf{E}_j$'s and $\mathbf{P}_j$'s, $j = 1, \ldots, p$. Since the meteorological and chemical variables are also space-time in nature, a function is needed to transform model variables into numerical expressions appropriate for modelling ozone features. Several designs of transform function are explored and

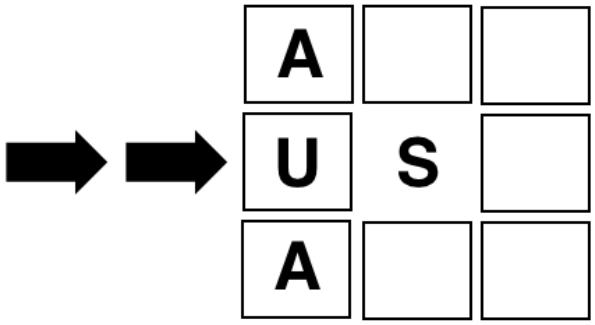


Figure 4.1: Illustration of how the arcsin weighting is done when the wind direction is 270°. There are 8 neighbours to the grid cell s, the NOx and VOC concentrations from the upwind neighbour "U" and two adjacent neighbours "A" will be used. The precursor concentration in "U" is scaled by $1/(1 + 2/\sqrt{2})$ and the concentrations in "A" is each scaled by $(1/\sqrt{2})/(1 + 2/\sqrt{2})$. Their weight sum is then calculated.

the two best functions (in terms of modelling capability) are presented in this section.

Model I: Covariates are Spatial or Temporal Means of Variables

With random response variables representing either spatial or temporal ozone patterns, their corresponding covariates can be spatial and temporal means of the model variables. A $t \times n$ space-time ozone dataset has meteorological and chemical data from the same spatial-temporal domain. An n-length vector of temporal variable means is obtained by averaging $t \times n$ variable data by column (across time): such a vector represents the mean field of each variable. A t-length vector of spatial variable means forms by averaging the variable data by row (across space): this is the hourly time series of LFV variable means.

The n-length vectors of temporal variable means (variable mean fields) are model covariates of $\mathbf{E}_j$'s, i.e., spatial ozone features. The t-length vector of LFV variable means are the covariates of $\mathbf{P}_j$'s, i.e., temporal ozone features. Covariate selection is not necessary with such a formulation: within any ozone feature model, each variable is represented by one covariate only.

Moreover, all spatial feature models $\mathbf{E}_j$ have the same set of covariates, and likewise for the temporal feature models $\mathbf{P}_j$.

Model II: Covariates are EOFs or PCs of Variables

As with the decomposition of ozone data, I can use the PCA of variable data to extract spatial and temporal features of model variables. EOFs of model variables are the covariates for spatial ozone feature models, while PCs of model variables are the covariates for temporal ozone feature models.

To better understand how variable EOFs and PCs can be incorporated into GP models, one needs to first analyze the PCA results of model variables. The results from PCA of model variable data are presented in Appendix C.1. The data are CMAQ-WRF-SMOKE outputs from 2006, the spatial domain is the "rectangular" LFV (Section 2.1, Figure 2.1) and the data contain all 96 hours of output. For all model variables, the first 3 EOF-PC pairs capture over 90%, or in certain cases, close to 100% of data variation, hence the full model covariate set is comprised of the first 3 EOFs/PCs extracted from all 7 variable data. The purpose of model variable PCA is not to closely analyze the decomposition of model variables. The objective is to confirm that the model variable EOFs and PCs do indeed capture noticeable space-time structures and variations, interpretability is not a priority.

The next step of analysis is covariate selection: determine the number of useful EOFs and PCs from each variable. Intuitively, one may view covariate selection as a procedure in which I determine the "useful number" of data features from each variable for the modelling of ozone features. After extensive analysis and experimentation with a variety of covariate selection schemes, I decided on a forward selection method which I refer to as *Iterative Improvement*. Relative to other selection methods I explored, iterative improvement delivered the best combination of model goodness-of-fit and ozone feature forecasting capability. This iterative approach is based on the method of maximum likelihood (4.12). Take the GP model of any ozone EOF as an example, the iterative procedure proceeds as follow:

Step 1: The starting covariate set contains longitude, latitude, elevation and

$\mathbf{E}_1$'s of all 7 meteorological and precursor variables. As a reminder, these are temperature, wind speed, boundary layer height, NOx and VOC emission rates and antecedent concentrations.

Fit the GP model containing the starting covariate set and record the maximized likelihood. Denote this starting likelihood value as ℓ_{start}. The likelihood maximizing method (thus model estimate method) is described in Section 4.2 and the detailed implementation will be described during data analysis in Sections 4.5 to 4.6.

All 1st order model-variable features are used since they capture/recover the most fundamental features of each variable. The iterative improvement procedure is in essence, a means of incorporating useful additional covariate information into each GP model.

Step 2: The candidate set contains the remaining Q covariates. Initially, $Q = 14$ ($\mathbf{E}_2$'s and $\mathbf{E}_3$'s from 7 meteorological and precursor variables). Fit Q GP models, where model q, $q = 1, \ldots, Q$, is the GP model with the q-th covariate added to the starting covariate set in the previous step. Denote each maximized likelihood as ℓ_q, the largest one as ℓ_{max}, and its corresponding covariate as x_{max}.

Step 3: If the likelihood test statistic $\delta_{max} = 2(\ell_{max} - \ell_{start})$ is *larger* than the critical value $\chi^2_{0.95,\, df=2}$, then I say that the addition of covariate x_{max} is a significant improvement over the starting model. Update the starting covariate set by incorporating x_{max}, and update the candidate set by removing x_{max}. Note that the χ^2 degree of freedom is 2 because there are correlation and power parameters attached to each additional covariate in the GP model for an ozone EOF.

Step 4: Repeat step 1 to 3 until the likelihood test statistic in step 3 is smaller than the critical value. This is the point where the addition of more candidate covariates fails to improve the GP model fit in a statistically significant way. If the initial step 3 fails to add any covariates, then I conclude that the original starting set is the best.

The alternative to iterative improvement is its backward selection counterpart: start with a full covariate set with all possible model covariates and iteratively remove covariates one-by-one until further omission of covariates results in a statistically significant drop in the log-likelihood.

The iterative covariate selection for ozone PC models proceed in the same manner using the PCs of model variables as candidate covariates.

The iterative improvement algorithm is built upon rejecting the null hypothesis (type I error), whereas the backward selection is based on *not* rejecting the null (type II error). Intuitively, the iterative improvement algorithm should arrive at the optimized covariate set quickly if the set is small. Conversely, the backward selection approach is recommended if the final covariate set is expected to be large: it requires less iterations to delete a small number of candidate covariates than to add a large number of them.

The likelihood test is especially useful since I have multivariate (correlated) data, where many statistical tests are inapplicable. A likelihood-based test takes into account the inherent correlation structure within the data. Furthermore, it can be shown that maximizing the likelihood is the same as "minimizing the expected predictive deficiency" (Currin et al., 1991). Therefore, one may interpret my proposed covariate selection algorithm as a statistical procedure that searches for a model that is expected to deliver the best prediction.

4.3.3 Goodness-of-fit Statistics

Summary statistics based on cross-validation errors (or residuals) are helpful for assessing model quality from a practical, goodness-of-prediction perspective. As usual, let $\mathbf{Y}_{n\times 1}$ be a vector of n responses and $\mathbf{X}_{n\times k}$ its corresponding covariate matrix. Denote an entry of the response-covariate set as y_i and $\mathbf{x}_i$, $i = 1, \ldots, n$. Let $\mathbf{Y}_{-i}$ and $\mathbf{X}_{-i}$ be the response and covariate data without the i-th entry. In cross validation, I fit GP model using $\mathbf{Y}_{-i}$ and $\mathbf{X}_{-i}$, and predict $y_i|\mathbf{x}_i$. Repeating for all entries in the data, one obtains n cross-validation predictions $\hat{y}_i|\{\mathbf{x}_i, \mathbf{Y}_{-i}, \mathbf{X}_{-i}\}$, $i = 1, \ldots, n$, each with accompanying prediction error. Cross-validation root mean squared error, or

CVRMSE, is calculated as

$$\text{CVRMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i|\{\mathbf{x}_i, \mathbf{Y}_{-i}, \mathbf{X}_{-i}\})^2}{n}}.$$

I also use mean percentage error (MPE), which during cross-validation is calculated as

$$\text{MPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i|\{\mathbf{x}_i, \mathbf{Y}_{-i}, \mathbf{X}_{-i}\}}{y_i} * 100\%\right).$$

In the MPE calculation, the positive and negative errors can offset each other, and it can be applied as a measure of prediction bias. When the CV-residuals are taken as absolute values, MPE becomes Mean Absolute Percentage Errors (MAPE).

Furthermore, a prediction $\hat{y}_i|\{\mathbf{x}_i, \mathbf{Y}_{-i}, \mathbf{X}_{-i}\}$ and its prediction error are estimates of the conditional mean and variance of a normal distribution $y_i|Y_{-i}$. In theory (Gao et al., 1996; Bastos and O'Hagan, 2009), I can check the assumptions of a Gaussian Process model by analyzing plots of standardized cross validation residual $(y_i - \hat{y}_i)/\text{se}(\hat{y}_i)$ (to be implemented in Section 4.6).

4.4 The Framework of Feature-based Ozone Modelling

In the remainder of this chapter, I will estimate individual spatial and temporal ozone feature models and assess their goodness-of-fit and predictive capabilities. It is worth re-emphasizing that the ozone models developed in this chapter are intended to both serve as the basis for air quality model evaluation and as an efficient means of modelling a large space-time air pollution process.

The framework below details the steps for (1) estimating individual spatial and temporal ozone feature models, and (2) applying said models to forecast LFV ozone features and subsequent space-time ozone fields.

- **Step 1:** Partition a complete space-time ozone dataset into the *training set* and the *predictive set*. The data can be separated by spatial locations, time periods, or both. In my analyses, I partition ozone data by the time periods within an episode.

  The training dataset is used to fit the ozone feature models and carry out model diagnostics. The predictive set contains variable data *future* to the training set. It provides statistical model inputs to make forecasts of ozone features and space-time ozone fields. This helps us to further assess the capabilities of ozone feature models.

  In this thesis, the term "prediction" is used to describe the procedure of estimating unobserved values of an ozone process, whereas the term "forecast" is specific to prediction of future ozone process in relation to the training set.

- **Step 2:** Model selection. In Section 4.3, I proposed a type of ozone feature model where the covariates are selected from the leading 3 EOFs or PCs of model variables. The covariate selection is done using an iterative improvement procedure. This model selection method will be implemented using the ozone feature data and the model covariate data obtained from the PCA of the training dataset.

- **Step 3:** With model covariates for each GP determined, fit individual GP models (estimate model parameters) for $\mathbf{E}_j$ and $\mathbf{P}_j$, $j = 1, 2, 3$, using their respective training data. Forecasts of regional ozone features can then be made from the fitted models using the EOFs and PCs of model variables from the predictive dataset.

  The predictive data contain information on the LFV's atmospheric and pollution conditions during the time period following the training set. Hence, I am making *ozone-feature forecasts* for the LFV region. Kriging-based prediction methods were discussed in Section 4.2.

- **Step 4:** Combine EOF and PC forecasts via (4.13) below to forecast the hourly LFV ozone fields. Denote the predictive set covariate data by $\mathbf{x}_{\mathbf{E}\_j}$ and $\mathbf{x}_{\mathbf{P}\_j}$, $j = 1, \ldots, p$, and the estimated model parameters

by $\hat{\boldsymbol{\xi}}_{E,j} = \{\hat{\boldsymbol{\theta}}_{E,j}, \hat{\boldsymbol{\alpha}}_{E,j}, \hat{\boldsymbol{\beta}}_{E,j}, \hat{\boldsymbol{\sigma}}_{E,j}\}$ and $\hat{\boldsymbol{\xi}}_{P,j} = \{\hat{\boldsymbol{\theta}}_{P,j}, \hat{\boldsymbol{\alpha}}_{P,j}, \hat{\boldsymbol{\beta}}_{P,j}, \hat{\boldsymbol{\sigma}}_{P,j}\}$ (parameter notations from Section 4.2). The predictive equation for a space-time ozone field is the "prediction based" form of (4.3):

$$\hat{\mathbf{O}}_{pred}|\mathbf{x}_{pred} = \sum_{j=1}^{p} (\hat{\mathbf{P}}_{j\_pred}|\mathbf{x}_{\mathbf{P}\_j}, \hat{\boldsymbol{\xi}}_{P,j})(\hat{\mathbf{E}}_{j\_pred}|\mathbf{x}_{\mathbf{E}\_j}, \hat{\boldsymbol{\xi}}_{E,j})^{\mathrm{T}}. \quad (4.13)$$

The data $\mathbf{x}_{pred}$ are the variable data of a predictive set: the data from which $\mathbf{x_E}$ and $\mathbf{x_P}$ are extracted. $\mathbf{O}_{pred}$ is the prediction target, i.e, the space-time ozone field of the predictive set. Furthermore, the training set and predictive set can have different data dimensions (different numbers of locations and time points).

In its entirety, my proposed ozone modelling scheme follows the path: decompose ozone training data into separate ozone features ⇒ select model covariates and fit the ozone feature models ⇒ apply the covariate inputs from the predictive data to forecast ozone features ⇒ combine predicted features to forecast the complete space-time ozone field.

4.4.1 Details of Training Set and Predictive Set

The data analysis in this chapter uses the 2006 CMAQ output across the entire "rectangular" LFV domain shown in Figure 4.2 (originally from Figure 2.1 in Chapter 2). This LFV domain includes north shore mountains as well as the valley floor analyzed in the last Chapter. In total, this region contains $n = 229$ CMAQ grid cells. The training data is the first 2 full days of 2006 episode (June 24th and 25th), and the prediction is made for the remaining full day (June 26th), so $t_{train} = 48$ and $t_{pred} = 24$. The training ozone field is dominated by a type I wind regime and the predictive day is driven by a type III regime (Table 3.1, Ainslie and Steyn (2007)).

As such, I am using ozone feature models estimated for an ozone field under one wind regime type to forecast an ozone field under a different regime. In addition, the modelling is implemented on a complex pollution field that includes a large metropolitan area, farm lands (Abbotsford), and surrounding mountains with elevations sometimes exceeding 500m. A

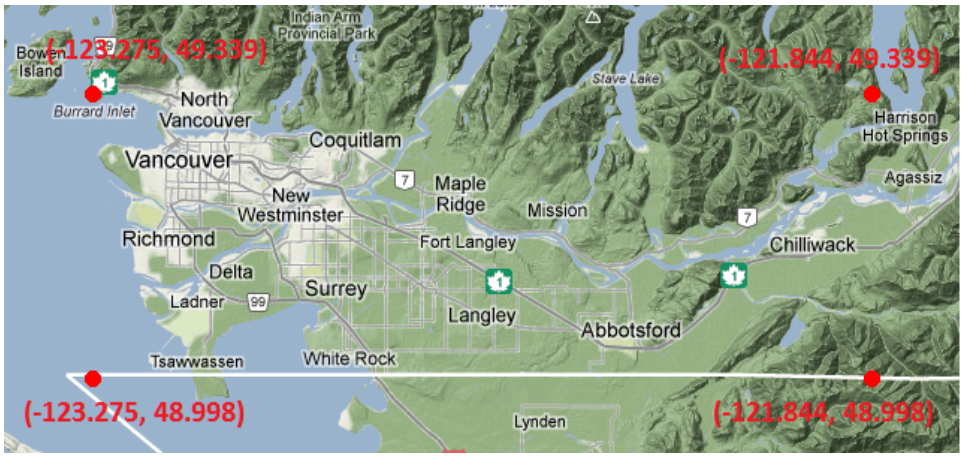


Figure 4.2: Map of the complete "rectangular" LFV domain being modelled in this Chapter. The red dots indicate the corners of the modelled domain, the coordinate of each corner is also shown.

simpler modelling exercise can be based on the low-elevation LFV field analyzed in Chapter 3, and have both training and predictive set under the same wind regime type, e.g., 1995 or 2001 episodes. Therefore, the analyses in this chapter is a rather tough assessment of the modelling capability of the proposed ozone feature models.

One also needs to produce the antecedent NOx and VOC data for the predictive set. As discussed in Section 4.3, the antecedent precursor data are processed from CMAQ outputs. The implication here is that one cannot obtain the actual antecedent precursor data without running CMAQ, or in the case of predicting physical observations, making measurements during the prediction hours. In practice, it is redundant to run CMAQ or take measurements to obtain NOx and VOC data before predicting ozone fields, because the real ozone values are also obtained. One straightforward solution is to use the previous-day antecedent precursor data as a proxy.

However, this is not an issue for the analyses in this thesis. Here, the predictions/forecasts are done for the sole purpose of evaluating the capability of ozone feature models. In fact, to properly evaluate these statistical models, it is essential to apply the *actual* NOx and VOC antecedent con-

centrations from the day of prediction. Hence, all forecasts in this chapter are done using NOx and VOC data from June 26th, 2006; the day for which the ozone forecasts are made.

4.4.2 PCA of CMAQ Output over the "Rectangular" LFV

Figures 4.3 and 4.4 show that the type of spatial-temporal ozone features over the "rectangular LFV" domain are more or less consistent with what were learned from Chapter 3, where the analyses are done for the low-elevation "valley floor" of LFV. However, the 2nd EOF now captures the dynamic spatial contrast between the lower valley and the mountains, whereas in Chapter 3 for the valley floor, the spatial contrast was between eastern and western LFV. The leading temporal features are still interpreted the same: $\mathbf{P}_1$ is the hourly spatial mean, and $\mathbf{P}_2$ is a daytime-nighttime temporal contrast that interacts with $\mathbf{E}_2$ to capture a diurnal eastward ozone transport.

Figure 4.5 shows the eigenspectra from the ozone PCA of training dataset. The ozone features form degenerate multiplets starting at $j = 4$. Based on the eigenspectrum interpretations discussed in Sections 3.2 and 3.3, one may conclude that for the "rectangular" LFV ozone field, the leading 3 ozone features are identifiable as individual features and separable from the others.

4.5 Covariate Selection

This section presents the results from covariate selection using the training data. A covariate is denoted using the abbreviation of the variable name, with subscripts indicating the order of EOF or PC. For example, $\text{Temp}_{E,2}$ and $\text{NOx-lag}_{P,1}$ denote respectively, the 2nd EOF of temperature and the 1st PC of antecedent (lagged) NOx concentration.

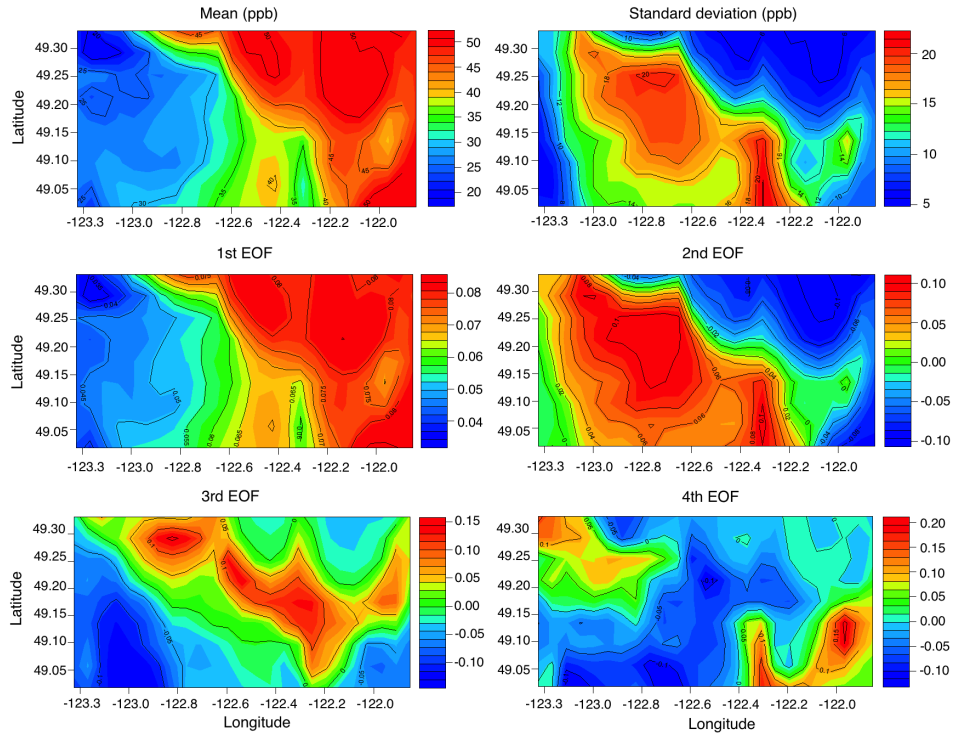


Figure 4.3: From the training set, day 2 and 3 of the 2006 Ozone episode: Spatial plots of temporal ozone means and standard deviations (calculated across time), and the first 4 EOFs. The ozone mean and standard deviation have units *ppb*, while $\mathbf{E}_1, \ldots, \mathbf{E}_4$ are unitless.

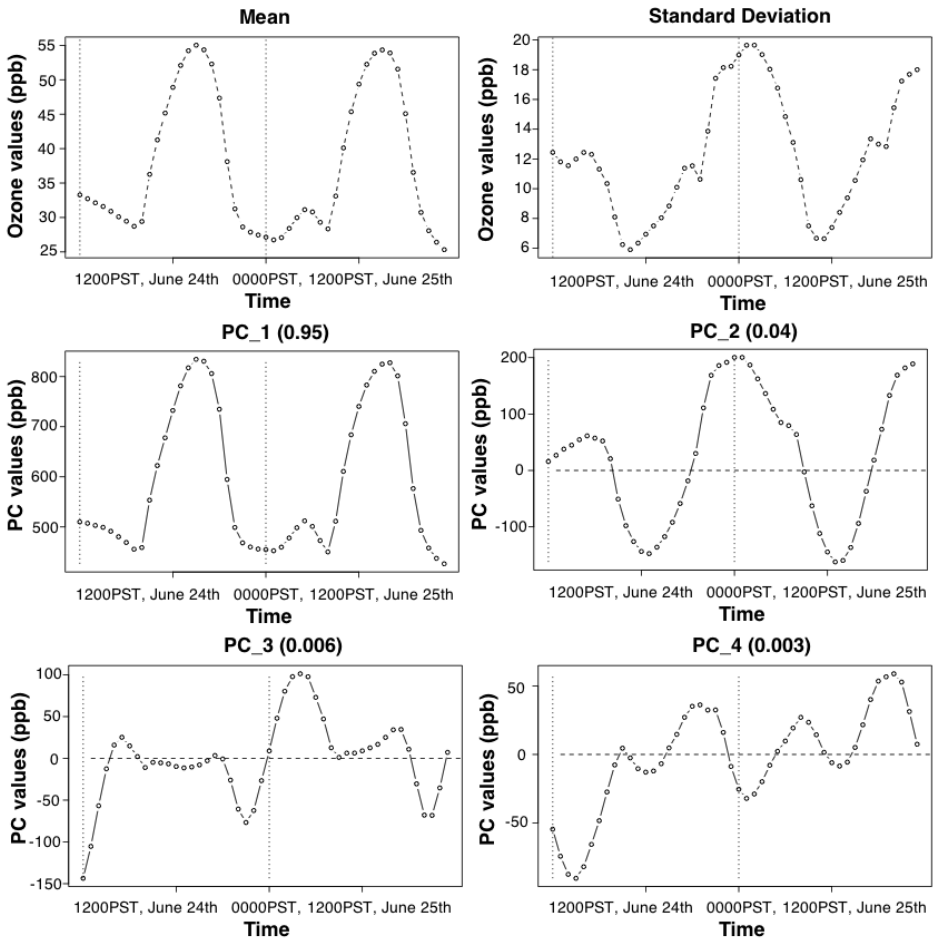


Figure 4.4: From the training set, day 2 and 3 of the 2006 Ozone episode: Time series of hourly spatial ozone means, standard deviation and the first four PCs. The number in each PC plot heading is the "proportion of data variation explained". All plotted data have units *ppb*.

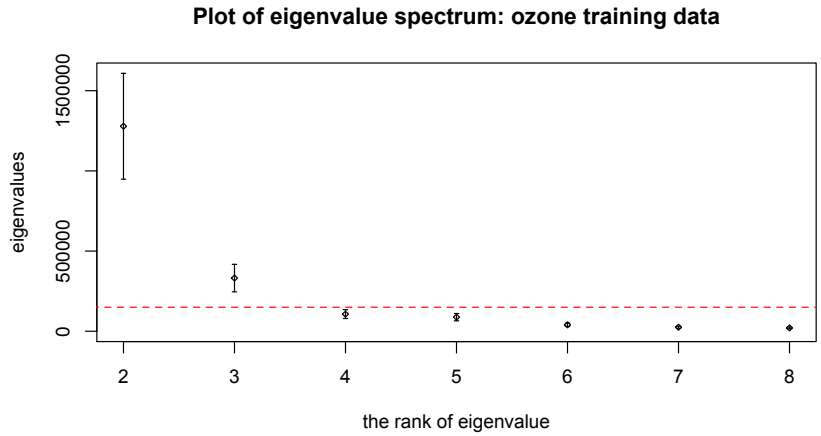


Figure 4.5: Eigenspectrum from the PCA of the training data: 2006 CMAQ output for June 24th and 25th, over the rectangular LFV domain (area including the mountains). The dashed line indicates the ozone feature when feature degeneracy occurs. No spectrum is shown for λ_1, which has a much larger eigenvalue that is distinct from those shown.

4.5.1 Implementation Details

As discussed in Section 4.2, the multivariate normal based profile likelihood (4.12) is maximized to find estimates for the GP model parameters, and the GP correlation functions are power exponential. I use the program `GASP` written by William J. Welch to optimize all my Gaussian-based profile likelihoods. This program iterates through different non-linear optimization approaches like the Nelder-Mead method to maximize a likelihood function. It is a reliable GP optimizer that has been applied in many published works (Jones et al., 1998; Aslett et al., 1998).

`GASP` outputs a number of optimization summaries, one of which is called the *Condition Number*. It shows the number of significant figures lost to numerical error in the maximum likelihood results. In other words, it indicates the precision and reliability of optimization: the *higher* the Condition Number, the *smaller* the number of accurate significant digits, which in term indicates a *lessened* degree of optimization quality. As such, it is desir-

able to have a small condition number. In practice, a condition number of $> 1 \times 10^6$ gives cause for concern. All the optimizations presented in this thesis have smaller condition numbers.

The regression part of $\mathbf{E}_j$ GP models (4.1) contains longitude, latitude, elevation and an intercept term: $\mathbf{f}(\mathbf{X_E}) = (\mathbf{1}, \text{lon}, \text{lat}, \text{elev})^{\text{T}}$. The covariates in the regression term are fixed. I only select covariates for the stochastic process $\mathbf{Z_E}$, i.e., covariates that define spatial correlations within each $\mathbf{E}_j$. As mentioned in Section 4.1, the regression term is often treated as a constant. Based on my past experience working with spatial Gaussian Processes and existing literature (Gao et al., 1996; Jones et al., 1998), the inclusion of spatial linear regression may slightly improve the models' predictive qualities. Hence, I chose to use the 4-term regression function, make its form constant, and focus my covariate selection analyses on the stochastic components of GPs.

The starting covariate set contains longitude, latitude, elevation and the 1st EOF of all 7 variables: NOx and VOC emission rates, temperature, wind speed, boundary layer height, NOx and VOC antecedent concentrations. Once again, at any time point t, the antecedent values (NOx and VOC) of each grid cell are the *sum* of antecedent concentration from that cell and neighboring cells weighted via Arcsine Weighting (Section 4.3). Given the starting covariate set, I initiate the iterative improvement algorithm outlined in Section 4.3. This algorithm is stopped once no *statistically significant* improvement can be made by introducing more covariates into the model, and the covariate set at termination is my covariate set of choice. This iterative operation is implemented for the ozone feature models $\mathbf{E}_j$, $j = 1, 2, 3$.

For PC covariate selection, I fix the regression term at $\mathbf{f}(\mathbf{X_P}) = (\mathbf{1}, hour)^{\text{T}}$. I experimented with different functions of $hour$, e.g., $hour^2$ and auto-regressive time series of lag 1. However, added model complexities came with no noticeable improvements in models' predictive qualities. I decided that a simple 2-term regression is sufficient for the $\mathbf{P}_j$ GP models (4.2).

The focus here is once again, selecting covariates for the stochastic process $\mathbf{Z_P}$. For selection of parameters in $\mathbf{Z_{P\_1}}$, $\mathbf{Z_{P\_2}}$ and $\mathbf{Z_{P\_3}}$, the starting

covariates are respectively $\text{Temp}_{P,1}$, $\text{VOC}_{P,1}$ and $\text{NOx-lag}_{P,3}$. They are selected because single-covariate models of $\mathbf{Z}_{\mathbf{P}\_1}$, $\mathbf{Z}_{\mathbf{P}\_2}$ and $\mathbf{Z}_{\mathbf{P}\_3}$ with these covariates have the highest model-fit likelihood. Iterative improvement is then implemented to identify additional model covariates.

4.5.2 Selection Results

Table 4.1 shows the results of covariate selection from iterative improvement using the 2006 training data mentioned in Section 4.4. The covariates shown are in addition to *longitude, latitude and elevation* for the $\mathbf{E}_j$ models. As shown, for both $\mathbf{E}_j$ and $\mathbf{P}_j$ models, the variable *temperature* consistently has multiple EOFs and PCs selected, reflecting its expected importance in an ozone process. Moreover, for $\mathbf{E}_j$ models the chemical precursor variables are more likely than meteorological variables to have covariates included. Table 4.2 shows the cross-validation RMSEs from the models selected by iterative improvement and the standard deviations of their respective training data. The RMSEs can be compared with those of a null model, where the prediction is just the mean of the training data, i.e., the training data standard deviations also shown in Table 4.2. As shown, the fitted ozone feature models have much lower CVRMSE than the data standard deviation.

Figure 4.6 plots the goodness-of-fit results from various model fits for $\mathbf{E}_1$. Performances are compared from four covariate sets:

- Covariates selected through iterative improvement.
- The covariates are the 1st EOFs of model variables.
- All candidate model covariates.
- *Only* longitude, latitude and elevation as covariates.

A circle in the plot indicates that using a likelihood-ratio test with a significance level of 0.05, that there is *no* significant difference between a simpler model and the full model containing all possible covariates. An "x" indicates otherwise.

	Model Covariate
EOF 1	$NOx_{E,1}$, $NOx_{E,2}$, $VOC_{E,1}$, $Temp_{E,1}$, $Temp_{E,2}$, $Wind_{E,1}$, $Wind_{E,2}$, $BL_{E,1}$, NOx-$lag_{E,1}$, VOC-$lag_{E,1}$, VOC-$lag_{E,2}$
EOF 2	$NOx_{E,1}$, $VOC_{E,1}$, $Temp_{E,1}$, $Temp_{E,2}$, $Wind_{E,1}$, $BL_{E,1}$, NOx-$lag_{E,1}$, VOC-$lag_{E,1}$
EOF 3	$NOx_{E,1}$, $VOC_{E,1}$, $Temp_{E,1}$, $Temp_{E,3}$, $Wind_{E,1}$, $BL_{E,1}$, NOx-$lag_{E,1}$, NOx-$lag_{E,2}$, NOx-$lag_{E,3}$, VOC-$lag_{E,1}$
PC 1	$Temp_{P,1}$, $Temp_{P,3}$, $Wind_{P,2}$, $BL_{P,1}$, VOC-$lag_{P,1}$
PC 2	$VOC_{P,1}$, $Temp_{P,3}$, $Wind_{P,1}$, $Wind_{P,3}$, NOx-$lag_{P,1}$
PC 3	NOx-$lag_{P,3}$, $BL_{P,2}$, NOx-$lag_{P,1}$, $Temp_{P,2}$, $Temp_{P,3}$, $Wind_{P,1}$

Table 4.1: Covariate selection results from the iterative improvement algorithm. The data are the 2006 training data mentioned in Section 4.4. "BL" is the boundary layer, "NOx-lag" and "VOC-lag" indicate antecedent precursor concentration. For the EOF models, the covariates shown are those used *in addition* to longitude, latitude and elevation.

	EOF 1	EOF 2	EOF 3
CVRMSE of iterative improvement	0.0006	0.0039	0.0064
Standard deviation	0.015	0.065	0.066
	PC 1	PC 2	PC 3
CVRMSE of iterative improvement	5.60	5.92	3.58
Standard deviation	142.6	117.23	48.98

Table 4.2: Training data Cross-validation RMSE of the models chosen by the iterative improvement method and the standard deviation of the ozone feature data. The units of $\mathbf{P}_j$ model CVRMSE and standard deviation are *ppb*. The CVRMSE and standard deviation are unitless for the $\mathbf{E}_j$'s.

Figure 4.6 shows that the model with only the location variables fits noticeably worse than models containing meteorological and precursor variables. These results highlight that, due to the complex non-linear structures of ozone features, longitude and latitude are not sufficient for ozone feature modelling. Meteorological and ozone precursor variables are statistically important for ozone modelling here.

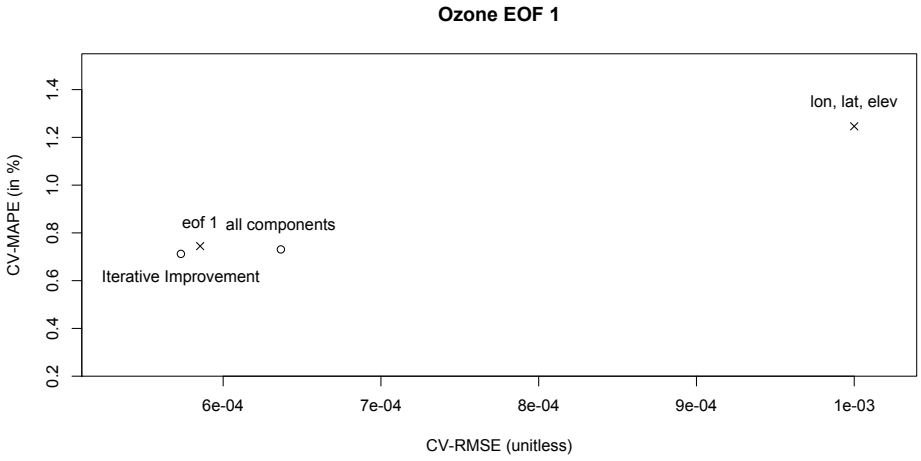


Figure 4.6: Cross-validation MPE versus cross-validation RMSE for four models with different covariate sets. A circle indicates via a likelihood-ratio test, that there is *no* significant difference between a simpler model and the full model.

4.6 Modelling and Forecasting Ozone Features

This section will present results including diagnostic tests of model assumptions and evaluation of the predictive quality of the ozone feature models. Here, "prediction" refers to the forecasts of *future* ozone-feature patterns and space-time ozone fields for the predictive set. This is not to be confused with the cross-validation done in the previous section, which is prediction *within* the training set. Forecasting is a more stringent test of a model's capabilities than cross-validation.

This section presents the results from two types of ozone feature models discussed in Section 4.3:

1. Models where the covariates are spatial and temporal means of model variables, which I refer to as the *Variable Mean* (VM) models. For VM models of $\mathbf{E}_j$, the covariates are model variables' temporal means (averaged across time): these are spatial fields of variable means. For the VM models of $\mathbf{P}_j$, the covariates are model variables' spatial means

(averaged across space): these are time series of variable means.

2. Models where covariates are variable data decomposed into EOFs/PCs and selected via the iterative improvement procedure, the results of which are shown in the previous section. I refer to such ozone feature models as *Covariate Iterative Improvement* (CII) models.

A further note on terminology: the ozone EOF models based on formulations VM and CII are referred to as EOF-VM and EOF-CII models. Similarly, the ozone PC models are PC-VM and PC-CII.

In Section 4.2, I presented the Best Linear Unbiased Predictor (BLUP) for the unobserved response Y_0 given observations $\mathbf{y}$. Its two main properties are:

- Let $\hat{y}_0$ denote the BLUP of Y_0, then by the unbiasedness property $\mathrm{E}(\hat{y}_0) = \mathrm{E}(Y_0)$, and the mean of random variable $Y_0 - \hat{y}_0$ is 0.
- The predictor $\hat{y}_0$ is also the mean of the conditional Normal (Gaussian) distribution $f(Y_0|\mathbf{y})$.

Summarizing the above properties, one would expect that, if Y_0 (and $\mathbf{Y}$) indeed follow a Normal (or MVN) distribution, then the random variable $(Y_0 - \hat{y}_0)/\mathrm{SE}(\hat{y}_0)$ would follow a standard normal distribution (ignoring estimation of the parameters). Therefore, the assumption of Gaussian Processes for EOFs and PCs can be tested by analyzing how closely the distribution of $(Y_0 - \hat{y}_0)/\mathrm{SE}(\hat{y}_0)$ resembles a standard normal. By running cross-validations on the training data and obtaining standardized Cross Validation (CV) errors, I obtain samples of $(Y_0 - \hat{y}_0)/\mathrm{SE}(\hat{y}_0)$. I then plot and examine the standard normal Q-Q plots of the $(Y_0 - \hat{y}_0)/\mathrm{SE}(\hat{y}_0)$ samples to assess the appropriateness of the Gaussian Process assumption (Gao et al., 1996; Jones et al., 1998).

4.6.1 Modelling and Forecasting of Spatial Ozone Features: EOFs

Figure 4.7 shows for the 3 $\mathbf{E}_j$ models based on CII (EOF-CII), standard normal QQ-plots of the standardized cross-validation errors. Given how closely

the scatter plot of "theoretical quantile vs. sample quantile" lie along the line y-axis= x-axis, I conclude that based on this quantile-to-quantile criterion, the assumption of a Gaussian Process is appropriate. The model fitting and diagnostic results are similar for the EOF-VM models (not shown).

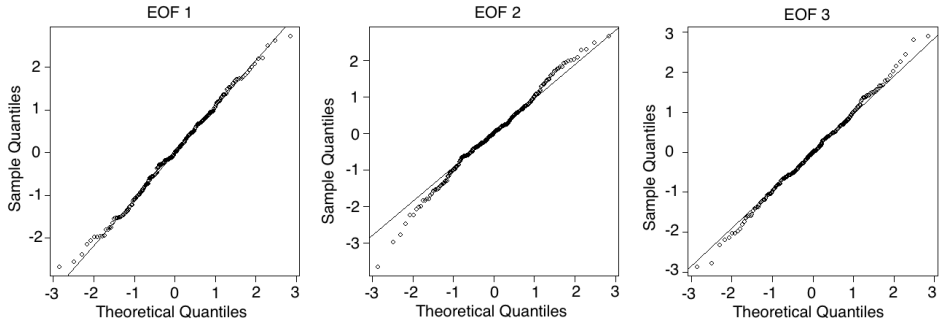


Figure 4.7: For the three EOF-CII models: standard normal QQ-plots of standardized CV-errors. The x-axis is the theoretical quantiles of a standard normal distribution and the y-axis is the sample quantiles from cross-validation.

An important model fitting result is that the fitted correlation smoothness (power) parameters $\hat{\alpha}$ are either equal or very close to 2. A value $\alpha = 2$ indicates an infinitely differentiable and smooth correlation function. In addition, at $\alpha = 2$, a *power-exponential* function becomes a *Gaussian* correlation.

Forecasting the Spatial Features

With GP model parameters estimated from the two days of training data, the corresponding covariate data from the predictive set are used in the multivariate form of BLUP (4.7) to make forecasts of spatial ozone features for the LFV on June 26th, 2006.

Table 4.3 contains the forecast RMSEs calculated as

$$\text{RMSE}_j = \sqrt{\frac{\sum_{i=1}^{n}(E_{ij} - \hat{E}_{ij})^2}{n}},\ j = 1, 2, 3. \tag{4.14}$$

Here, $\hat{E}_{ij}$ denotes the forecast made at location i for j-th ozone feature, and $n = 229$ is the number of forecast locations. When $\hat{\mathbf{E}}_j = \bar{\mathbf{E}}_j$ in (4.14), one obtains the standard deviations of the *true* EOFs being predicted (also in Table 4.3). The true EOFs are decompositions of the ozone data from the predictive set, i.e., they are the real-life spatial ozone features we try to forecast. These true standard deviations provide useful reference points when assessing the scale of RMSE. Note that the response variables are the ozone EOFs, which are unitless.

When compared to the standard deviations of the *true* ozone EOFs, the estimated ozone feature models delivered reasonably low prediction RMSEs. However, it is worth noting that the RMSEs of both $\mathbf{E}_3$ models (VM and CII) are about 77% of the true EOF's standard deviation. As discussed in Chapter 3, the 3rd-order and 4th-order ozone features capture the space-time behaviour of the nocturnal ozone process, which is not as influenced by the model variables as the daytime ozone. In other words, they may not be process-driven enough to be adequately modelled as ozone features. This translates to the difficulties we see in the modelling of $\mathbf{E}_3$. Table 4.3 shows, however, that $\mathbf{E}_1$ and $\mathbf{E}_2$ have smaller forecasting error relative to the standard deviations of the true EOFs.

	EOF 1	EOF 2	EOF 3
RMSE of EOF-VM models	0.009	0.020	0.051
RMSE of EOF-CII models	0.008	0.016	0.047
S.D. of the *true* EOFs	0.020	0.065	0.066

Table 4.3: Prediction RMSE of the EOF models. The last row contains the standard deviations of the *true* EOFs being predicted. The $\mathbf{E}_j$'s are unit less.

Figures 4.8, 4.9 and 4.10 display the spatial patterns of the true EOFs and their predictions. For reference, refer to Figure 4.2 in Section 4.4 for the map of modelled LFV domain. As shown, the model forecasts captured the true EOFs' gross regional-scale spatial patterns as well as some of the finer-scale spatial variations. This type of visual test informs us how closely the

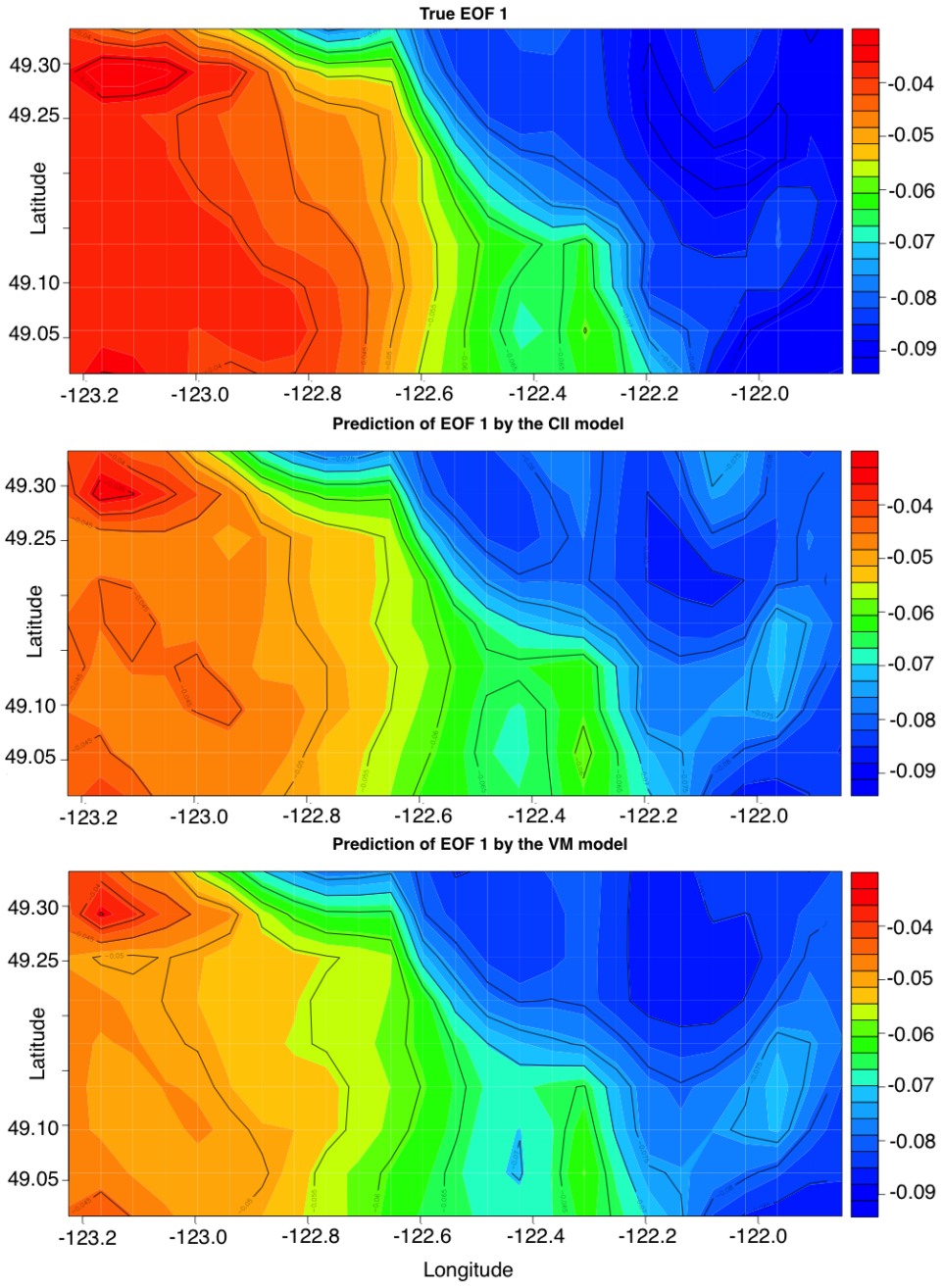


Figure 4.8: Spatial plots of *true* $\mathbf{E}_1$ to be predicted and its GP model predictions (all unitless). The colour scales are the same between plots.

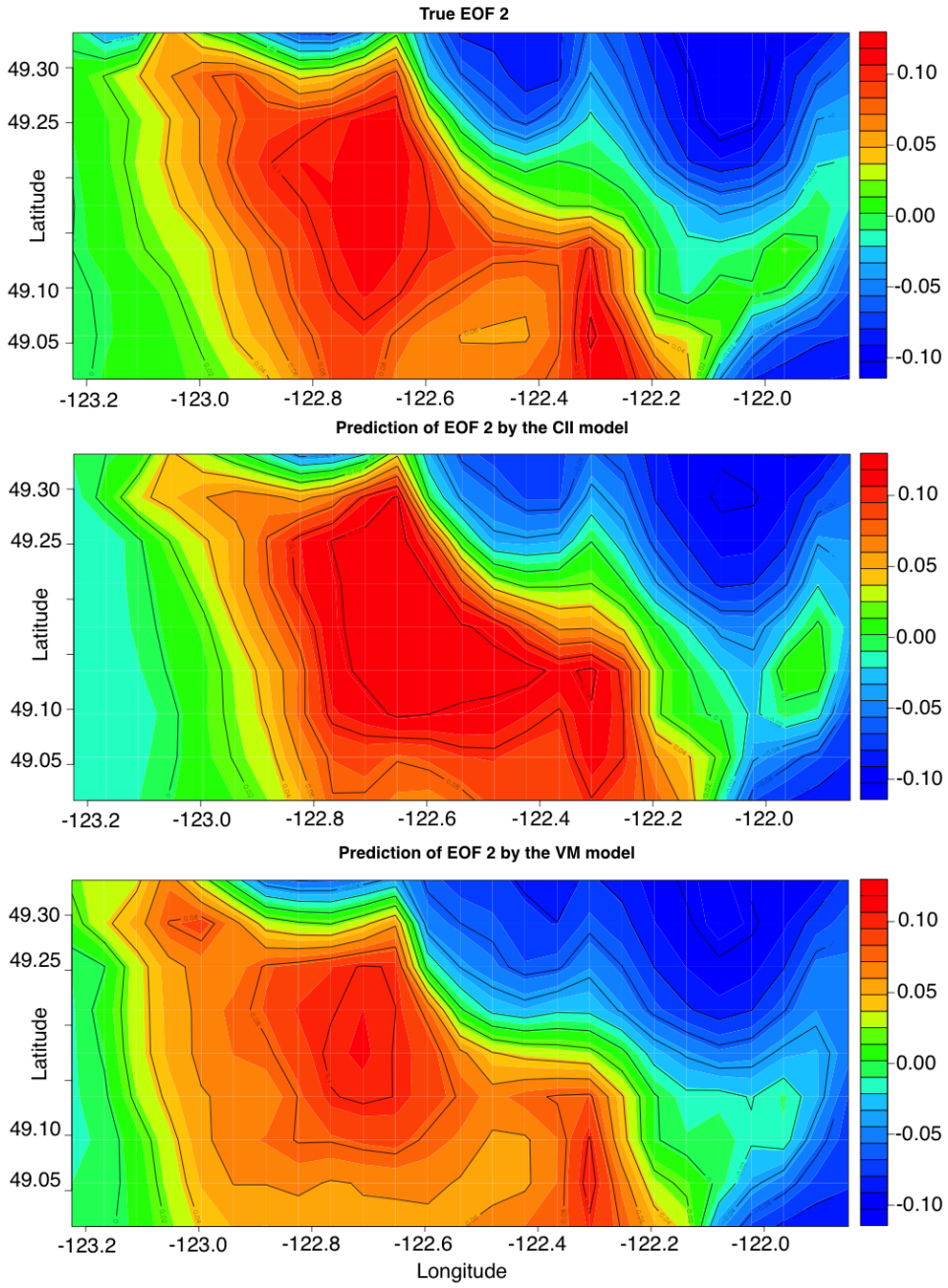


Figure 4.9: Spatial plots of *true* $\mathbf{E}_2$ to be predicted and its GP model predictions (all unitless). The colour scales are the same between plots.

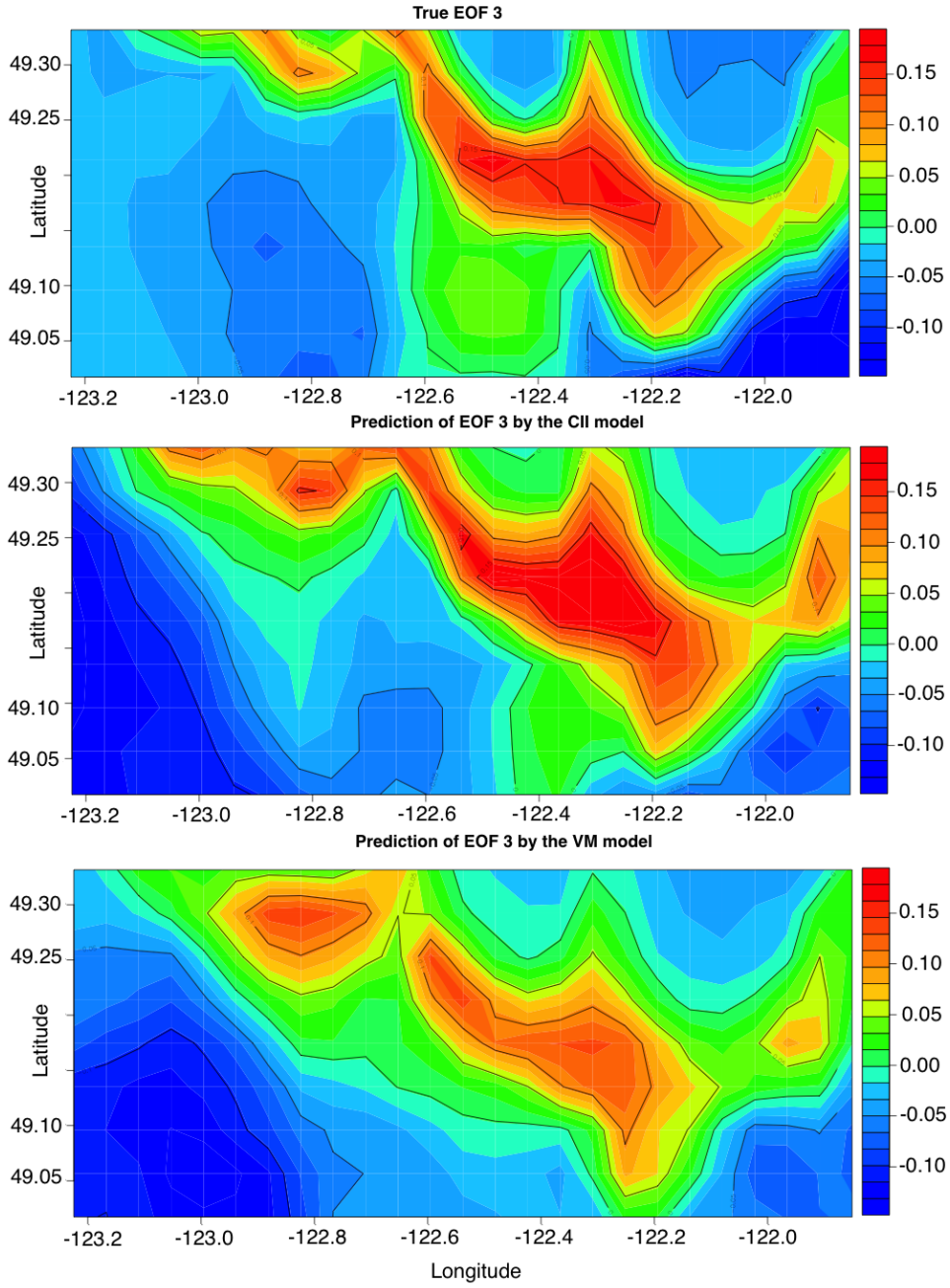


Figure 4.10: Spatial plots of *true* $\mathbf{E}_3$ to be predicted and its GP model predictions (all unitless). The colour scales are the same between plots.

statistical models can predict future spatial variations of the ozone features, which is especially useful in this application of ozone feature modelling. The results show that despite moderate RMSEs, the ozone feature models are capable of forecasting the complex non-linear structure of the leading ozone features.

4.6.2 Modelling and Forecasting of Temporal Ozone Features: PCs

Figure 4.11 shows for the three ozone PC-CII models (the covariate formulations shown in Table 4.1), the standard normal QQ-plots of standardized CV-errors. While the centres of the sample distributions correspond closely to that of a standard normal, the lower tails of $\mathbf{P}_2$ and $\mathbf{P}_3$ are both higher than expected for a standard normal. There are also 3 upper-tail standardized errors that deviated noticeably from the standard normal assumption. Therefore, except for its deficiency in modelling the extremities of higher-order PCs, models with a Gaussian assumption do a satisfactory job of describing the distribution of temporal ozone features. The QQ-plots from the PC-VM model fits (not shown) delivered similar results. In the end, I found little reason to doubt the appropriateness of the Gaussian assumption when modelling ozone features.

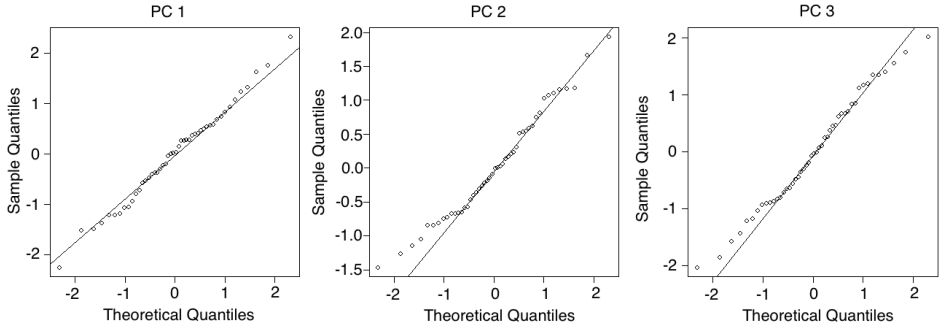


Figure 4.11: For the three PC-CII models: standard normal QQ-plots of standardized CV-errors. The x-axis is the theoretical quantiles of a standard normal distribution and the y-axis is the sample quantiles from cross-validation.

Forecasts of the Temporal Features

Table 4.4 shows the prediction RMSE's along with the standard deviations of the *true* PCs for reference. It should be noted again that the PCs are the weighted row-sums of $\mathbf{O}_{t\times n}$ (Section 3.2), which explains the high RMSEs and standard deviations shown. The space-time ozone field $\hat{\mathbf{O}} = \hat{\mathbf{P}}\hat{\mathbf{E}}^{\mathrm{T}}$ will contain appropriate ozone values once scaled by $\mathbf{E}$. Relative to the standard deviations of the *true* PCs, both $\mathbf{P}_1$ and $\mathbf{P}_2$ predictions have lower RMSEs. Again, the prediction of a higher-order ozone feature, $\mathbf{P}_3$, is relatively less accurate, with an RMSE of about 70% of the standard deviation of true $\mathbf{P}_3$.

	PC 1	PC 2	PC 3
RMSE of PC-CII models	92.21 *ppb*	45.95 *ppb*	42.22 *ppb*
RMSE of PC-VM models	74.29 *ppb*	49.46 *ppb*	40.70 *ppb*
S.D. of the *true* PCs	177.23 *ppb*	113.60 *ppb*	56.64 *ppb*

Table 4.4: Prediction RMSEs of the PC models, where predictions are made using the "real" predictive set on antecedent precursors. The standard deviations of the *true* PCs are shown for comparison.

Figure 4.12 plots the temporal patterns of predicted PCs overlaid with the true PCs. As shown, the temporal patterns of the forecasts reflect the general trends of true PCs. Figure 4.12 does show a few exceptions: my models for $\mathbf{P}_1$ over-predict the LFV ozone means in the early-morning and after 1900PST; there are also slight discrepancies between the true $\mathbf{P}_2$ and predictions during early-morning and late-night. However, my models forecasted the day-time temporal ozone features very well, which is the most important conclusion drawn.

Figure 4.13 shows from the CMAQ 2006 ozone output, the temporal plots of hourly spatial mean, standard deviation (both summarized across space) and the 1st 4 PCs over the course of the whole ozone episode (including both the training and predictive days). There is a trend of decreasing night-time hourly LFV means through the episode, and by the 4th day (predictive set), the hourly LFV mean experiences a dramatic decline from which

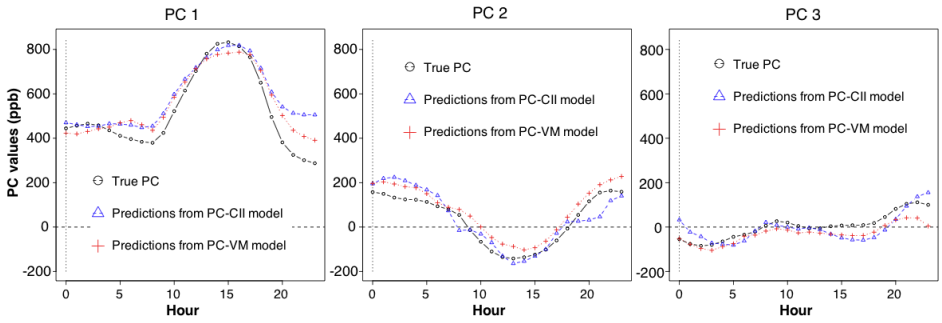


Figure 4.12: Time-series plots of the *true* temporal ozone features in the predictive set (black), their predictions using ozone feature models PC-CII (blue) and PC-VM (red).

it barely recovers as this episode concludes. This daily trend is naturally reflected in the temporal pattern of $\mathbf{P}_1$. Recall that days 2 and 3 are used as the training set, in which this "sudden decline" in night-time ozone is not observed. Therefore, such within-episode variation of the ozone process is not "learned" when estimating the $\mathbf{P}_1$ models, and $\mathbf{P}_1$'s numerical relations with available covariates are not sensitive enough to forecast this night-time feature. In short, this is a problem of extrapolation.

It is worth noting that the magnitude of $\mathbf{P}_1$ is noticeably larger than the others (Figure 4.12). From the ozone prediction function (4.13), one can deduce that $\mathbf{E}_1$ prediction receives a larger weighting in the final ozone modelling/prediction, an expected result given the importance of the spatial/temporal mean to an ozone (or any air pollution) process. The $\mathbf{P}_3$ values are smallest in magnitude, hence assigning the smallest weight towards the $\mathbf{E}_3$ prediction. As a result, the effect of any prediction errors in the 3rd EOF-PC models are subsequently alleviated.

4.7 Forecast of Space-Time Ozone Fields

With the ozone features forecasted, (4.13) is used to forecast the regional ozone fields for the last day of the ozone episode: June 26th, 2006. Once again, the ozone field being modelled is the CMAQ produced output, not

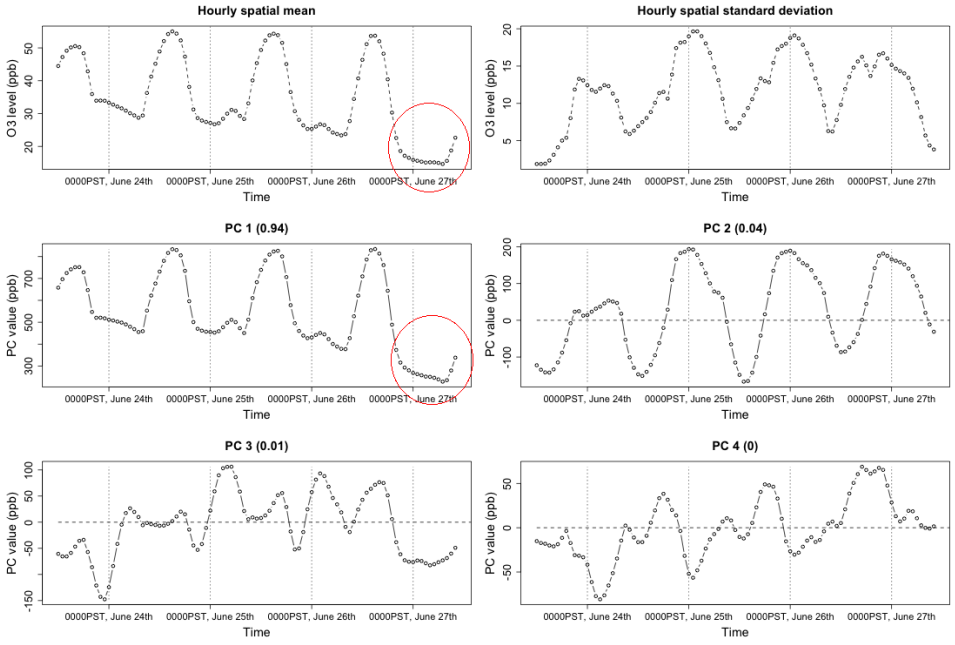


Figure 4.13: Temporal plots of hourly spatial ozone mean, standard deviation and the 1st 4 PCs over the course of the entire episode. Notice the sharp night-time decline of $\mathbf{P}_1$ during the 4th day (predictive set), as highlighted by a red circle. The vertical dashed lines indicate the hour 0000, and the value in the heading of each PC plot is the "proportion of data variation explained".

physical observations.

Figures 4.14 and 4.15 show for selected hours, the scatter plots of the model forecasts against the true CMAQ output. The feature-based model gave good forecasts during the afternoon peak hours: the points are scattered near the line $x = y$. As expected from the over-prediction of $\mathbf{P}_1$ in the last section, nighttime forecasts are higher than the true ozone level at a number of locations.

Figures 4.16 to 4.19 present forecasts for selected hours as regional ozone fields visualized through 3-dimensional plots. There are four types of ozone field: *true* CMAQ output, *true* CMAQ output constructed from only the leading 3 EOFs/PCs, and my predictions using the CII model (the covariates are variable EOFs and PCs selected via iterative improvement) and the VM model (covariates are variables' spatial and temporal means).

There are two sources of prediction (or forecast) error inherent to the feature-based ozone model: (1) the error from directly predicting ozone features, and (2) the error from using $p = 3$, or in general, $p \ll \min(t, n)$ ozone features to predict a complete space-time ozone field. The second error source is extensively discussed in Chapter 3. Hence, I believe it is useful to present the patterns of the *true* regional ozone reconstructed with only 3 EOFs and PCs. Comparison between forecasted hourly ozone fields (bottom plots in each set of four plots in Figures 4.16 to 4.19) to the corresponding "true CMAQ with first 3 EOFs/PCs" (upper right of a set) evaluates the feature-based ozone model's capability from the sole perspective of error source (1) mentioned above.

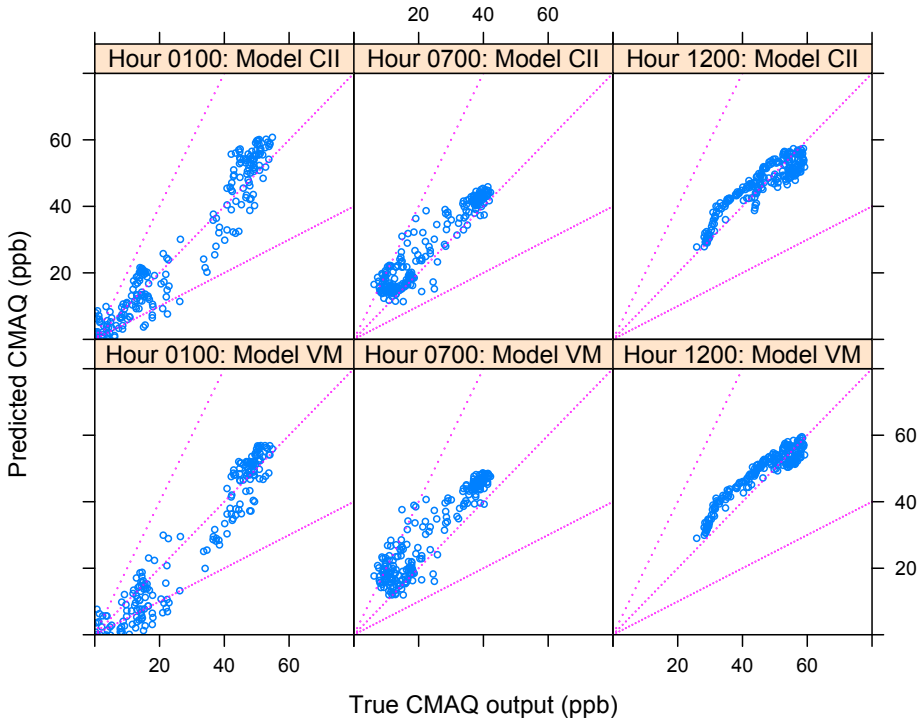


Figure 4.14: For hours 0100, 0700 and 1200 of June 26th, 2006 (the predictive set): the scatter plots of predictions from the CII model and the VM model versus the *true* CMAQ output. The three lines are y=x, y=2x and y= $\frac{1}{2}$x.

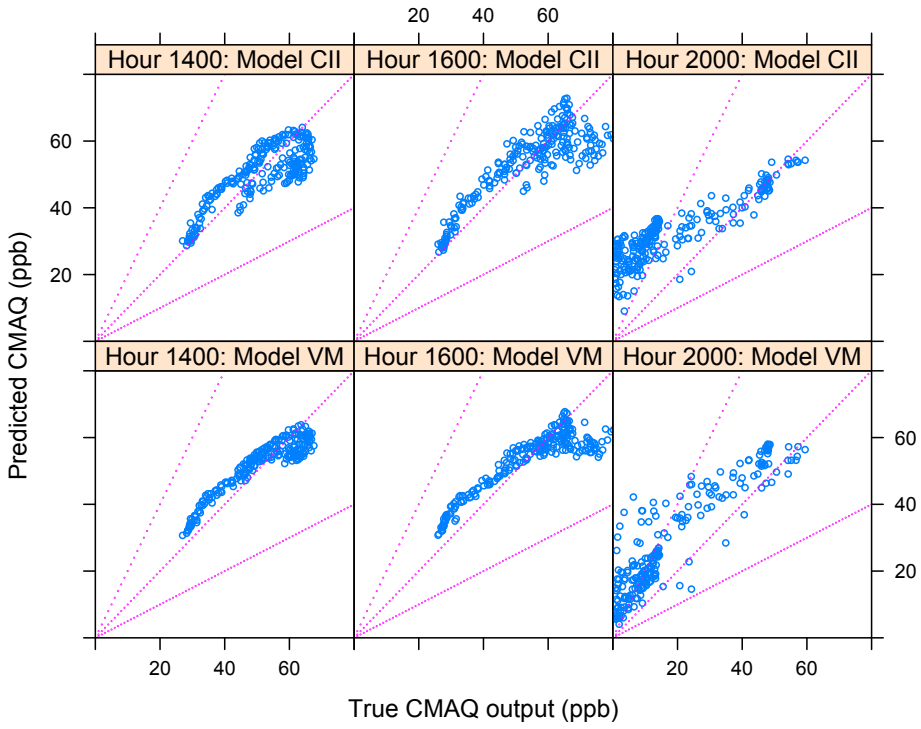


Figure 4.15: For hours 1400, 1600 and 2000 of June 26th, 2006 (the predictive set): the scatter plots of predictions from the CII model and the VM model versus the *true* CMAQ output. The three lines are y=x, y=2x and y= $\frac{1}{2}$x.

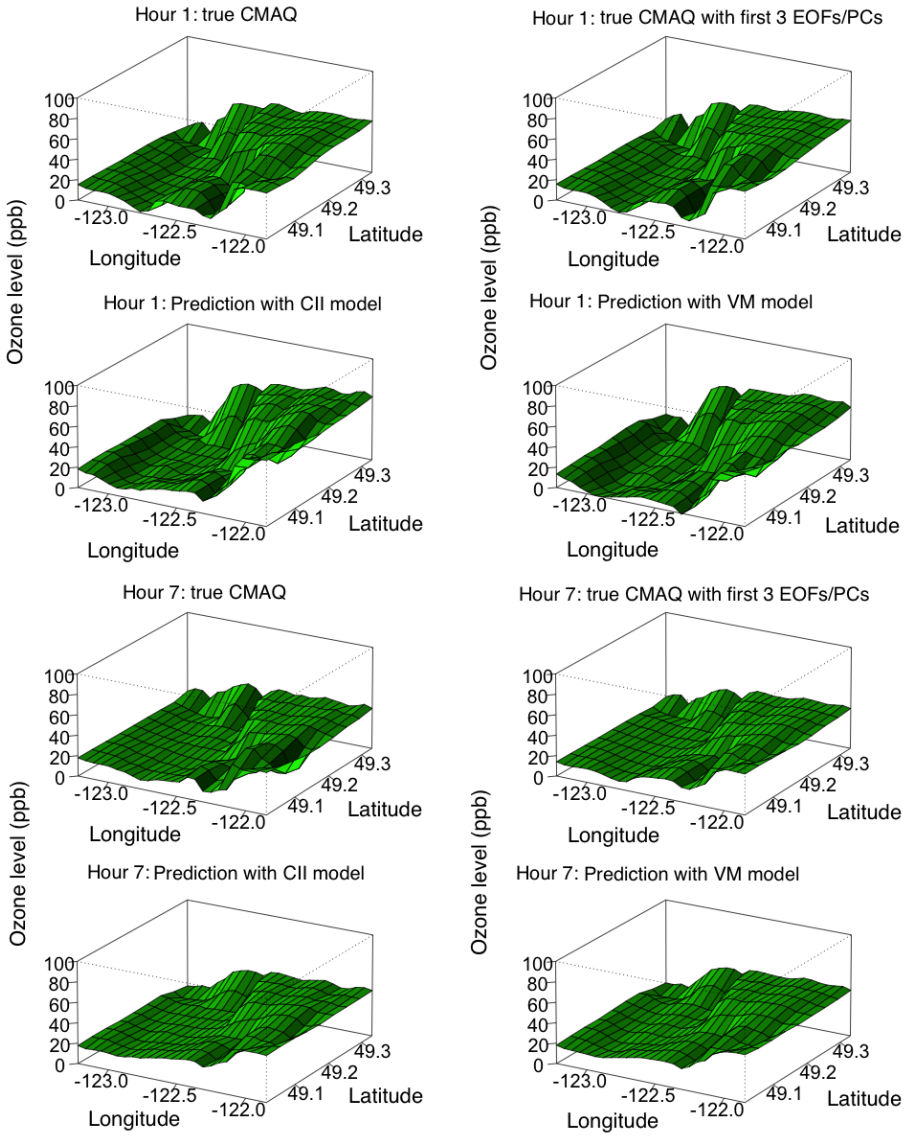


Figure 4.16: Hour 0100 and 0700 of June 26th, 2006 (the predictive set): 3-D spatial ozone fields of *true* CMAQ output (upper-left), *true* CMAQ output with only the first 3 EOFs and PCs (upper-right), forecasts using CII model (lower-left) and VM model (lower-right).

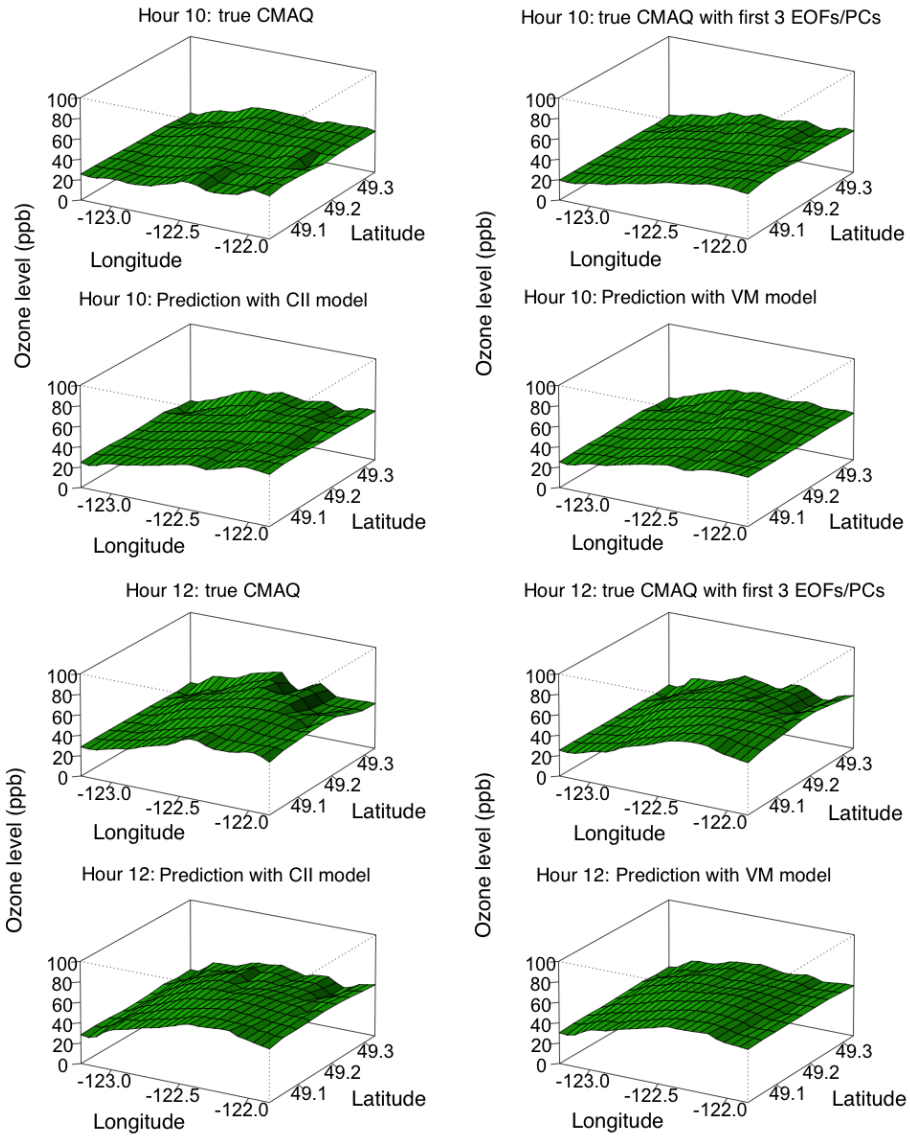


Figure 4.17: Hour 1000 and 1200 of June 26th, 2006 (the predictive set): 3-D spatial ozone fields of *true* CMAQ output (upper-left), *true* CMAQ output with only the first 3 EOFs and PCs (upper-right), forecasts using CII model (lower-left) and VM model (lower-right).

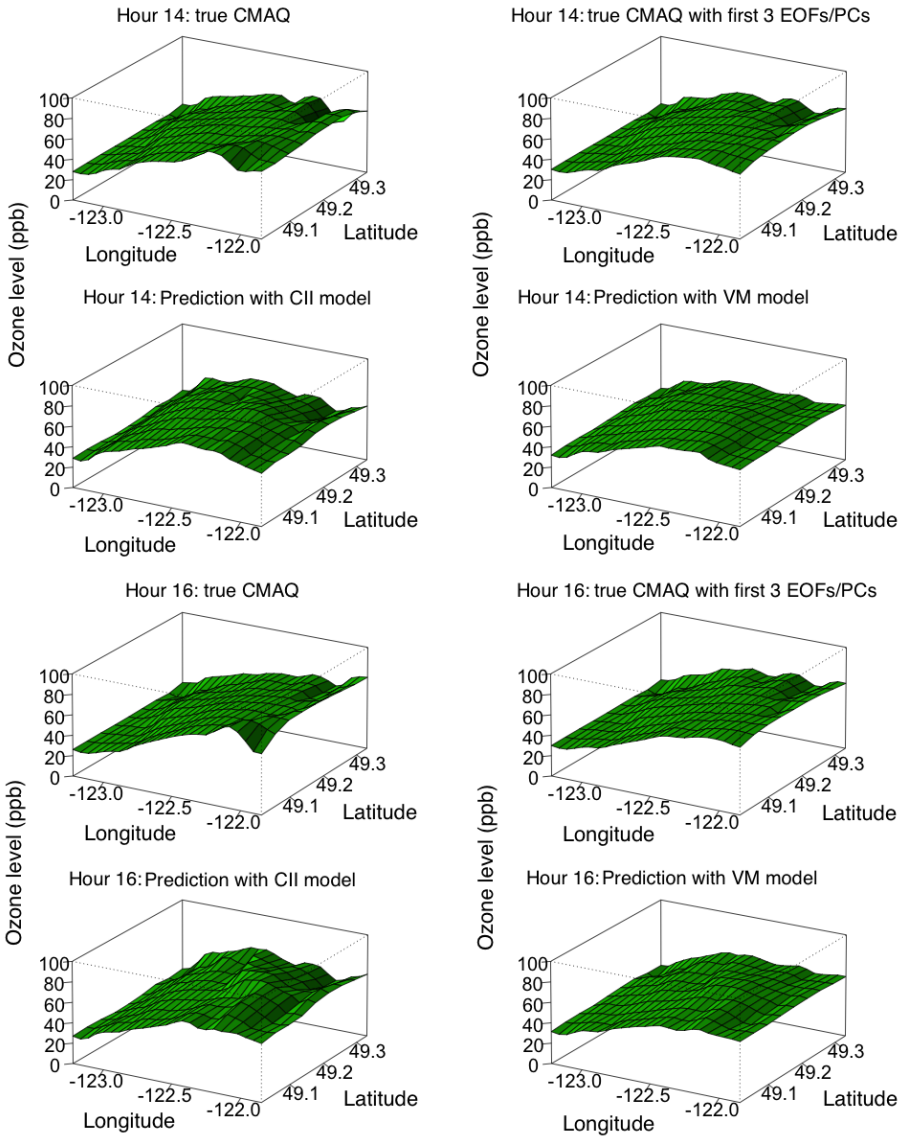


Figure 4.18: Hour 1400 and 1600 of June 26th, 2006 (the predictive set): 3-D spatial ozone fields of *true* CMAQ output (upper-left), *true* CMAQ output with only the first 3 EOFs and PCs (upper-right), forecasts using CII model (lower-left) and VM model (lower-right).

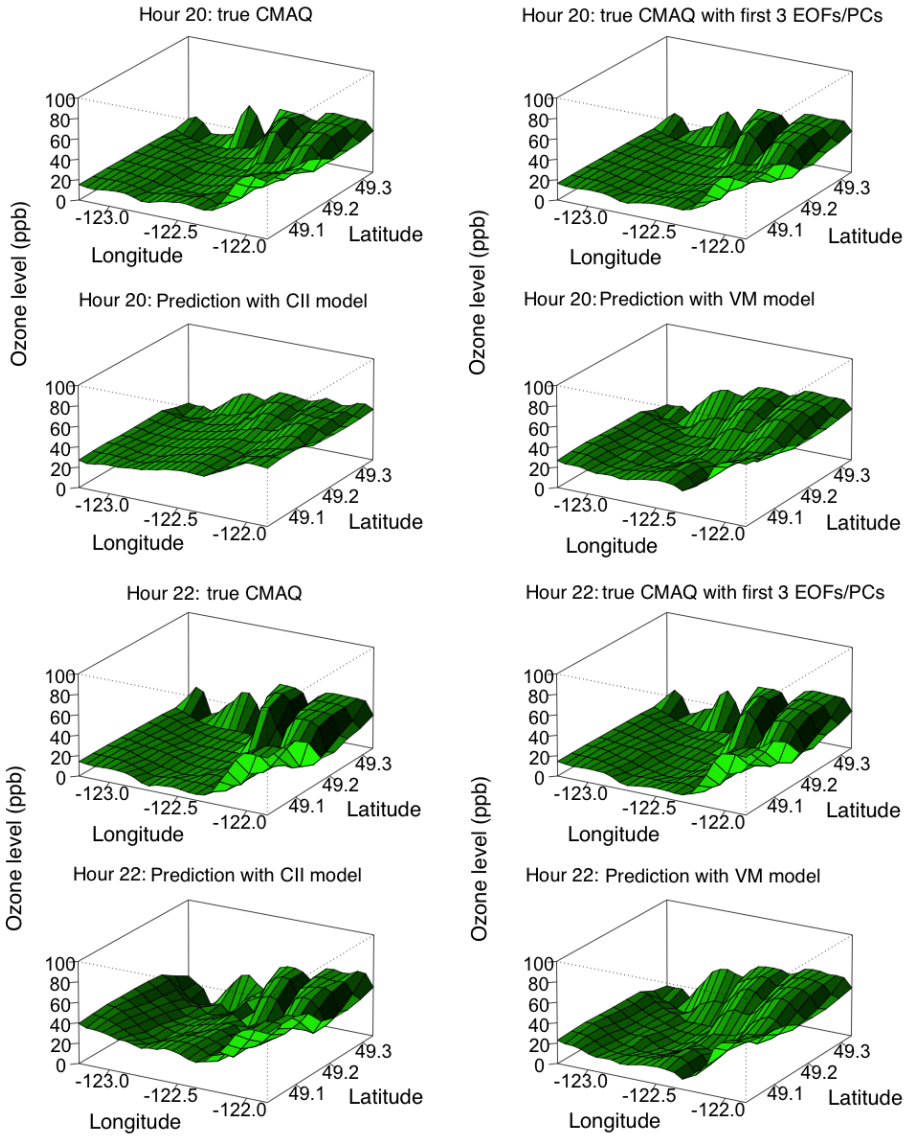


Figure 4.19: Hour 2000 and 2200 of June 26th, 2006 (the predictive set): 3-D spatial ozone fields of *true* CMAQ output (upper-left), *true* CMAQ output with only the first 3 EOFs and PCs (upper-right), forecasts using CII model (lower-left) and VM model (lower-right).

At 0100PST and 0700PST, the spatial patterns of ozone fields resemble those of the regional topography: near background ozone level across the valley, and a high-level ozone plume blanketing the north shore mountains. During these hours my predictions effectively capture this reality, including higher-resolution spatial details such as the "peaks-and-troughs" along the north shore mountains. As one might expect from the forecasts of $\mathbf{P}_1$ (previous section), my CII model over-predicts the southwest LFV ozone fields during the night-time period of 2000PST-2300PST. This over-prediction can be seen from Figure 4.19. Although the VM model also over-predicted the 2000PST CMAQ ozone field, its predictive quality improved noticeably in the following hours, as seen in the prediction for 2200PST (Figure 4.19). Both statistical models' over-prediction at 2000PST is also evident from the scatter plots in Figure 4.15: some of the *true* CMAQ output are near 0 *ppb* while their corresponding predictions are noticeably higher.

However, day-time is the most important period for ozone modelling. Much ozone research focuses on the *mean 8-hour daily maximum*: the average ozone levels during the highest 8-hour window of each day, and government policies and regulations are based on compliance with this statistic (CCME, 2000; Yarwood et al., 2005; Reuten et al., 2012). My *feature-based* ozone models delivered good predictions during the day-time, as evident from both the 3-D spatial plots (Figures 4.17 and 4.18) and scatter plots of *true* CMAQ vs. prediction (hours 1200PST, 1400PST and 1600PST in Figures 4.14 and 4.15).

Table 4.5 shows for the CII and VM models, the hourly prediction RMSEs and MPEs at 3 mid-day hours and the RMSE/MPE summarized over the 8-hour maximum. The table also shows the hourly LFV standard deviations of the *true* ozone fields being predicted: these values help to put in context the scale of the CII/VM model forecasting accuracy. Hourly LFV standard deviation is also mathematically the same as hourly RMSE of predictions made by the true ozone mean of that hour (averaged across space). Using RMSE as the reference, both models displayed similar prediction accuracy during the afternoon ozone peak hours. However, the CII model gave predictions with noticeably smaller MPE throughout the 8-hour daily

maximum as well as the entire 24 forecasting hours (not shown). Furthermore, the hourly LFV standard deviations of the true ozone are more than double the hourly RMSEs of ozone feature models. The prediction MPE is near 0% at 1300PST for the CII model, and the MPEs of the VM model are consistently higher than the CII model in magnitude throughout the diurnal cycle.

	Hour (PST)			Hours from
	1200	1300	1400	8-hour maximum
RMSE of CII (in *ppb*)	4.08	4.24	7.19	7.60
RMSE of VM (in *ppb*)	4.34	5.24	6.88	7.50
Std. deviation of true data	14.78	15.59	16.24	16.84
MPE of CII (in %)	-2.08	0.33	1.18	-4.95
MPE of VM (in %)	-6.52	-3.09	-1.28	-7.90

Table 4.5: Prediction RMSE and MPE from the two *feature-based* ozone models and the standard deviation of the *true* ozone field. The prediction statistics are presented as hourly value for hours 1200PST, 1300PST, 1400PST and summarized across the hours during the 8-hour ozone maximum. The forecasting day is June 26th, 2006 and the spatial domain is the rectangular LFV field (Figure 4.2).

	RMSE in *ppb*		MPE in %	
	day 1	day 2	day 1	day 2
CII model	2.84	2.80	-0.85	-0.63
VM model	3.45	3.47	-0.82	-0.47

Table 4.6: RMSE and MPE of cross-validation predictions made on complete ozone fields. The statistics are summarized over the 8-hour ozone maximum of June 24th and 25th, 2006 (day 1 and 2 from training data). The spatial domain is the rectangular LFV field (Figure 4.2).

Overall, the ozone feature models delivered space-time ozone forecasts with reasonably low error-statistics, and their predictions manage to capture the complex spatial structures of LFV's hourly ozone fields through a diurnal cycle. One caveat is their over-predictions during night-time hours

(especially for the CII model). This is the result of over-predicting the tail of $\mathbf{P}_1$, which as discussed in the previous section, is due to extrapolation. However, this over-prediction is limited to a particular area of LFV (the southwest) and the hours of 2000PST and 2100PST.

Table 4.6 shows the RMSE and MPE of cross-validation predictions of complete ozone fields for the two days of training data. This is analogous to the Table 4.2, where CV-RMSE and CV-MPE of individual ozone feature models are shown. The RMSEs and MPEs are summarized over the 8-hour maximum of each day. The cross-validation statistics are described in Section 4.3, but now applied to reconstructed ozone rather than the features. As shown, compared to the predictions done for ozone fields *outside* of the training data (Table 4.5), both RMSE and MPE are noticeably smaller for the cross-validation. This is especially true for the percentage of mean prediction error, where the MPEs are $< 1\%$ for cross-validation compared to -4.96% and -7.90% for ozone forecasting (Table 4.5).

The accuracy of cross-validation RMSE and MPE may be viewed as a goodness-of-fit statistics for the ozone feature models: CV-RMSE and CV-MPE show how well a statistical model can emulate the space-time structure of LFV ozone in the training set. As results from cross-validations have shown, both types of ozone feature models are capable of modelling a complex weather and pollution driven regional ozone process.

4.8 Model Fits from other Episodes

Ozone feature models like those developed in this chapter will be applied to AQM evaluations in Chapters 5 and 6. As I will discuss in Section 5.1, the evaluations to be presented in this thesis are done for individual episodes. Hence, the ozone feature models were fitted *per-episode* in this chapter, and the 2006 model in particular was analyzed in detail between Sections 4.4 to 4.7. Tables 4.7 and 4.8 show the cross-validation RMSEs of the fitted models for the other four episodes along with the standard deviations of the training data. As shown, all fitted models have CVRMSEs that are much lower than the standard deviations, and these CVRMSEs are comparable

across episodes between 1985 and 2001.

	Episode			
	1985	1995	1998	2001
$\mathbf{E}_1$	0.0008	0.0007	0.001	0.0008
	(0.016)	(0.011)	(0.011)	(0.016)
$\mathbf{E}_2$	0.0033	0.0026	0.0038	0.0037
	(0.065)	(0.066)	(0.065)	(0.065)
$\mathbf{E}_3$	0.0071	0.0062	0.0071	0.005
	(0.066)	(0.066)	(0.066)	(0.066)

Table 4.7: Cross-validation RMSE of the fitted $\mathbf{E}_1$, $\mathbf{E}_2$ and $\mathbf{E}_3$ models, whose covariates are selected by iterative improvement. The numbers in parentheses are standard deviations of the training data. The CVRMSE and standard deviation are unitless for the $\mathbf{E}_j$'s.

	Episode			
	1985	1995	1998	2001
$\mathbf{P}_1$ (*ppb*)	17.46	17.29	19.62	15.9
	(216.02)	(252.87)	(335.12)	(230.63)
$\mathbf{P}_2$ (*ppb*)	17.91	20.33	10.67	15.86
	(159.49)	(153.31)	(188.21)	(138.16)
$\mathbf{P}_3$ (*ppb*)	10.77	15.15	10.37	11.06
	(71.31)	(75.10)	(91.27)	(76.51)

Table 4.8: Cross-validation RMSE of the fitted $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ models, whose covariates are selected by iterative improvement. The numbers in parentheses are standard deviations of the training data. The CVRMSE and standard deviation have units *ppb*.

An alternative model-fitting procedure is to merge all CMAQ outputs and fit ozone feature models describing all episodes. This is a reasonable approach if the objective is to estimate a statistical emulator of CMAQ process. However, as mentioned the ozone feature models in this chapter were estimated for *per-episode* CMAQ evaluations. Therefore, I decided to not pursue the "merged-data" approach to model estimation.

4.9 Chapter Conclusion

In this chapter, statistical models of spatial-temporal ozone features are developed. These ozone feature models displayed notable capability in modelling the non-linear structures of the ozone features.

Individual features are modelled as GPs driven by a set of variables describing background temperature, wind speed, planetary boundary layer height, ozone precursor emission rates and ambient concentrations. The covariates of each feature model are selected through a forward selection algorithm based on a combination of goodness-of-fit statistics. The models are then fitted by maximizing the GP profile-likelihood. The fits and predictive capabilities of individual ozone feature models are evaluated through diagnostic tests, cross-validation and feature forecasting. Here, the forecasts are made for the 4th day of the 2006 CMAQ output across a complex spatial domain including LFV and surrounding mountains.

The ozone feature models proved their capability in forecasting the complex non-linear structures of the spatial ozone features, where both the regional-scale patterns and localized details of the *true* features are captured by model forecasts with good numerical accuracy (Section 4.6). Temporal ozone feature models displayed appropriate goodness-of-fits through low values of cross-validation RMSE (Section 4.5). With the exception of the night-time forecast of $\mathbf{P}_1$, the temporal ozone feature models satisfactorily forecasted the temporal patterns and values of the *true* features (Section 4.6).

By combining the predicted ozone features via equation 4.13, forecasts were also made for the complete space-time ozone field. The *feature-based ozone model* is able to forecast the hourly LFV ozone fields at great spatial resolution: compared to the *true* ozone fields being forecasted, the statistical model predictions captured the detailed local ozone variations both in the lower-valley region and across north shore mountains. The forecasting accuracy was especially good during the important daily ozone peak hours, with low RMSEs of about 4 *ppb* to 7 *ppb* and near 0 prediction biases.

Use of ozone feature models in Chapters 5 and 6

Compared to the spatial domain of the ozone field modelled in this chapter, the CMAQ evaluation analyses in the following chapters involve the modelling of a much smaller LFV sub-domain: area within the boundary of Metro Vancouver monitoring stations. Hence, the ozone feature models developed here should be well qualified to model a simpler ozone field, thus serving their original intended purpose of CMAQ evaluation.

An efficient and capable model of space-time ozone process

The analyses in this chapter showed that a complex space-time ozone can be modelled through a few ozone features, i.e., data components with simpler structures. This *feature-based* ozone model combines the methods of PCA and GP, and it is a novel approach for modelling a space-time air pollution process. Furthermore, several variables are identified to be useful for modelling ozone features, and data on wind direction can be used to create an useful new variable representing the space-time field of antecedent ozone precursor concentrations.

In addition to the already established modelling capability, the feature-based ozone model is also a computationally efficient means of modelling a large air quality dataset. Let N be the size of the data used in modelling; the ozone feature models has $N = n$ or $N = t$, while the direct modelling of "raw" data has $N = n * t$. The GP modelling is a $\mathcal{O}(N^3)$ function (Sacks et al., 1989), which means that the rate of increase in the computational load is defined by the cube of N. Hence, the computational efficiency of the statistical models is sensitive to data size. In this case, we are comparing the computation loads involving $N = 229$ or $N = 48$ with $N = 229 * 48 = 10992$. This notable computational efficiency will prove useful when emulating an AQM process, which typically generates large datasets.

Lastly, statistical theory suggests that feature-based predictions of ozone via reconstruction may be biased. As discussed, the predictions of individual ozone features are unbiased in the sense that they are based on BLUPs. However, the equation (4.13) for the *ozone field* is not a BLUP. This topic

is also explored in my research and I found that the problem of prediction bias is not an overriding concern here. Appendix C.2 presents statistical analysis of prediction bias.

Chapter 5

AQM Evaluation I: Comparison of Ozone Features and Modelling of Feature Differences

The conventional way of evaluating air quality models is to compare model outputs and observations at a point location and time (Dennis et al., 2010). The output-observation differences are then summarized by statistics such as RMSE and MPE (mean percentage error). Preisendorfer and Barnett (1983) and Willmot et al. (1985) further used sampling methods to estimate the statistical confidence and significance of error statistics. However, while point-based comparison can be useful "up to a point", without a process-level understandings of the compared air pollution fields at hand, any observation-model agreement (or disagreement) should be deemed "fortuitous" (Dennis et al., 2010). The reasoning behind this assertion is discussed extensively in the introduction Section 1.2.

As mentioned in the introduction, this research is motivated by the need for a more informative means of air quality model evaluation (Galmarini and Steyn, 2010). This thesis proposes two general AQM evaluation approaches, both based on ozone features. These methods are then implemented to evaluate CMAQ outputs for LFV ozone episodes. Both methods apply the statistical tools formulated in Chapters 3 and 4: methods for analyzing and modelling ozone features.

Evaluation I: Ozone Feature Comparison and Feature Difference Model

First, I propose to compare individual spatial and temporal ozone features between AQM output and observations. In Chapter 3, I interpreted the types of spatial-temporal ozone features that define an LFV ozone process. In addition to capturing the structure of space-time ozone mean relationships, I also identified the most dominant pattern(s) of ozone advection (movement of the ozone plume) across the LFV.

The comparison of features between AQM and observations is a means of evaluation that addresses the need for a more insightful model evaluation. Feature based observation-AQM comparison allows for evaluations of underlying space-time structures and dynamic processes, e.g., evaluate whether AQM can capture observed patterns of ozone advection, or model the east-west variation of ozone means caused by westerly wind regime (typical of an ozone episode in LFV).

Secondly, I propose to statistically model the *ozone feature differences* using the GP model and covariates determined in Chapter 4. Feature difference between AQM and observation can be summarized into error statistics such as RMSE and other tests for significance. However, such analysis only provides one summarized value of observation-AQM distance without providing insight into the pattern and behaviour of observation-AQM difference.

By analyzing the statistical association between observation-AQM feature difference and various conditions of an AQM run, one may (1) identify the specific AQM input(s) most responsible for its modelling deficiencies, and (2) associate the observation-AQM difference with the spatial or temporal variations of said model input(s). This is another means of informative evaluation of AQM.

AQMs such as CMAQ model the grid-cell ozone average calculated from a set of initial and boundary conditions, whereas the observations are recordings of air pollution levels at points in space and time (Section 1.2). In other words, AQM outputs and physical observation are defined by discrepant spatial scales and processes (Dennis et al., 2010), which can make direct

observation-model comparisons questionable.

The spatial discrepancy between computer models and observations is not an obvious problem when comparing spatial ozone features $\mathbf{E}_j$. This is because the values in an $\mathbf{E}_j$ are no longer indexed by either grid-cell average or point location, the $\mathbf{E}_j$ are spatial weights that describe specific patterns of ozone variation in space. We are comparing data structures or "summaries" rather than data values, thereby avoiding the problems stemming from direct data comparison. This is an important point that I have not seen raised by existing literature in PCA-based AQM evaluation (those reviewed in Section 1.4).

In this chapter, the proposed "AQM Evaluation I" will be described in detail and implemented to evaluate the CMAQ performance against physical observations for 5 ozone episodes in 1985, 1995, 1998, 2001 and 2006. Although the chapter focuses on the first method, the second method is also outlined here to give an overview.

Evaluation II: Comparison of AQM and Observation as Stochastic Ozone Processes

Another proposed AQM evaluation approach aims to answer the question: given the same basic conditions in background weather and precursor pollution, will AQM and the physical process produce similar ozone features?

This evaluation is implemented by first building GP-based ozone feature models for both AQM ozone and physical observation. Comparisons are then made between the ozone features produced by the two processes (model predictions) under the *same* covariate settings that represent various background conditions.

At a covariate setting, GP model can produce the estimated *process mean* and *standard deviation* at this particular condition. By comparing the outputs of two GP models under the same setting, one compares the statistical properties of two ozone processes that generated AQM outputs and observation data. This point will be discussed more extensively in Chapter 6.

Figure 5.1 shows in diagrammatic form the central idea behind the proposed AQM evaluation (in the context of CMAQ) and its departure from the conventional point-to-point AQM evaluation. As I discussed in Section 1.2, AQM output and observation data are generated by different air pollution processes, and direct data comparison only serves to inform the deviation in their values, not the difference in their behaviour as air pollution processes.

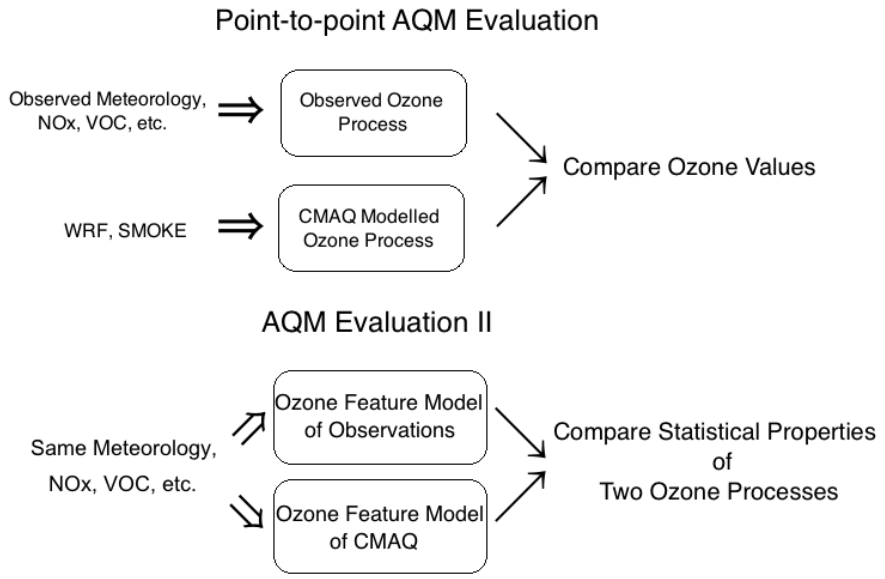


Figure 5.1: Schematics of the idea behind the "AQM/CMAQ Evaluation II" and the "traditional" point-to-point approach.

My second proposed evaluation method provides the type of insights not obtainable from either point-to-point evaluations or mere comparison of air pollution features. The advantage of such analysis is that it evaluates AQM simulated ozone field against observations as comparable stochastic processes: both are described by GPs that are governed by the *same* sets of background conditions.

Feature Correspondence during AQM Evaluation

To successfully implement my proposed AQM evaluation, it is crucial to ensure feature correspondence: the evaluated AQM and observed ozone features are indeed comparable. This point was discussed in Chapter 3, and the analyses in that chapter formulated a set of PCA procedure and interpretations of ozone features that address the topic of feature comparability. These analyses are means of addressing the complications from EOF/PC sampling uncertainty. The following is a recap.

First, I found that PCA of original uncentered $\mathbf{O}_{t\times n}$ will extract the spatial and temporal means as the first feature. Specifically, $\mathbf{E}_1$ will capture the structure of the mean field (temporal means) and $\mathbf{P}_1$ represents the time-series of hourly spatial ozone means. Later analyses will also show that the observed $\mathbf{E}_1$ and $\mathbf{P}_1$ also capture the mean structures of observation data. Therefore, the PCA of uncentered $\mathbf{O}_{t\times n}$ ensures that the first and the most important ozone features are indeed comparable, and that they are well-estimated.

Secondly, I interpreted the other leading features as dynamic data modes that capture the space-time patterns of diurnal ozone advection, as well as the area and magnitude of ozone plume formation. These dynamic structures are interactions between spatial and temporal ozone contrasts ($\mathbf{E}_j$ and $\mathbf{P}_j$). I also found that certain leading features capture more localized ozone variations during the less important nocturnal hours. Moreover, PCA sampling stability analysis indicated that the aforementioned LFV feature interpretations are applicable to smaller sub-domains. Detailed understanding of ozone features allows for an informative and defensible evaluation of AQM features: the evaluated features are comparable because we already understood what they are.

Lastly, the test of ozone feature degeneracy can inform on which features can be analyzed individually or jointly. The analysis of feature degeneracy is a means of measuring the extent of EOF estimation error during AQM evaluation.

The Purpose of the Proposed Evaluation Methods

In summary, my proposed AQM evaluation approaches are designed to address the shortcomings of conventional methods: lack of of informative comparison due to observation-AQM discrepancies in physical scale, ozone-producing conditions and the underlying stochastic process. The proposed methods in this thesis also aim to add to the existing "statistical toolset" for PCA-based AQM evaluations, via rigorous statistical analysis and modelling of air pollution features.

Usually, simulation or sampling-based approaches are used to assess the usefulness of a statistical model evaluation method. However, such an exercise is mainly beneficial when developing a statistical measure that summarizes point-to-point differences between two dataset.

The proposed CMAQ evaluation involves the modelling of CMAQ and observation ozone features, as well as the feature differences as processes driven by background meteorology and atmospheric pollution. Although there are ways of simulating LFV ozone fields driven by temperature and wind (Appendix A.1), it is difficult to build an ozone simulation that accurately describes the complex interactions between LFV emission, meteorology and surrounding topography. In fact, the best ozone simulation is from CMAQ itself.

Therefore in this study, the means of assessing the usefulness of an evaluation method will be based on whether this method can provide results and insights into observation-CMAQ differences that are sensible and explainable by existing knowledge of the LFV pollution process.

5.1 Evaluation Methods and Strategy

I denote the CMAQ ozone output as $\mathbf{O}^c_{t\times n}$ and observed ozone data as $\mathbf{O}^o_{t\times n}$. Suppose either dataset is decomposed into $\mathbf{E}_{n\times n}$ and $\mathbf{P}_{t\times n}$, $\mathbf{E}_j$ and $\mathbf{P}_j$ then represent spatial and temporal features. Let $\mathbf{E}^c_j$ and $\mathbf{P}^c_j$ denote the features of CMAQ, and $\mathbf{E}^o_j$ and $\mathbf{P}^o_j$ the features of observations. Ozone feature differences will be denoted by $\mathbf{E}^d_j = \mathbf{E}^o_j - \mathbf{E}^c_j$ and $\mathbf{P}^d_j = \mathbf{P}^o_j - \mathbf{P}^c_j$ for

any j. Furthermore, PCA will be implemented on the original data *without* column-centering, and rotation of $\mathbf{E}_j$ is *not* applied (Section 3.2).

In the previous 2 chapters, $\mathbf{E}_j$ is analyzed and modelled as unitless (normalized) eigenvectors of $\mathbf{O}^{\mathrm{T}}\mathbf{O}$. In the context of CMAQ evaluation, the comparison of $\mathbf{E}_1^c$ and $\mathbf{E}_1^o$ is the unitless comparison of *spatial variations* of temporal ozone means. That is, the spatial weights in $\mathbf{E}_1$'s are compared without considering the magnitude of ozone values from both data.

In order to evaluate CMAQ's capability in capturing both the spatial patterns and the numerical values of the observed mean fields, one may multiply $\mathbf{E}_1^c$ and $\mathbf{E}_1^o$ by their respective PCA eigenvalues $\sqrt{\lambda_1^c}$ and $\sqrt{\lambda_1^o}$. This way, the magnitudes of data values are incorporated into the spatial ozone features. I denote this scaled $\mathbf{E}_1$ as $\tilde{\mathbf{E}}_1 = \mathbf{E}_1\sqrt{\lambda_1}$, and this can be done for all $j \geq 2$ ozone features under evaluation. The observation-CMAQ difference of scaled-EOF is denoted as

$$\tilde{\mathbf{E}}_j^d = \mathbf{E}_j^o\sqrt{\lambda_j^o} - \mathbf{E}_j^c\sqrt{\lambda_j^c}, \;\; j = 1, \ldots, p.$$

It is worth noting that $\sqrt{\lambda_1}$-scaling is done for the purpose of this AQM evaluation. The eigenvalues are always incorporated into $\mathbf{P}_j$ for my specific data decomposition (Section 3.2), so the reconstruction of $\mathbf{O}_{t\times n}$ still requires the use of $\mathbf{E}_1$.

5.1.1 Model of Feature Differences $\tilde{\mathbf{E}}_j^d$ and $\mathbf{P}_j^d$

Kennedy and O'Hagan (2001) formulated an often used mathematical relationship between computer model outputs and their associated physical data. Using the ozone feature notation of this thesis, for scaled-EOF and PC the formulas are:

$$\tilde{\mathbf{E}}_j^o = \tilde{\mathbf{E}}_j^c|\mathbf{X}_{\tilde{E}^c\_j} + \tilde{\mathbf{E}}_j^d|\mathbf{X}_{\tilde{E}^d\_j} + \varepsilon_{E\_j}, \text{ and} \tag{5.1}$$

$$\mathbf{P}_j^o = \mathbf{P}_j^c|\mathbf{X}_{P^c\_j} + \mathbf{P}_j^d|\mathbf{X}_{P^d\_j} + \varepsilon_{P\_j}. \tag{5.2}$$

Here, $\tilde{\mathbf{E}}_j^d$ and $\mathbf{P}_j^d$ are the spatial and temporal ozone feature difference: they are multivariate random processes representing the modelling deficiencies of

CMAQ. Equation 5.1 is also applicable to $\mathbf{E}_j$, the following CMAQ evaluation will be focused on $\tilde{\mathbf{E}}$ due to its easier interpretability.

A few more comments regarding the observation-CMAQ relationship formulated by equations (5.1) and (5.2):

- $\mathbf{X}_{\tilde{E}^c\_j}$ and $\mathbf{X}_{P^c\_j}$ are the covariates representing CMAQ ozone features.
- $\mathbf{X}_{\tilde{E}^d\_j}$ and $\mathbf{X}_{P^d\_j}$ are covariates of the random processes representing CMAQ modelling deficiency; they are either equal to $\mathbf{X}_{\tilde{E}^c\_j}$ and $\mathbf{X}_{P^c\_j}$ or are subsets of them. The latter case indicates that CMAQ inadequacies are influenced by a few particular CMAQ inputs.
- ε_{E_j} and ε_{P_j} are random observation errors are assumed to be $i.i.d$ $N(0, \sigma_{Ej})$ and $N(0, \sigma_{Pj})$.
- The above formulation implies that $\tilde{\mathbf{E}}_j^c + \tilde{\mathbf{E}}_j^d$ and $\tilde{\mathbf{P}}_j^c + \tilde{\mathbf{P}}_j^d$ are the *true* underlying mean processes representing the j-th ozone feature. Physical observation is the sum of true ozone feature and random error.
- The $\tilde{\mathbf{E}}_j^d$ (similarly for $\mathbf{P}_j^d$ and $\mathbf{E}_j^d$) is defined as $\tilde{\mathbf{E}}_j^o - \tilde{\mathbf{E}}_j^c$: a negative difference is interpreted as CMAQ over-estimate of observed feature and vice versa.

The model diagnostic and goodness-of-fit assessments in Sections 4.5 and 4.6 supported the GP assumption for individual ozone features of orders $j = 1, 2, 3$. If one regards $\mathbf{E}_j$ and $\mathbf{P}_j$ as Gaussian Processes, then it is reasonable to model $\tilde{\mathbf{E}}_j^d$'s and $\mathbf{P}_j^d$'s as GPs.

For temporal computer models, Guttorp and Walden (1987) further proposed that the model deficiency be represented by two terms: one for inadequacy in describing the physical system, one for not tracking the extreme observations due to model outputs being temporal averages. The second source of model deficiency should not be a concern here: the CMAQ runs are specifically used to simulate an extreme event (ozone episode), and the outputs are generated on a high-resolution spatial grid during the *same* time

period as the physical process. Hence, a single term AQM deficiency process ($\tilde{\mathbf{E}}_j^d$ or $\mathbf{P}_j^d$) is sufficient for my evaluation.

By analyzing the importance of each covariate in $\mathbf{X}_{\tilde{E}^d\_j}$ and $\mathbf{X}_{P^d\_j}$ for the processes $\tilde{\mathbf{E}}_j^d$ and $\mathbf{P}_j^d$, one may formulate insights into the underlying association between the magnitude of ozone feature difference and specific covariate/input of AQM modelling run. It provides the CMAQ modeller with a statistical reference for calibrating and interpreting CMAQ outputs.

With feature differences $\tilde{\mathbf{E}}_j^d$ and $\mathbf{P}_j^d$ modelled as Gaussian Processes, their covariates $\mathbf{X}_{\tilde{E}^d\_j}$ and $\mathbf{X}_{P^d\_j}$ are selected through the *iterative improvement* algorithm proposed in Chapter 4. This algorithm adds one-by-one, the model covariates (from a candidate set of covariates) into a GP model until no more statistically significant covariates are left. In addition to finding the most parsimonious form of model formulation, this model selection method is also designed to *rank* the model covariates in terms of their statistical importance in modelling a GP. It has shown proficiency in finding the appropriate forms of GP models for ozone EOFs and PCs (Sections 4.5 to 4.7). In this AQM evaluation, candidate covariate sets for $\mathbf{X}_{\tilde{E}^d\_j}$ and $\mathbf{X}_{P^d\_j}$ are respectively, the temporal and spatial means of CMAQ model variables: NOx and VOC emission rates and antecedent concentrations, temperature, wind speed and planetary boundary layer height.

This iterative improvement procedure can also be viewed as significance tests of the statistical associations between individual AQM model inputs and feature differences.

5.1.2 PCA of CMAQ Outputs and Observation Data

The PCA is implemented individually for CMAQ output and observation data, then $\mathbf{E}_j$ and $\mathbf{P}_j$ of the same order between CMAQ and observations are compared. Here, CMAQ-WRF-SMOKE outputs for the 5 episodes (Table 2.1) are interpolated onto $n = n_{obs}$ locations, where n_{obs} is the number of LFV observation locations available for a given year. The interpolation is done for both the ozone data and model variable data: temperature, wind speed, boundary layer height, NOx and VOC emission rates and ambient

concentration. This way, the original computer model outputs will be placed on a comparable space and time domain as the physical observation. Readers may refer back to Section 2.2 for the way data are processed for evaluation use. Furthermore, for both CMAQ and observation of all episodes, I used the data from the middle 3 full days (complete diurnal cycles) of the episode. When I use the term "episode mean", I am referring to the ozone averaged over this 3 day period.

Table 5.1 shows the portion of data variation explained by the first 8 ozone features of $\mathbf{O}^c_{t\times n}$ and $\mathbf{O}^o_{t\times n}$. As the table shows, the underlying mean structure ($\mathbf{E}_1$) dominates the data variation, and the amount of variation explained by features of $j \geq 2$ orders decrease rapidly to ≈ 0 from $j = 5$ onward. CMAQ and observed features of the same order explain similar proportions of their data variations.

	The order j of $\mathbf{E}_j$ and $\mathbf{P}_j$						
Episode	1	2	3	4	5	...	8
1985 CMAQ	0.94	0.03	0.02	0.01	0.00	...	< 0.00
1985 Obs.	0.94	0.02	0.01	0.01	0.00	...	< 0.00
1995 CMAQ	0.94	0.03	0.02	0.01	0.00	...	< 0.00
1995 Obs.	0.94	0.02	0.01	0.01	0.00	...	< 0.00
1998 CMAQ	0.94	0.03	0.02	0.01	0.00	...	< 0.00
1998 Obs.	0.95	0.02	0.01	0.01	0.00	...	< 0.00
2001 CMAQ	0.93	0.03	0.01	0.01	0.00	...	< 0.00
2001 Obs.	0.94	0.02	0.01	0.01	0.01	...	< 0.00
2006 CMAQ	0.96	0.02	0.01	0.01	0.00	...	< 0.00
2006 Obs.	0.93	0.02	0.01	0.01	0.01	...	< 0.00

Table 5.1: Proportion of data variation explained by ozone features of orders $j = 1, 2, 3, 4, 5, \ldots, 8$.

Table 5.2 shows the RMSEs from data reconstruction $\sum_{j=1}^{p} \mathbf{P}_j \mathbf{E}_j^{\mathrm{T}}$ for increasing value of p. These results show from a prediction perspective the amount of data variation that can be recovered by successive ozone features. Here, the improvement in RMSE gradually decreases between $p = 2$ to $p = 4$, and from $p = 4$ onward the improvement in RMSE becomes ≤ 1 *ppb* for both CMAQ outputs and observations of all episodes. At $p = 1$ and $p = 2$,

the RMSE of observation is smaller than CMAQ for 3 out of 5 episodes. However, starting from $p = 4$, the reconstruction RMSEs of CMAQ become smaller than the RMSEs of observation for all episodes except for 1995.

	The number p used for data reconstruction						
Episode	1	2	3	4	5	...	8
1985 CMAQ	8.62	6.07	4.13	2.76	2.01	...	0.60
1985 Obs.	9.84	7.73	6.02	4.94	4.12	...	2.10
1995 CMAQ	8.88	6.62	4.56	3.73	3.14	...	1.53
1995 Obs.	5.34	4.21	3.39	3.01	2.61	...	1.77
1998 CMAQ	10.14	7.71	5.51	3.79	3.14	...	1.87
1998 Obs.	7.55	5.98	4.86	4.19	3.68	...	2.54
2001 CMAQ	9.42	6.59	5.10	3.92	3.29	...	1.95
2001 Obs.	7.53	6.41	5.45	4.78	4.21	...	2.91
2006 CMAQ	6.59	4.70	3.77	2.92	1.95	...	1.16
2006 Obs.	7.01	5.87	4.89	4.09	3.49	...	2.27

Table 5.2: Data reconstruction RMSE at $p = 1, 2, 3, 4, 5, \ldots, 8$. The units are *ppb*.

As discussed in Chapter 3, each $\mathbf{E}_j$ has an associated eigenvalue λ_j. If λ_j is not statistically significantly different from λ_{j+1} or even high-ordered eigenvalues, then $\mathbf{E}_j$ forms a degeneracy set with higher-order EOF(s). The consequence is that these "degenerate EOFs" may suffer mixing of data feature/patterns, making them difficult to analyze (North et al., 1982; Monahan et al., 2009) and the order of these EOFs may also be arbitrary (Cohn and Dennis, 1994; Hannachi et al., 2007).

This question of "ozone feature separability" is important for feature-based CMAQ evaluation. Suppose $\mathbf{E}_2^c$ from CMAQ is clearly distinguishable from the rest, but the observed $\mathbf{E}_2^o$ and $\mathbf{E}_3^o$ form a degeneracy set, then it is not clear whether $\mathbf{E}_2^c$ should be compared to $\mathbf{E}_2^o$ or $\mathbf{E}_3^o$. This matter is made worse by the orthogonality constraint of EOFs, because any mismatch between $\mathbf{E}_j^c$ and $\mathbf{E}_j^o$ may be carried-over onto higher-order features.

The eigenspectrum (North et al., 1982) is often used to assess the level of EOF degeneracy of a given dataset. The idea and methodology of eigenspectrum are discussed in Section 3.2 and implemented in Section 3.3 to assess

the orders of feature degeneracy of LFV ozone. For CMAQ evaluation, I also produced eigenspectra for the $n = n_{obs}$ interpolated CMAQ output and observed ozone data. The results are summarized in Table 5.3. As shown, ozone feature separability can be categorized into two groups: one that defines the 1985, 1995 and 1998 episodes, one that defines the 2001 and 2006 episodes.

Episode (wind regime)	Orders of ozone feature separability
All episodes	$\mathbf{E}_1$ is separable from higher-order features.
1985 (I-IV-IV)	For both CMAQ and observations: features of orders $j = 2, 3$ form a couplet, and the feature of order $j = 4$ is separable from the rest.
1995 (III-III-III)	The same as 1985.
1998 (II-III-II)	The same as 1985.
2001 (II-II-II)	$j = 3, 4$ features form a couplet for CMAQ, feature inseparability starts at $j = 2$ for observations.
2006 (I-I-III)	The same as 2001.

Table 5.3: The types of ozone features separability of both CMAQ and observations of all episodes, the parentheses shows the wind regime type(s) of each episode. The conclusions are drawn based on eigenspectra obtained from the PCA of individual episodes.

5.1.3 Evaluation Strategy

It was concluded in Section 3.3 that ozone features of order $j \leq 4$ should be the focus of analysis. In Chapter 4, I further built ozone feature models for $\mathbf{E}_j$ and $\mathbf{P}_j$ of orders $j = 1, 2, 3$. There, the $j = 4$ feature is not modelled due to degeneracy with $j > 4$ features. The modelling and forecasting exercises in Sections 4.6 and 4.7 showed that by modelling only the first 3 spatial and temporal ozone features, one can closely model a complex space-time ozone process.

Considering the aforementioned results, the CMAQ evaluation will be focused on $\mathbf{E}_j$ and $\mathbf{P}_j$ at $j = 1, 2, 3$. Based on Table 5.3, the proposed feature-to-feature evaluation between CMAQ and observation is implemented in the

following individual analyses:

- Both original and $\sqrt{\lambda_1}$-scaled $\mathbf{E}_1$ are compared directly between CMAQ and observations for all episodes. The 1st-order *spatial ozone feature difference* $\mathbf{E}_1^d = \mathbf{E}_1^c - \mathbf{E}_1^o$ and $\tilde{\mathbf{E}}_1^d = \mathbf{E}_1^c\sqrt{\lambda_1^c} - \mathbf{E}_1^o\sqrt{\lambda_1^o}$ will be modelled as Gaussian Processes driven by CMAQ input conditions.

- The same evaluation will be done for 1st-order temporal feature $\mathbf{P}_1$.

- For 1985, 1995 and 1998 episodes, the 2nd and 3rd-order features will be compared *jointly*. The methods in Krzanowski (1979) and Cohn and Dennis (1994) will be used to calculate the "joint distance measures" between CMAQ modelled and observed features. I will also compare the joint spatial-temporal features $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}} + \mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ to examine any difference in their ozone advection patterns and other underlying dynamic processes.

- For the two most recent episodes during 2001 and 2006, I will refrain from feature-to-feature comparison at orders $j \geq 2$.

5.1.4 Discussion of Evaluation Methods

The combined results in this section have revealed that regional ozone field of LFV is simple enough to be dominated by one leading ozone feature, and have shown high levels of ozone feature inseparability that makes higher-order feature-based CMAQ evaluation difficult. However, this should not detract from the main purpose of this chapter as well as Chapter 6, which is the development of AQM evaluation methods. Furthermore, if an ozone process is simple enough to be dominated by one leading feature, then one may say that the CMAQ evaluation based on this feature alone is a near complete evaluation of CMAQ.

In the following sections, ozone feature comparison statistics will be shown for all episodes. A more detailed ozone feature comparison and modelling of feature differences will be focused on either 1995 or 2001: one episode from each type of feature degeneracy shown in Table 5.3.

The CMAQ evaluations in Chapters 5 and 6 are done individually for each episode, an alternative is to use combined CMAQ outputs and observations from all available episodes. Due to the highly consistent nature of LFV's dominant features (Section 3.4), *per-episode* analysis still allows for a systematic evaluation of CMAQ features against the observed features. Per-episode evaluation further has the potential to uncover episode specific deficiencies in the CMAQ system: each CMAQ run is done under episode specific atmospheric conditions, emission levels and spatial patterns (Steyn et al., 2013).

5.2 Comparison of the Mean Fields, $\tilde{\mathbf{E}}_1$ and $\mathbf{E}_1$

Figures 5.2 and 5.3 show the the spatial plots of temporal ozone means (the mean fields) from CMAQ output and observation data. The spatially continuous plots are created by applying a cubic-spline smoothing that interpolates the irregular spatial data into a smooth spatial field within the longitude-latitude bound of the n_{obs} locations. The between-episode difference in spatial domain is due to the different locations of available observations. The colour scale is set to be the same for each episode, because the focus of comparison here is between CMAQ output and observations from the same episode, not across episodes. As Figures 5.2 and 5.3 shown, with the exception of 1985, the episode means (ozone averaged across time) of CMAQ are near uniformly higher than observed throughout LFV.

5.2.1 General Features of $\tilde{\mathbf{E}}_1^d$ and $\mathbf{E}_1^d$

Figure 5.4 compares $\mathbf{E}_1^c$ and $\mathbf{E}_1^o$ for the 2001 episode, where the bottom plot shows the spatial feature difference $\mathbf{E}_1^d = \mathbf{E}_1^o - \mathbf{E}_1^c$. Figure 5.5 shows $\mathbf{E}_1^d$ from the other 4 episodes. As shown for all episodes, the observed $\mathbf{E}_1$ varies over a wider range of values than the CMAQ feature: higher maximum in the east and lower minimum in the west. However, both features exhibited similar patterns of east-west variation. These results imply that CMAQ modelling is able to capture similar spatial variation of ozone means as the physical

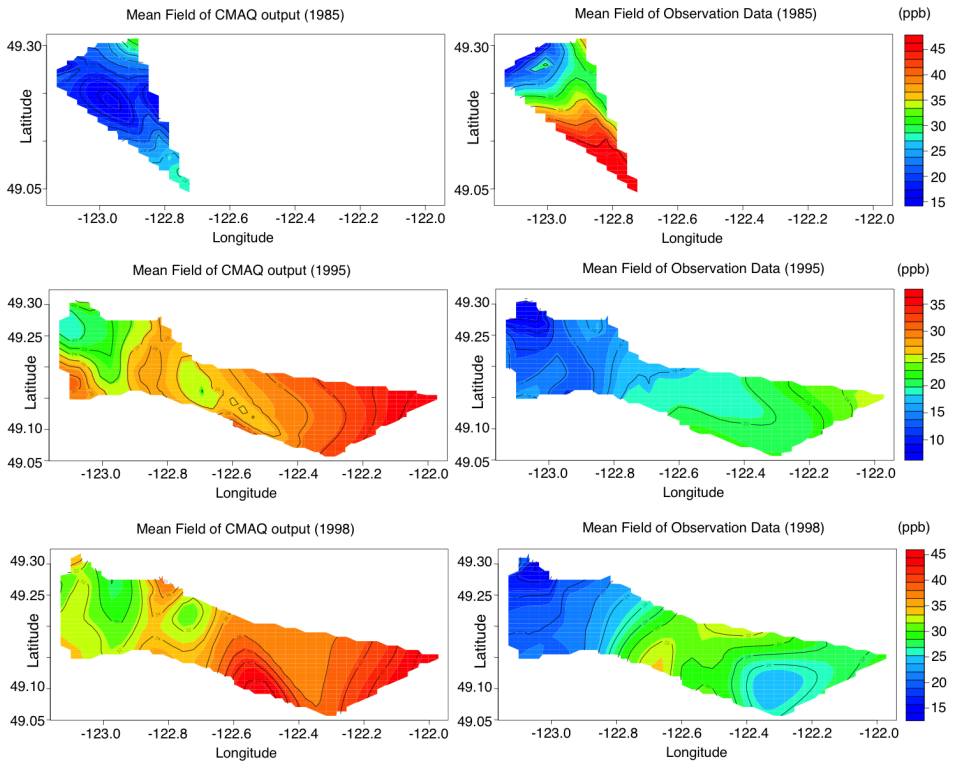


Figure 5.2: Spatial plots of temporal ozone means (the mean fields) of CMAQ outputs and observation data of the 1985, 1995 and 1998 episodes. For the same episode, the colour scale is the same in order to aid comparison.

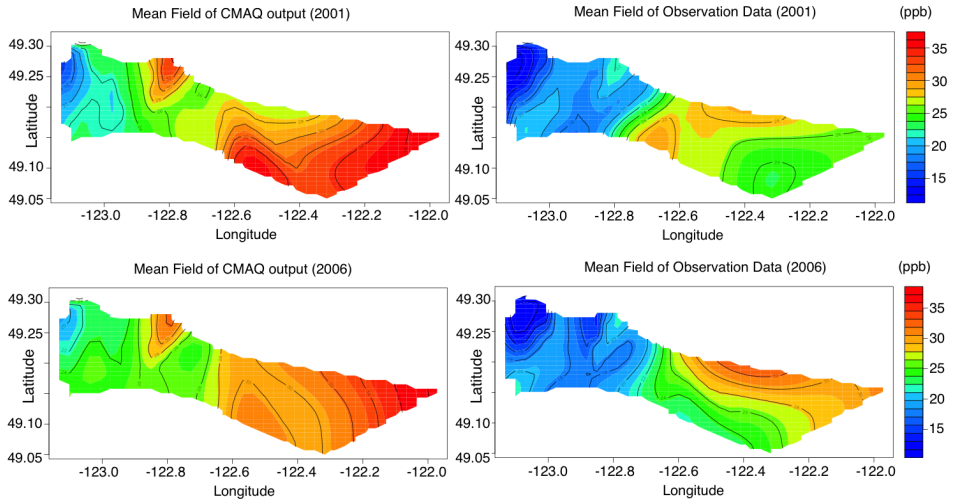


Figure 5.3: Spatial plots of temporal ozone means (the mean fields) of CMAQ outputs and observation data of the 2001 and 2006 episodes. For the same episode, the colour scale is the same in order to aid comparison.

observation. However, compared to CMAQ modelled ozone, the observed ozone process is governed by a more pronounced space-time variation: $\mathbf{E}_1^o$ has a larger range and variance than $\mathbf{E}_1^c$.

The scaled-EOF $\tilde{\mathbf{E}} = \mathbf{E}_1\sqrt{\lambda_1}$ captures the spatial variation of the mean field while taking the magnitude of $\mathbf{O}_{t\times n}$ into account. Figure 5.6 compares $\tilde{\mathbf{E}}_1^c$ and $\tilde{\mathbf{E}}_1^o$ from 2001, and Figure 5.7 shows the $\tilde{\mathbf{E}}_1^d$ of all other episodes. As shown, the comparison of $\tilde{\mathbf{E}}_1$ is analogous to the comparison of mean fields calculated from the data (Figures 5.2 and 5.3). Hence, the comparison between $\tilde{\mathbf{E}}_1^c$ and $\tilde{\mathbf{E}}_1^o$ is a means to evaluate CMAQ's capability in capturing not only the spatial variations of temporal ozone means in LFV, but also their magnitudes. This is a spatial ozone feature evaluation using data information summarized across hours of the episode.

Figures 5.6 and 5.7 show that the $\tilde{\mathbf{E}}_1^d$ values are all negative for the 1995, 1998 and 2001 episodes, nearly all-negative on 2006, and mostly positive for 1985. Hence, when evaluated against observations, CMAQ almost systematically over-estimated the temporal ozone means throughout the triangular LFV region. In other words, the mean fields produced by CMAQ modelling

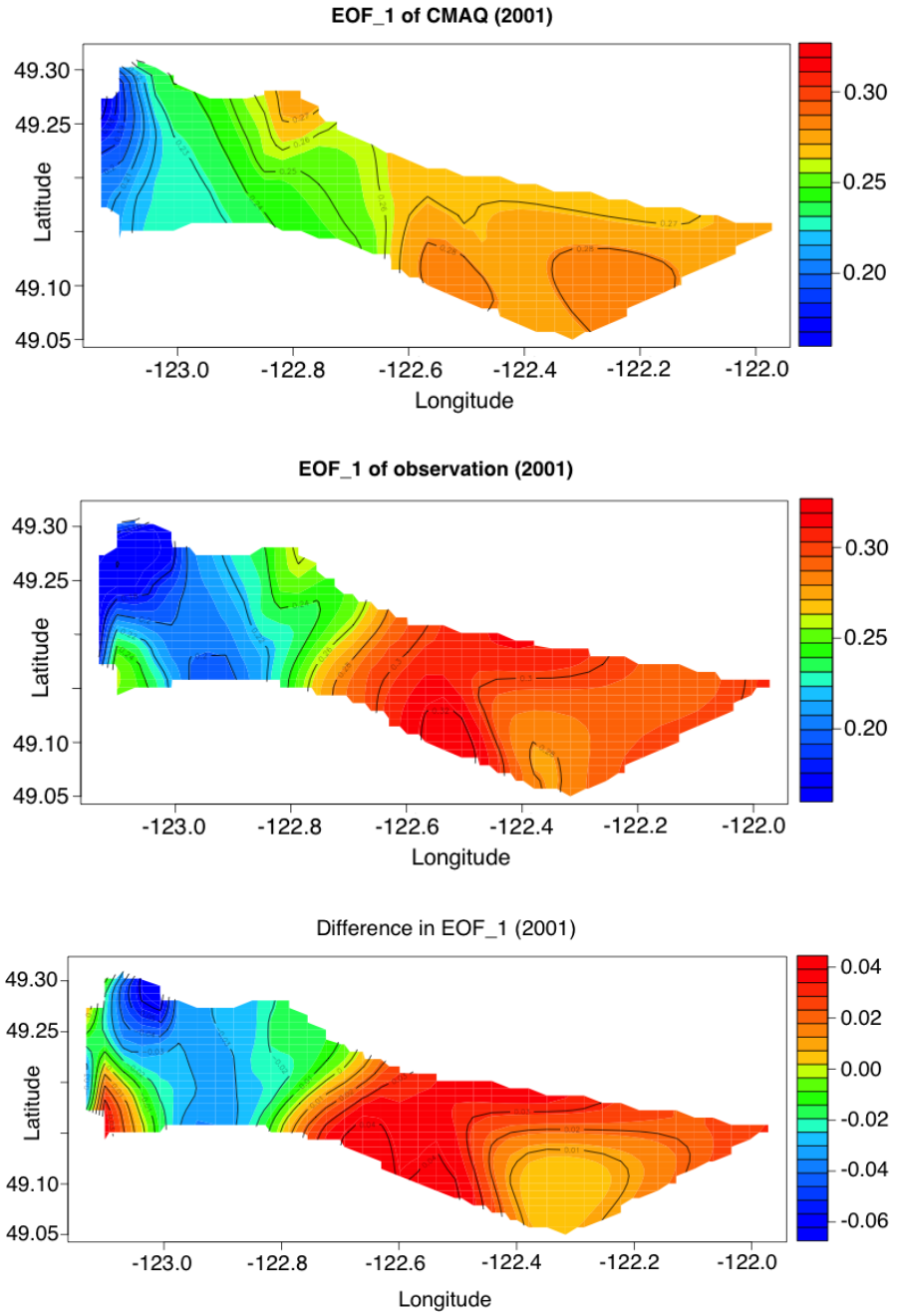


Figure 5.4: For the 2001 episode: plots of $\mathbf{E}_1^c$ (top), $\mathbf{E}_1^o$ (middle) and $\mathbf{E}_1^d$ (bottom).

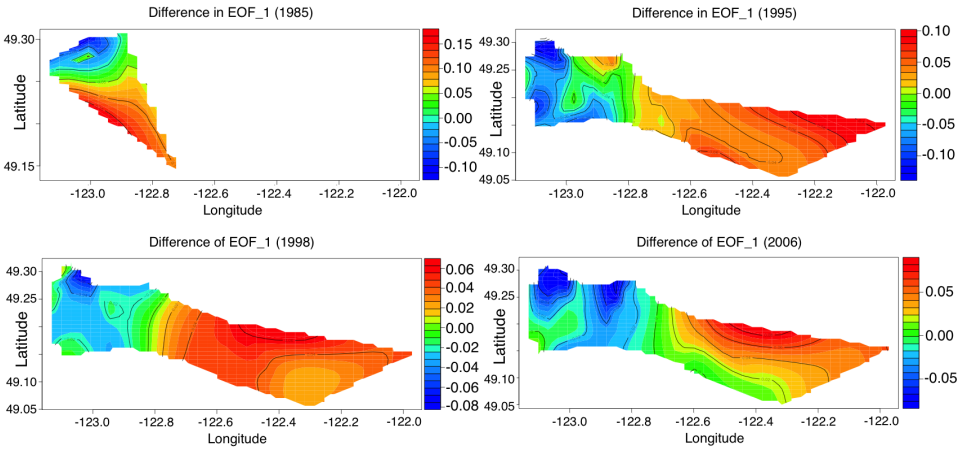


Figure 5.5: Plots of $\mathbf{E}_1^d$ of from the 1985, 1995, 1998 and 2006 episodes.

tend to have uniformly higher values. Moreover, the magnitude of CMAQ over-estimate is more pronounced in the west around the city of Vancouver and its suburbs than the eastern LFV.

Table 5.4 shows the angles between $\tilde{\mathbf{E}}_1^c$ and $\tilde{\mathbf{E}}_1^o$ for all 5 episodes. The angle between two $\mathbf{E}_j$ vectors is a distance measure between the CMAQ modelled feature and the corresponding observed feature. Cohn and Dennis (1994) used vector angle to quantify the closeness between the EOFs of acid deposition model outputs and observations. The authors regarded angles $\leq 15^\circ$ as reasonably low, which indicates good observation-model agreement. The values in table 5.4 show that the $\mathbf{E}_1$ angles are $8 - 11^\circ$ for all episodes. Hence, despite just discussed CMAQ over-estimates, there is consistent close correspondence between the $\tilde{\mathbf{E}}_1$'s (the mean fields) of CMAQ and observations.

	Episode				
	1985	1995	1998	2001	2006
Angle between $\tilde{\mathbf{E}}_1$'s:	10.73°	10.21°	9.01°	7.98°	10.29°

Table 5.4: Angles between $\tilde{\mathbf{E}}_1^c$ and $\tilde{\mathbf{E}}_1^o$.

In summary, the CMAQ tends to over-estimate the observed episode

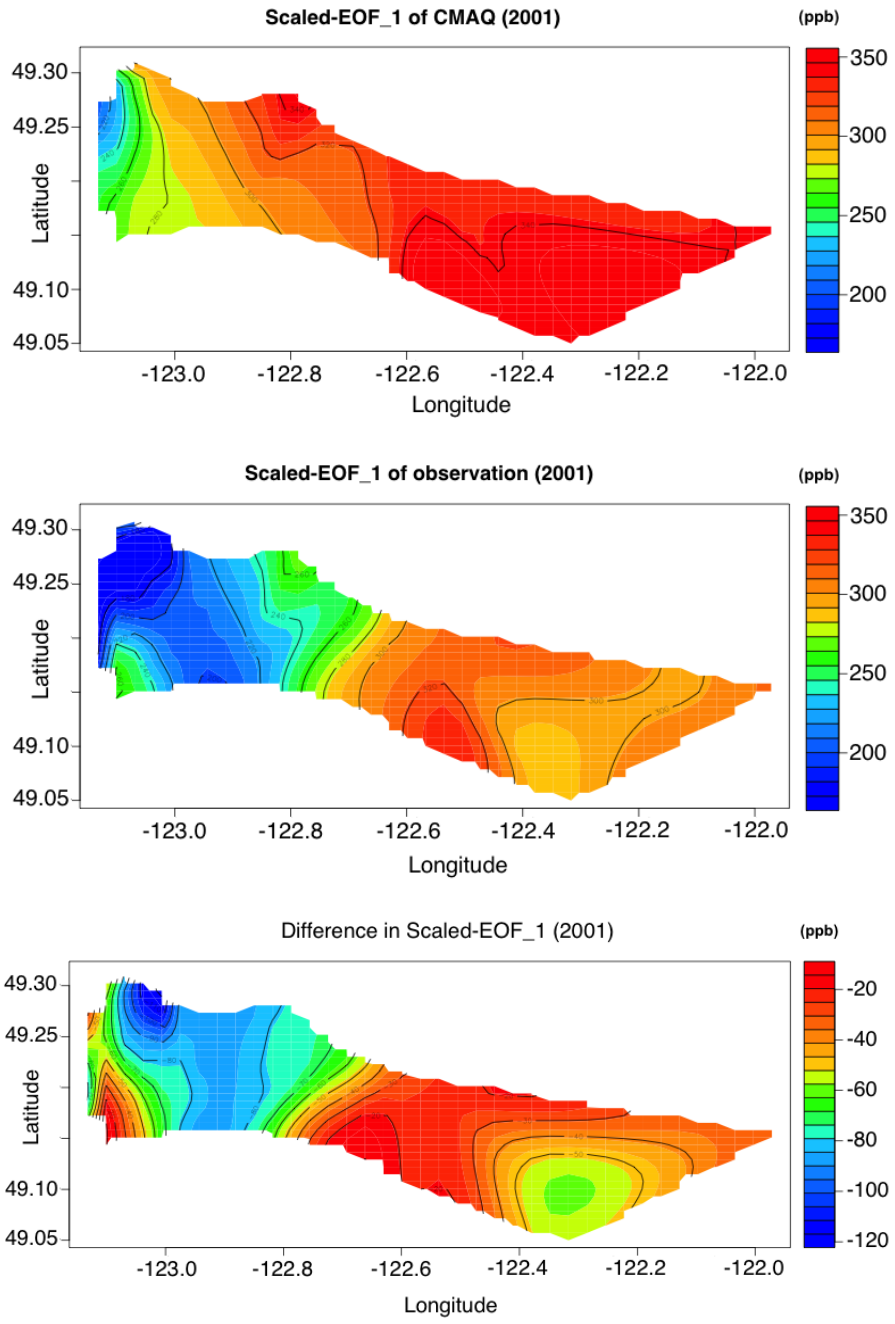


Figure 5.6: For the 2001 episode: plots of $\tilde{\mathbf{E}}_1^c$ (top), $\tilde{\mathbf{E}}_1^o$ (middle) and $\tilde{\mathbf{E}}_1^d$ (bottom).

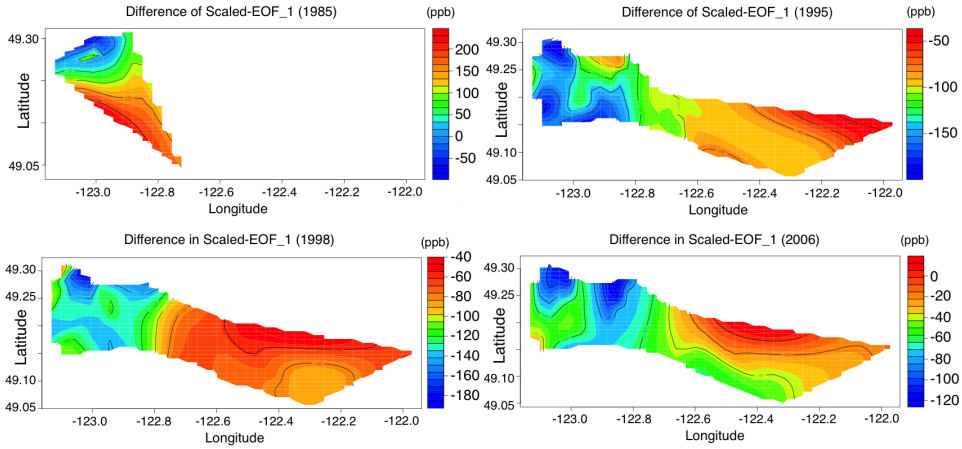


Figure 5.7: Plots of $\tilde{\mathbf{E}}_1^d$ of from the 1985, 1995, 1998 and 2006 episodes.

means across LFV, and this difference is captured by $\tilde{\mathbf{E}}_1$ for each episode. In addition, $\tilde{\mathbf{E}}_1$ has well defined spatial structures that are consistent between most of the episodes evaluated.

5.2.2 Covariate Selection for $\tilde{\mathbf{E}}_1^d$: the Difference in the Mean Fields

A follow-up to the preceding ozone feature comparison is to analyze the background factors that influence the feature differences between CMAQ modelled ozone and physical observation. These "background factors" are the model covariates of $\tilde{\mathbf{E}}_1^d$ selected from a set of candidate model covariates using iterative improvement algorithm (Section 5.1). The covariate set $\mathbf{X}_{\mathbf{E}}\_j$ = (longitude, latitude) is used to *start* the algorithm. Given the random observation error in (5.1), the optimization of GP functions will explicitly include a *nugget* term that captures the stochastic variation at fixed point location or time.

Table 5.5 shows the selected model covariates *in addition* to longitude and latitude. The results reveal that for 1985, 2001 and 2006 CMAQ modelling runs, the 1st-order observation-CMAQ feature difference is influenced solely by the mean VOC emission rates of the episodes, i.e., the spatial fields

of VOC emission rates averaged across time. The 1995 and 1998 feature differences are statistically associated with either the episode mean of NOx emission or antecedent NOx concentrations.

	Episode year				
	1985	1995	1998	2001	2006
$\mathbf{X}_{\tilde{E}^d\_1}$	VOC	NOx-lag	NOx	VOC	VOC

Table 5.5: Result of covariate selection for $\tilde{\mathbf{E}}_1^d$. The listed covariates are those in addition to longitude and latitude. The notation "-lag" represent the atmospheric (antecedent/lagged) concentration of the precursor (notation defined in Section 4.5.)

These model selection results indicate that the deviations in $\tilde{\mathbf{E}}_1$, or the mean fields, are heavily influenced by the spatial distributions of mean precursor emission rates or antecedent concentration. On the other hand, no meteorological variable was determined to be statistically significant through iterative likelihood testing.

SMOKE/CMAQ modelling deficiencies associated with emission inputs and chemical reaction modelling is identified by my proposed method of CMAQ evaluation for all episodes. As mentioned in Chapter 2, Steyn et al. (2011) and Steyn et al. (2013) described the efforts that went into producing the CMAQ-WRF-SMOKE data used in this thesis. The papers outlined the methods of estimating the space-time distributions of NOx and VOC emission across LFV, especially the way of estimating the year-specific spatial shift in emission sources. The task of estimating localized emission patterns is a difficult one, this is noticed from the descriptions in aforementioned papers as well as the the complexities of SMOKE operations in general (overview in Section 1.1). Moreover, detailed space-time emission is unobservable (unlike the weather), thus CMAQ users are unable to tune SMOKE outputs against observations.

In addition to CMAQ input uncertainties, there are further uncertainties when modelling the atmospheric chemical precursor concentrations. Uncertainties in the chemical model within CMAQ come from the fact that one

is typically unable to know and model every chemical reaction occurring. Rather, reactions involving one molecule serve as a proxy model for reactions of similar molecules (Finlayson-Pitts and Pitts Jr, 1999), thus modelling deficiencies unavoidably follow.

5.2.3 Detailed Analyses of $\tilde{\mathbf{E}}_1^d$ vs. VOC Emission for the 2001 Episode

In this section, I will evaluate the 2001 ozone episode by analyzing the sensitivity of $\tilde{\mathbf{E}}_1^d$ to the spatial variation of episode-mean VOC emission rate. This analysis will show over the course of an episode, how the mean VOC emission influences the difference between the mean fields of CMAQ and observations. The 2001 data are used to demonstrate that, by modelling $\tilde{\mathbf{E}}_1^d$ one can extract insightful information regarding CMAQ's modelling deficiency. Similar analysis can be done on any other ozone features from other episodes.

An estimated sensitivity or univariate effect plot of $\tilde{\mathbf{E}}_1^d$ against VOC emission rate is produced using the method described in Schonlau and Welch (2006). I first fitted the GP model of $\tilde{\mathbf{E}}_1^d$ using CMAQ-SMOKE outputs of 2001, where the covariates are longitude, latitude and the temporal mean VOC emission rates. I then produced $\tilde{\mathbf{E}}_1^d$ outputs at a range of VOC emission rates while integrating out the two location variables from the GP model. Thus, the *univariate* effect of mean VOC emission on $\tilde{\mathbf{E}}_1^d$ can be analyzed.

Figure 5.8 shows the estimated univariate effect of $\tilde{\mathbf{E}}_1^d = \tilde{\mathbf{E}}^o - \tilde{\mathbf{E}}^c$ against a range of mean VOC emission rate. The dots are $\tilde{\mathbf{E}}_1^d$ outputs over VOC emission rates of $0.2 - 1.9$ moles$\cdot$sec^{-1}. The dashed-lines are the 95% confidence interval calculated using the analytical expression derived in Schonlau and Welch (2006). The training data (2001 SMOKE output) have temporal mean VOC emission varying between $0.5 - 1.0$ moles$\cdot$sec^{-1} and one data point at 1.8 moles$\cdot$sec^{-1}. This distribution of values partially explains the large standard error (wide confidence interval) associated with $\tilde{\mathbf{E}}_1^d$ outputs between $1.3 - 1.8$ moles$\cdot$sec^{-1}.

The confidence interval indicates the statistical significance of mean fea-

ture difference at each VOC emission rate. Given a VOC emission rate, if the confidence interval for the univariate effect on $\tilde{\mathbf{E}}_1^d$ contains 0, then we fail to reject at significance level of 5% the hypothesis that the mean feature difference is 0. In other words, if the confidence interval is above or below the line $\tilde{\mathbf{E}}_1^d = 0$ at a given VOC emission, then we conclude that the estimated univariate effect of VOC emission on $\tilde{\mathbf{E}}_1^d$ is statistically significant.

We see from the sensitivity plot that:

- There is a negative trough at VOC = 0.82 moles·sec^{-1}, i.e., CMAQ over-estimate of temporal ozone mean in space. The confidence interval is below 0, indicating the statistical significance of this ozone feature difference.

- There is a positive peak at VOC = 1.18 moles·sec^{-1}. However, this peak value is predicted by the $\tilde{\mathbf{E}}_1^d$ model with lower confidence interval at near 0. Thus, the statistical significance of ozone feature difference maybe in question. As mentioned, a positive $\tilde{\mathbf{E}}_1^d$ indicates a CMAQ under-estimate of observed feature.

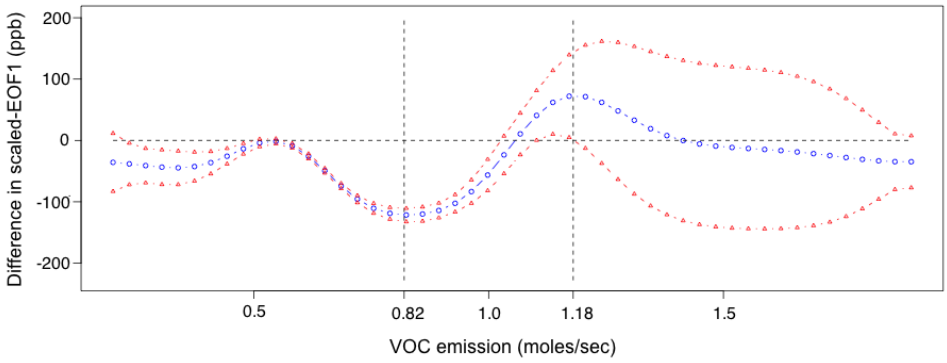


Figure 5.8: Sensitivity or univariate effect plot of $\tilde{\mathbf{E}}_1^d$ against episode mean VOC emission rate. The blue dotted line is the estimate of $\tilde{\mathbf{E}}_1^d$ averaged over locations for VOC emission rates of $0.2 - 1.9$ moles·sec^{-1}, and the red triangle lines are point-wise 95% confidence intervals. The GP model of $\tilde{\mathbf{E}}_1^d$ is fitted using 2001 CMAQ-SMOKE data.

Figure 5.9 shows the spatial field of mean VOC emission (averaged across time) from the 2001 episode, where the dataset is SMOKE output. In the same figure, the map of the LFV observation network from 2001 is also provided. Table 5.6 shows the station name associated with each location number in the network map. From Figure 5.8, the largest feature difference occurs for VOC $\approx \{0.7, 1.2\}$ moles·sec^{-1}. These two values can be identified to define three areas of LFV: (1) the area north of Abbotsford and west of Chilliwack has temporal (or episode) mean VOC emission at ≈ 0.7 to 0.8 moles·sec^{-1}, (2) the suburbs of eastern Metro Vancouver (Burnaby, etc,) have temporal mean VOC emission at ≈ 0.8 moles·sec^{-1}, and (3) the small area surrounding Vancouver's city core has temporal mean VOC emission at ≈ 1.2 moles·sec^{-1}.

Number	Longitude	Latitude	Name
1	-123.16	49.26	Kitsilano
2	-123.15	49.19	YVR
3	-123.12	49.28	Robson square
4	-123.11	49.14	Richmond south
5	-123.08	49.32	Mahon park
6	-123.02	49.30	North Vancouver
7	-122.99	49.22	Burnaby south
8	-122.97	49.28	Kenshington park
9	-122.90	49.16	North Delta
10	-122.85	49.28	Rocky Point Park
11	-122.79	49.29	Coquitlam
12	-122.71	49.25	Pitt Meadows
13	-122.69	49.13	Surrey east
14	-122.58	49.22	Maple Ridge
15	-122.57	49.10	Langley central
16	-122.31	49.04	Central Abbotsford
17	-121.94	49.16	Chilliwack

Table 5.6: Station names and coordinates of numbers 1 to 17 in Figure 5.9: the map of the 2001 LFV monitoring network.

A high observation-CMAQ feature difference is mostly associated with VOC emissions at the aforementioned three areas of LFV. Two areas of

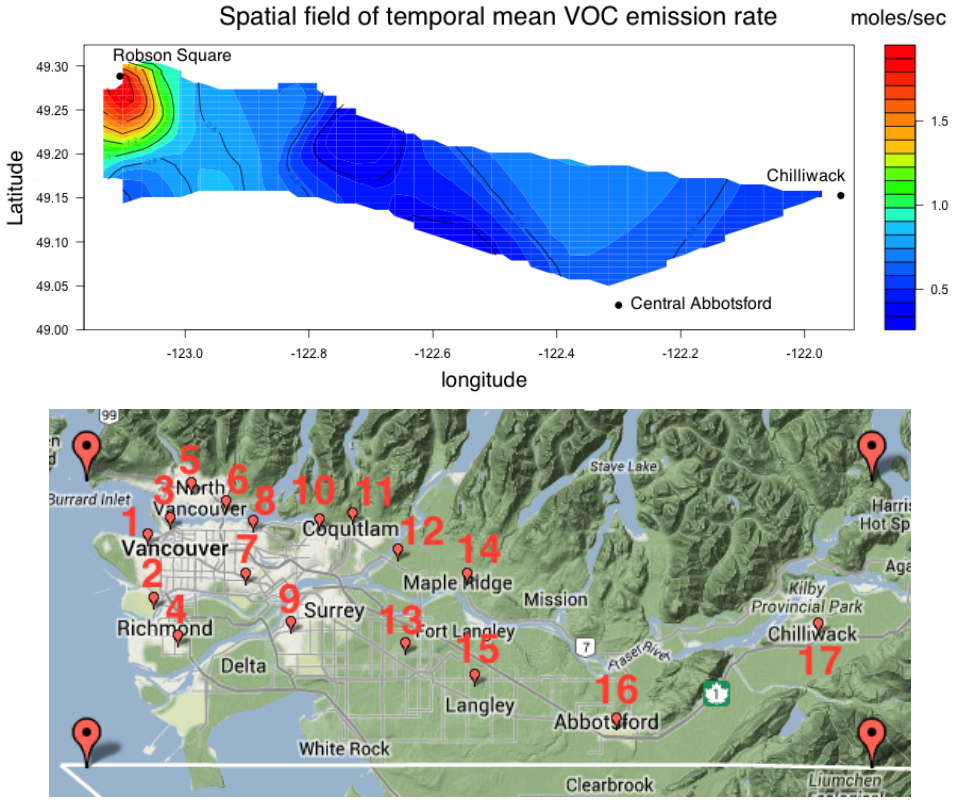


Figure 5.9: For the 2001 episode: the spatial plot of episode mean VOC emission rate within the LFV region defined by $n_{obs} = 17$ monitoring sites (top), and the map of the LFV monitoring network (bottom). The station names and coordinates associated with the numbers 1 to 17 are in Table 5.6.

interest, the suburbs of Metro Vancouver and locations around Vancouver's city core, are also areas where daily ozone plume forms (Section 3.1). Hence, evidence suggests that the CMAQ over-estimation of the observed ozone field is attributable to the production of a higher-than-observed initial ozone plume by CMAQ.

Furthermore, Steyn et al. (2011) and Steyn et al. (2013) mentioned that the city of Vancouver is a "VOC sensitive" region: NOx is the dominating ozone precursor and its concentration is near saturation, hence any variation in VOC causes a noticeable change in O_3 concentrations. On the other hand, the eastern LFV (Abbotsford and Chilliwack) are "NOx sensitive" area, i.e., high concentration of VOC, making O_3 pollution sensitive to variation in NOx. Combined results from preceding analyses indicate that the eastern Metro Vancouver, where the ozone process begin to transition from VOC to NOx sensitive, is the area of interest: the temporal mean VOC emission produced by SMOKE for this region showed strong statistical association with the observation-CMAQ difference in their mean ozone fields.

The univariate plot in Figure 5.8 is obtained by averaging out the effect of location. To uncover any bivariate or interaction effect of location *and* VOC emission on $\tilde{\mathbf{E}}_1^d$, one can produce "$\tilde{\mathbf{E}}_1^d$ versus VOC emission" plots at multiple locations across LFV. In each plot, the location covariates in the $\tilde{\mathbf{E}}_1^d$ model are fixed at a longitude-latitude setting, and $\tilde{\mathbf{E}}_1^d$ is estimated over a range of VOC emission rates appropriate for this location.

Figure 5.10 shows the $\tilde{\mathbf{E}}_1^d$ versus VOC emission plot at 5 locations across LFV. As shown, for locations across LFV, a negative trough at VOC ≈ 0.82 moles·sec^{-1} is a common feature. This result is representative of other LFV locations not shown. Figure 5.10 shows that there is little, if any interaction effect of VOC and location on CMAQ over-estimate (negative $\tilde{\mathbf{E}}_1^d$) of the episode mean. The over-estimate is most noticeable when the mean VOC emission is around 0.82 moles·sec^{-1}, and this is a feature of CMAQ modelling deficiency that is common across LFV locations.

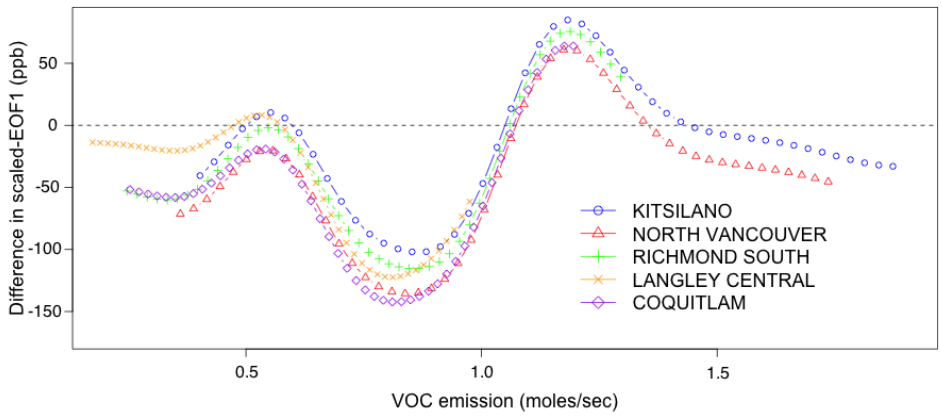


Figure 5.10: Sensitivity or univariate effect plot of $\tilde{\mathbf{E}}_1^d$ against VOC emission rate at 5 LFV locations. The GP model of $\tilde{\mathbf{E}}_1^d$ is fitted using 2001 CMAQ-SMOKE data.

5.3 Comparison of $\mathbf{P}_1$: Hourly LFV Mean Ozone

Figure 5.11 compares $\mathbf{P}_1^c$ and $\mathbf{P}_1^o$ for ozone episodes 1985, 1995 and 1998, while Figure 5.12 does the same for 2001 and 2006. Figures 5.13 and 5.14 show the time series of LFV mean ozone of CMAQ output and observations data. Comparison with $\mathbf{P}_1^c$ and $\mathbf{P}_1^o$ time series reveals that the 1st-order temporal features of both ozone data captured to near exact detail, the temporal patterns of their respective ozone means, and the scale of difference between two ozone means. Therefore, the comparison of $\mathbf{P}_1$ is equivalent to the comparison of hourly LFV mean ozone. It is worth repeating that the PCs are weighted row sums of $\mathbf{O}_{t\times n}$, hence the high values (in units *ppb*).

As shown in Figure 5.11, for the 1985 episode, the CMAQ modelled hourly LFV mean ozone corresponded closely with the observations. For all other 4 episodes, the pattern of observation-CMAQ differences can be defined as CMAQ over-estimate of observed hourly LFV ozone during both the early morning and afternoon peaks hours. The relatively close correspondence of the 1985 temporal features is noticeable from Table 5.7, which shows the angles between component vectors $\mathbf{P}_1^c$ and $\mathbf{P}_1^o$. The 1985 episode has slightly smaller angle than other episodes, while the 2001 episode has

the largest angle. However, the angles are low enough (Cohn and Dennis, 1994) that there is generally good agreement between the 1st-order temporal-features of CMAQ and observations.

	Episode				
	1985	1995	1998	2001	2006
Angle between $\mathbf{P}_1$'s:	9.77°	10.30°	12.09°	13.02°	10.52°

Table 5.7: Angles between $\mathbf{P}_1^c$ and $\mathbf{P}_1^o$.

5.3.1 Modelling $\mathbf{P}_1^d$

Covariate selection for $\mathbf{P}_j^d$ is initiated by a GP model with "hour of the day" as starting covariate, and the iterative improvement procedure is applied as before. The GP model optimizations are done by including the stochastic error term ε_{Pj} in (5.2).

Table 5.8 shows the statistically significant covariates associated with $\mathbf{P}_1^d$,

	Episode year		
	1985	1995	1998
$\mathbf{X}_{P^d\_1}$	Temp, Wind BL, VOC	BL, NOx-lag	Temp, Wind NOx-lag
	2001	2006	
$\mathbf{X}_{P^d\_1}$	Temp, Wind BL, NOx-lag	Temp, Wind NOx-lag	

Table 5.8: Result of covariate selection for $\mathbf{P}_1^d$. The listed covariates are those in addition to "hour of the day".

i.e, observation-CMAQ difference in hourly LFV mean. The descriptions "NOx-lag" and "VOC-lag" represent the hourly LFV means of NOx and VOC antecedent concentrations, and "BL" represent hourly LFV means of boundary layer height. The iterative improvement algorithm delivered a mixture of meteorological and chemical precursor variables. Unlike the modelling of $\tilde{\mathbf{E}}_1^d$ (Table 5.5), there are no clearly definable CMAQ inputs responsible for the difference in temporal ozone features between CMAQ

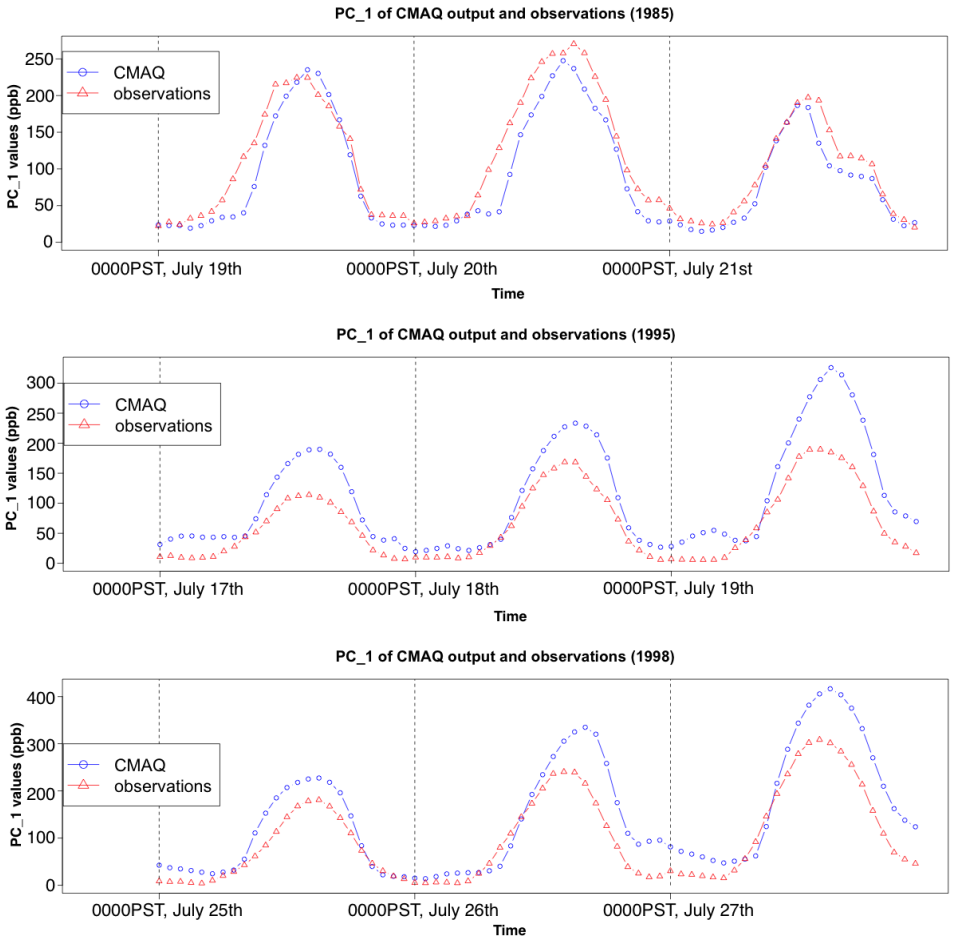


Figure 5.11: Time-series of $\mathbf{P}_1^c$ (blue) and $\mathbf{P}_1^o$ (red) for the 1985, 1995 and 1998 episodes.

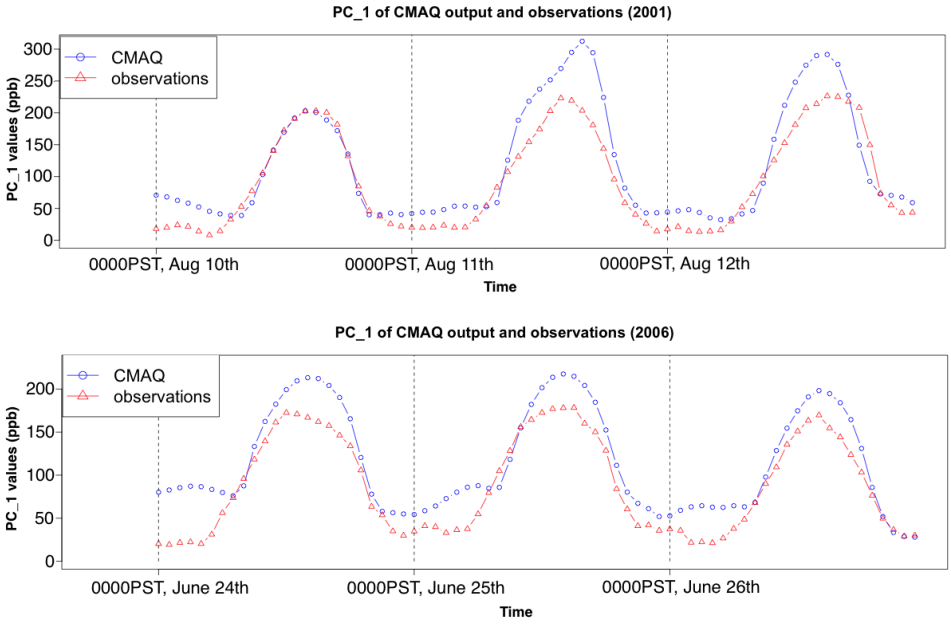


Figure 5.12: Time-series of $\mathbf{P}_1^c$ (blue) and $\mathbf{P}_1^o$ (red) for the 2001 and 2006 episodes.

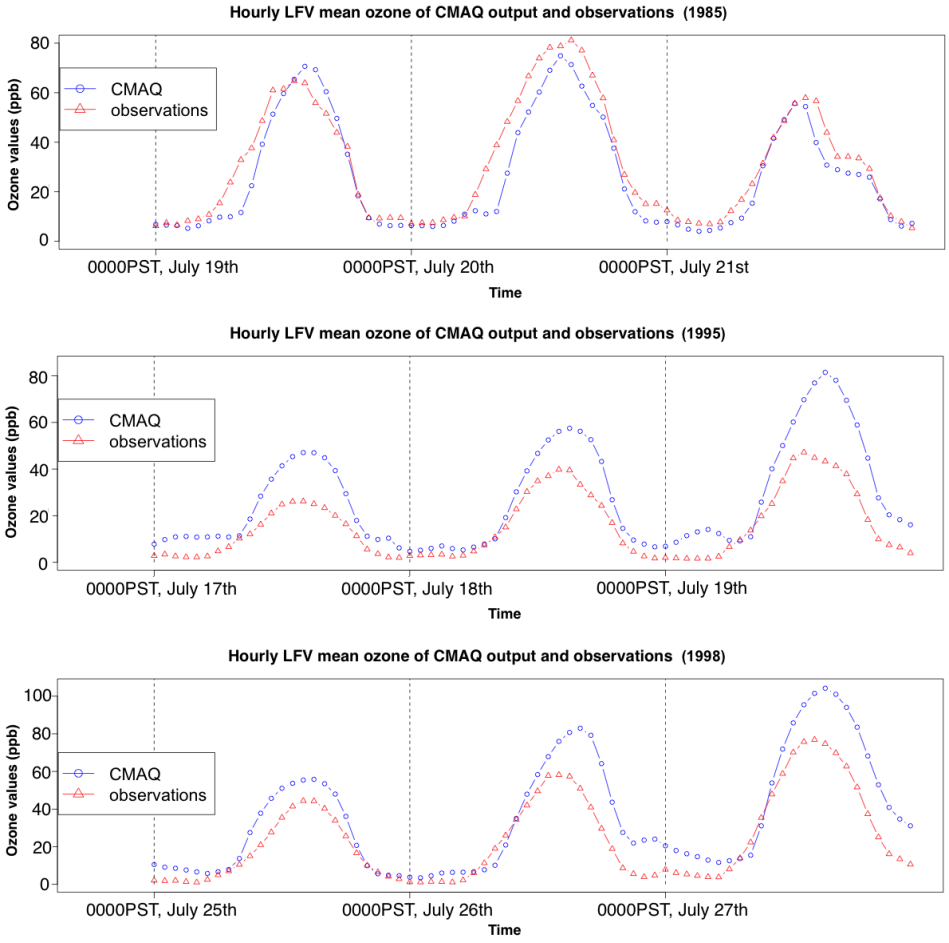


Figure 5.13: Time-series of hourly LFV mean ozone (averaged across space) of CMAQ output (blue) and observations (red) from the 1985, 1995 and 1998 episodes.

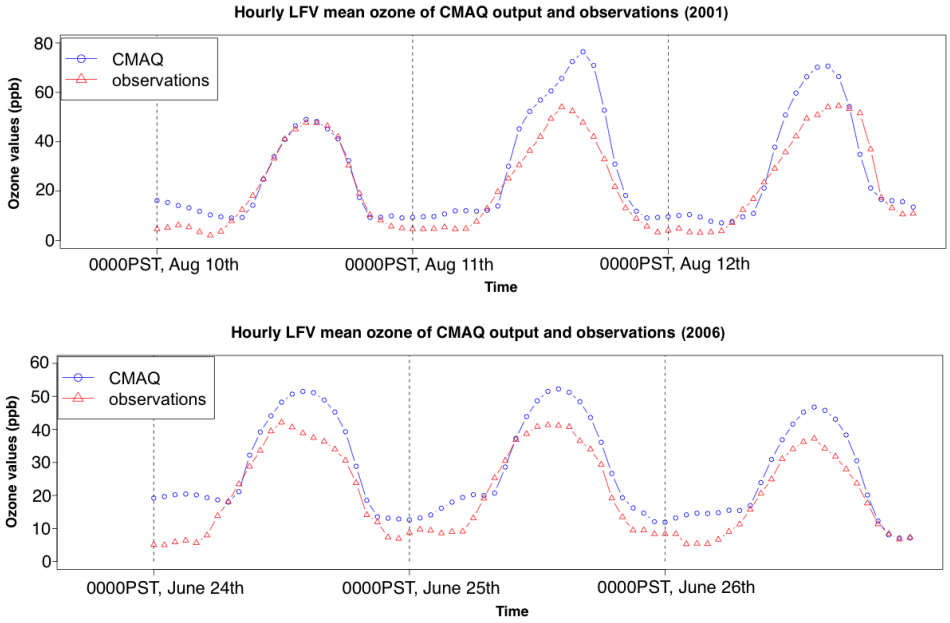


Figure 5.14: Time-series of hourly LFV mean ozone (averaged across space) of CMAQ output (blue) and observations (red) from the 2001 and 2006 episodes.

and observations. The forward selection method indeed found statistically significant covariates based on the Gaussian log-likelihood test (the selection criteria of iterative improvement, Section 4.3), but the selected covariates do not allow a clear explanation.

As the analysis in the next chapter will show, there is reason to believe that the difference in temporal ozone features is to a large degree, caused by CMAQ not modelling certain real-world ozone processes. In Section 2.2, I mentioned the phenomenon of nocturnal ozone down-mixing: during the nocturnal hours, vertical atmospheric mixing draws the upper-layer pollution downward, causing a short-term spike in the ground level ozone at some locations in the LFV. The follow-up ozone process is that ozone is consumed by NOx at the ground level, making surface-level ozone concentrations approximately 0 *ppb*. However, as Figures 5.11 and 5.12 showed, CMAQ does not seem to capture the reality of NOx-initiated ozone reduction the way observations do, and it tends to over-estimate the ozone levels between 0000PST to 0400PST. Furthermore, CMAQ and WRF do not account for the process of nocturnal ozone down-mixing.

The analysis of temporal ozone feature difference will arrive at some form of conclusion in the next chapter, where the statistical properties of ozone features are compared.

5.4 Comparison of Higher-order Features

As described in Section 5.1 and Table 5.3, for the 1985, 1995 and 1998 episodes, $\mathbf{E}_2$ and $\mathbf{E}_3$ have close enough eigenvalues that make the same-order feature comparison between CMAQ and observations questionable. Whereas for the 2001 and 2006 episodes, the degeneracy of the observed features starts from $j = 2$, which makes the feature-by-feature comparison even harder.

Krzanowski (1979) proposed a method of jointly comparing PCA components that was later applied by Cohn and Dennis (1994) to evaluate acid deposition models. Let $\mathbf{E}^c_{n\times p}$ and $\mathbf{E}^o_{n\times p}$ be a matrix with p *leading* EOFs, where $p = 3$ and $n = n_{obs}$ in this evaluation. Also consider a form of

joint-covariance matrix $\mathbf{M} = (\mathbf{E}^c)^{\mathrm{T}}\mathbf{E}^o(\mathbf{E}^o)^{\mathrm{T}}\mathbf{E}^c$ whose eigenvectors are $\mathbf{e}_j$ and eigenvalues are λ_j^e, $j = 1, 2, 3$. It was shown in Krzanowski (1979) that $\mathbf{e}_j^c = \mathbf{E}_j^c\mathbf{e}_j$ form an orthogonal basis for $\mathbf{E}^c$ and $\mathbf{e}_j^o = \mathbf{E}_j^o(\mathbf{E}_j^o)^{\mathrm{T}}\mathbf{e}_j$ is an orthogonal subspace for $\mathbf{E}^o$. Furthermore, $\mathbf{e}_1^c$ and $\mathbf{e}_1^o$ are the closest vectors between the subspaces defined by ozone features $\mathbf{E}^c$ and $\mathbf{E}^o$, and their angle is calculated as $\cos^{-1}(\sqrt{\lambda_1^e})$. The subsequent higher-order vectors are further apart with angles $\cos^{-1}(\sqrt{\lambda_j^e})$.

Table 5.9 shows for the episodes 1985, 1995 and 1998, the angles between the vectors in $\mathbf{E}_{n\times 3}^c$ and $\mathbf{E}_{n\times 3}^o$ calculated using the joint-comparison method just described. These angles indicate the observation-CMAQ difference when the leading $p = 3$ ozone features are compared jointly. As shown, the angles between $\mathbf{e}_1^c$ and $\mathbf{e}_1^o$ are smaller than 10° for all three episodes, indicating close agreement. While the angles between the 2nd vector sets are reasonably low, the angles between $\mathbf{e}_3^c$ and $\mathbf{e}_3^o$ increased significantly to near 45° for the 1998 episode. These results reveals that, when compared alone, the 1st-order ozone features have good agreement between CMAQ and observations, but when the leading 3 features are compared jointly, the good agreement quickly disappears. This implies a noticeable discordance of higher-order features.

	Angles between $\mathbf{e}_j^c$ and $\mathbf{e}_j^o$		
	j=1	j=2	j=3
Episode 1985	6.76°	10.2°	37.34°
Episode 1995	9.84°	15.0°	28.43°
Episode 1998	4.43°	15.83°	44.19°

Table 5.9: Angles from the joint comparison of the leading 3 ozone features from CMAQ output and physical measurements.

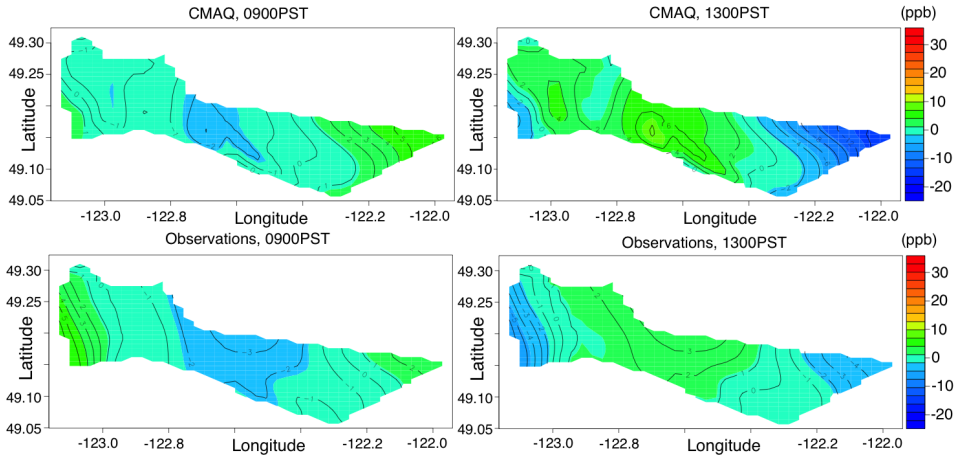
In Chapter 3, I have shown that some ozone features of orders $j \geq 2$ individually or jointly capture the dynamic patterns of ozone advection across LFV. However, difference statistics such as ones in Table 5.9 only gives one value summarizing the observation-CMAQ difference. A more informative approach is needed to compare dynamic ozone features between CMAQ and observations.

As discussed in Section 5.1, the results from eigenspectra pointed out the closeness between the 2nd and 3rd eigenvalues of both CMAQ and observations. This implies the possibility of feature degeneracy, which requires that $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ and $\mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ be analyzed jointly. Therefore, one way of performing the observation-CMAQ comparison of advection patterns is to compare the sum of ozone features $\mathbf{P}_2^c\mathbf{E}_2^{c\mathrm{T}} + \mathbf{P}_3^c\mathbf{E}_3^{c\mathrm{T}}$ between CMAQ and observation. Figures 5.15a and 5.15b compare the $\mathbf{P}_2^c\mathbf{E}_2^{c\mathrm{T}} + \mathbf{P}_3^c\mathbf{E}_3^{c\mathrm{T}}$ and $\mathbf{P}_2^o\mathbf{E}_2^{o\mathrm{T}} + \mathbf{P}_3^o\mathbf{E}_3^{o\mathrm{T}}$ at selected hours on the 3rd day of 1995. For ease of comparison all contour plots have the same range of values. In the morning hours between 0600PST-0900PST, both CMAQ and observation captured ozone contrasts that are similar both in pattern and magnitude: the contrast is slightly > 0 *ppb* in the eastern and western edge of LFV and slightly < 0 *ppb* in the middle.

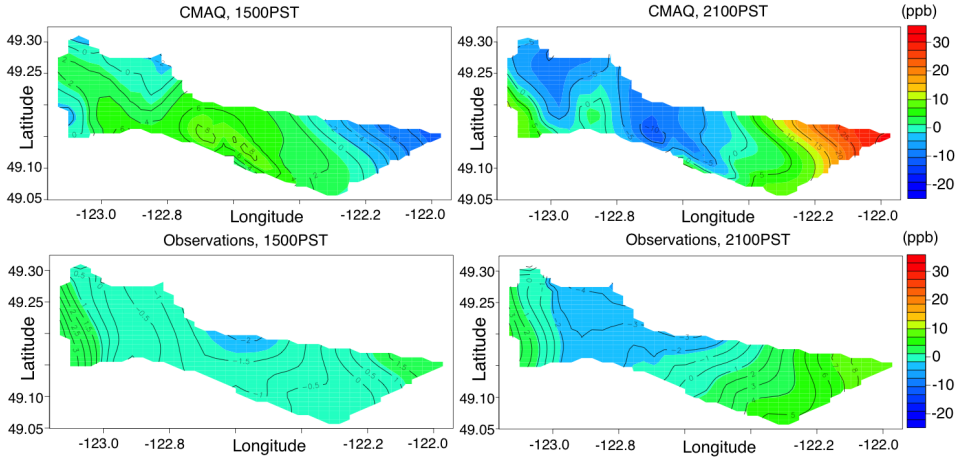
The observation-CMAQ difference emerges during the afternoon ozone peaks, where the east-west ozone contrast is more pronounced for CMAQ. At 1300PST the CMAQ ozone feature has contrast values ranging from -12 *ppb* to 6 *ppb* whereas the observation has contrast ranging from -5 *ppb* to 3 *ppb*. At 1400PST the CMAQ feature still has noticeable ozone contrast while nearly no spatial ozone contrast is noticeable in observed feature. Moreover, this contrast pattern of "positive in the west and negative in the east" lasted from 1100PST to 1500PST for CMAQ and 1100PST to 1300PST for observations. As discussed in Chapter 3, such dynamic east-west ozone contrast captures the formation of daytime ozone plume in LFV. Given the above results, one may conclude that CMAQ generates higher-than-observed level of ozone plume at western LFV between mid-day to early afternoon.

The night time dynamic ozone contrasts of CMAQ are also more pronounced than observations. At 2100PST, the positive contrast in the eastern LFV is up to $+25$ *ppb* for CMAQ and $+8$ *ppb* for observations, the negative contrast in the western LFV is down to -10 *ppb* for CMAQ and -4 *ppb* for observations. These results imply that aforementioned CMAQ's "overproduction" of daytime ozone caused higher-than-observed level of night time ozone in the east.

Earlier analysis in Section 5.2 showed that compared to physical observa-



(a) Hours 0900PST and 1300PST.



(b) Hours 1500PST and 2100PST.

Figure 5.15: From PCA of the 1995 CMAQ output (top) and ozone observations (bottom): dynamic spatial plots of joint ozone feature $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}} + \mathbf{P}_3\mathbf{E}_3^{\mathrm{T}}$ at hours 0900PST, 1300PST, 1500PST and 2100PST on the 3rd day of 1995.

tion, CMAQ persistently overestimated the temporal ozone means throughout LFV. The comparison of dynamic ozone contrasts in this section revealed that the problem lies primarily on the fact that CMAQ modelled process generated (thus transported) higher-than-observed level of ozone plume in the western LFV. This CMAQ over-production of ozone may further explain the daytime pattern of feature difference $\mathbf{P}_1^d$ we saw in last section (Figures 5.11 and 5.12). Since CMAQ produces more ozone than the physical process during the daytime, the spatial means of CMAQ would be higher than that of the observations, i.e., $\mathbf{P}_1^c > \mathbf{P}_1^o$ during some daytime hours.

Furthermore, the joint-comparison of the 2nd and 3rd-order ozone features also reveals that the computer model is able to capture the overall pattern of east-to-west ozone advection that is observed physically. This is an important result that highlights WRF's capability of accurately modelling the wind patterns across LFV.

5.5 Chapter Conclusion

In this chapter, I developed and implemented means of CMAQ evaluation by combining methods of ozone PCA (Chapter 3) and ozone feature modelling (Chapter 4). Although the statistical analyses are done to evaluate CMAQ's capability to model space-time ozone, the overall methodology should apply to the evaluations of other AQMs. The central idea behind the proposed AQM evaluation is based on the observation-model comparison of data features (space-time structures of an air pollution field), and statistical modelling of the feature differences. The specific purpose of this chapter is to (1) develop the exact methods of feature-based AQM evaluation, and (2) implement these methods using CMAQ-WRF-SMOKE outputs and observation data to show the usefulness and advantages of feature-based model evaluation.

Implementation of my proposed evaluation methods revealed a few "big picture" similarities and differences between CMAQ ozone and observations. Compared to physical measurements, CMAQ tends to over-estimate episode means (average across hours of the episode) throughout LFV. This is ob-

served for 4 out of 5 episodes analyzed. However, the pattern of ozone variation across LFV is similar between CMAQ output and observations: the mean ozone levels are highest in the east and gradually decrease towards the west. Comparison of ozone features and GP modelling of feature differences identified two "sources" of this feature discrepancy:

- For all episodes, the difference in temporal means in LFV are statistically associated with the episode means of either the emission rates or antecedent concentrations of one ozone precursor (NOx or VOC).

  A detailed evaluation of the 2001 episode showed that the main source of discrepancy lies in the area of LFV where the episode averaged VOC emission rates are between 0.7 to 1.2 moles·sec^{-1}. This corresponds to the middle LFV, especially the eastern Metro Vancouver. This region is where much of daily ozone plume forms, and also an area where the local ozone process transitions from VOC-sensitive to NOx-sensitive.

  Furthermore, CMAQ over-estimation of observed episode mean is expected to be the most pronounced when VOC $\approx$ 0.82 moles·sec^{-1}. This is a feature of CMAQ deficiency expected from all LFV locations.

- Certain ozone features are what I refer to as "dynamic ozone contrasts" (Section 3.4), they capture the most dominant patterns of ozone plume advection across LFV and the magnitude (in *ppb*) of ozone formation/destruction. Comparison of these features has shown that CMAQ tends to produce higher-than-observed level of ozone pollution around Metro Vancouver during the ozone formation stage of a diurnal cycle. Thus transports a "bigger" ozone plume eastward across LFV.

  However, my analyses have also shown that WRF (the weather component of CMAQ) is able to simulate close-to-observed patterns of diurnal ozone transport across LFV.

In the end, the available evidence suggest that the source of observation-CMAQ difference lies primarily in the computer models' deficiencies in sim-

ulating processes of ozone precursor emission and photochemical reactions.

Furthermore, the ozone feature analyses and CMAQ evaluations are done for five LFV ozone episodes spanning two decades. I found that LFV ozone process is dominated by a few recurring spatial-temporal ozone features, and the episode-by-episode CMAQ evaluation resulted in similar, i.e., systematic, sets of conclusions.

These model evaluation results are made possible by the statistical comparison and analysis of ozone features. This highlights the important point that CMAQ (or any AQM) evaluation based on ozone features is more informative than direct observation-model comparison of data values. By deconstructing CMAQ output and observation data into informative ozone features, I was able to (1) evaluate how closely CMAQ can emulate the observed structure of space-time ozone means, and (2) how close-to-reality the CMAQ-WRF-SMOKE system can model the defining patterns of ozone advection, as well as the magnitude of ozone creation and destruction across LFV. The combination of ozone PCA and ozone feature comparison is a means to extract "maximum information" out of two compared ozone data. With the point-to-point comparison of data values, many important data structures are simply "hidden" from analysis.

Secondly, I proposed to model the ozone feature differences. As I have already summarized, this analysis not only revealed a definitive statistical association between CMAQ deficiency and SMOKE output, it also quantified the non-linear structure of this association. In turn, I was able to highlight a few specific areas in LFV where future SMOKE modelling effort should pay close attention. This description of SMOKE's spatial deficiency is especially useful for AQM modellers. In practice, due to the scarce availability of detailed emission measurements, one cannot simply analyze SMOKE-observation comparison data (Steyn et al., 2013). Therefore, the type of detailed CMAQ/SMOKE evaluations presented in this chapter are the unique outcomes of my proposed AQM evaluation approaches.

Lastly, in the existing literature of PCA-based AQM evaluation, informative discussions of AQM capabilities are based on authors' prior knowledge of the AQMs. The analyses in this chapter showed that using combined

methods of PCA and GP modelling, one can also achieve informative and systematic evaluation of AQM.

Chapter 6

AQM Evaluation II: Comparison of AQM and Observations as Stochastic Ozone Processes

In this chapter, I will implement the second proposed AQM evaluation method that was briefly explained at the beginning of Chapter 5. A more detailed description of this method, written in the context of CMAQ evaluation, follows:

1. Using the methods from Chapter 4, fit CMAQ ozone feature models using the CMAQ-WRF-SMOKE outputs, and fit separate observation ozone feature models using data from physical measurements.

2. Use the fitted ozone feature models to produce GP model outputs (make predictions) under *common* covariate settings. These model outputs are the estimated CMAQ and observation ozone features under the same covariates settings that capture the basic conditions of background weather and precursor pollution.

3. The estimated "common background" ozone features of CMAQ and observations are then compared.

Figure 5.1 from Chapter 5 showed in diagrammatic form the central idea behind my CMAQ evaluation.

To discuss the purpose of above evaluation approach, I will use the CMAQ evaluation of 1st-order feature $\mathbf{E}_1$ as an example. Let $\mathbf{E}_1^c$ and $\mathbf{E}_1^o$ be

the random processes representing the features of CMAQ ozone and physical observations. Further, let $\mathbf{x}^c$ and $\mathbf{x}^o$ denote the covariate sets that represent individual background conditions that the two ozone processes occur under. I propose to model $\mathbf{E}_1^c$ and $\mathbf{E}_1^o$ as GPs with $\mathbf{x}^c$ and $\mathbf{x}^o$ as model covariates: $\mathbf{E}_1^c(\mathbf{x}^c)$ and $\mathbf{E}_1^o(\mathbf{x}^o)$. After fitting the GP models, one can produce model outputs at a new *common* covariate setting $\mathbf{x}_0$: $\hat{\mathbf{E}}_1^c(\mathbf{x}_0)$ and $\hat{\mathbf{E}}_1^o(\mathbf{x}_0)$. These GP model outputs are statistically, the process *means* (or expected values) estimated at $\mathbf{x}_0$ given $\mathbf{E}_1^c(\mathbf{x}^c)$ and $\mathbf{E}_1^o(\mathbf{x}^o)$. They are the ozone features that are statistically expected to be produced by the two GPs.

By applying the same input $\mathbf{x}_0$, the outputs $\hat{\mathbf{E}}_1^c(\mathbf{x}_0)$ and $\hat{\mathbf{E}}_1^o(\mathbf{x}_0)$ are only different due to the parameters, i.e., the stochastic structures of $\mathbf{E}_1^c(\mathbf{x}^c)$ and $\mathbf{E}_1^o(\mathbf{x}^o)$. Hence, comparison of $\hat{\mathbf{E}}_1^c(\mathbf{x}_0)$ and $\hat{\mathbf{E}}_1^o(\mathbf{x}_0)$ is a means of comparing the overall statistical properties of CMAQ ozone and physical process under the same condition. Moreover, every point of model output has an associated standard error, which allows one to assess the significance of feature difference in space and time.

My proposed evaluation approach is a statistical means of addressing the need for a "process level understanding" between AQM and observations (Dennis et al., 2010; Galmarini and Steyn, 2010), and it is related to the "Probabilistic Evaluation" approach they mentioned. It should be noted that I am not evaluating the underlying physical or chemical processes governing an air pollution system, such as the chemical kinetics of specific reactions. Such detailed AQM evaluation are beyond the scope of my thesis.

The parameters in a Gaussian Process model quantify the influence of model covariate(s) on the random response variable. As described in Chapter 4, my GP model is the sum of a fixed regression component and a stochastic Gaussian Process. The regression coefficients model the linear association between each spatial/temporal ozone feature and variables such as longitude, latitude and hour of the day. The GP correlation parameters model the spatial or temporal behaviour of each ozone feature (pattern) as a non-linear function of model covariates. Hence by comparing the correlation and regression parameters between GP models $\mathbf{E}_j^c$ and $\mathbf{E}_j^o$, or $\mathbf{P}_j^c$

and $\mathbf{P}_j^o$, one can analyze how the ozone features of CMAQ and observation behave differently (or similarly) across a range of meteorological conditions and precursor pollution.

Such comparison can be purely numerical using statistical testing. However, as I shall demonstrate in the following analyses, the comparison of statistical models is more informative when done as I proposed.

6.1 Pre-analysis Comments

In Chapter 4, I estimated the CMAQ ozone feature models as Gaussian Processes, performed model diagnostics and assessed the models' performance in modelling both individual spatial/temporal ozone patterns and complete space-time ozone fields. Each ozone feature model is able to emulate the behaviour of the corresponding spatial or temporal processes. I also concluded that, once combined using equation (4.13), the ozone feature models are well-suited for forecasting hourly ozone across the entire spatial domain of the "rectangular" LFV, especially during the important period of daily 8-hour maximum. I believe the feature-based ozone model is an appropriate foundation upon which to implement my proposed CMAQ evaluation.

This chapter will use the same notations as the previous chapters. Suppose CMAQ or observational data of dimensions $t \times n$ are decomposed into $\mathbf{E}_{n\times n}$ and $\mathbf{P}_{t\times n}$, and GP models (4.1) and (4.2) are fitted to the decompositions. Let $\mathbf{E}_j^c$ and $\mathbf{P}_j^c$ denote the feature-based GP of CMAQ, and $\mathbf{E}_j^o$ and $\mathbf{P}_j^o$ denote the corresponding features of ozone observations. I further denote the statistical ozone model of CMAQ and physical observation as $\mathbf{O}^c$ and $\mathbf{O}^o$, hence

$$\begin{aligned} \mathbf{O}^c &\approx \sum_{j=1}^{3} \mathbf{P}_j^c \mathbf{E}_j^{c\mathrm{T}} \text{ and} \\ \mathbf{O}^o &\approx \sum_{j=1}^{3} \mathbf{P}_j^o \mathbf{E}_j^{o\mathrm{T}}. \end{aligned} \tag{6.1}$$

I will implement the proposed CMAQ evaluation using the 2001 and

2006 data. Specifically, the middle 3 days (72 hours) of the interpolated $n = 17$ CMAQ-related data and observation data from 2006 will be used as the "training dataset" to estimate the ozone feature models. The 96-hour 2001 CMAQ-WRF-SMOKE outputs at the same 17 locations will be used as the "common model covariate inputs" to produce the ozone feature model outputs. The 2006 data is already used as the training dataset in Chapter 4 to fit ozone feature models, so it is used here again for the same purpose. The 2001 CMAQ data captured a set of meteorological and precursor pollution conditions similar to the 2006, so it is used as model inputs to minimize uncertainties due to model extrapolation.

It is important to note that in the following data analysis, the ozone feature models are fitted using 2006 data. This means that CMAQ evaluation is done for the 2006 episode, because the compared stochastic models describe the 2006 ozone processes. The 2001 CMAQ data are simply used as common model inputs, but the evaluation is *not* done for the year 2001. Moreover, the 2001 CMAQ data are used to provide model inputs based on realistic combinations of weather conditions and precursor pollution. One may also use the observed meteorology and precursor data, but as discussed in Chapter 2, observations data may suffer from missing observations or measurement error.

As an alternative, one may input into the CMAQ ozone feature models the covariates from the corresponding observation data. The output from the CMAQ statistical model can then be compared to the observed ozones features. In such a test, the physical observations are regarded as the "benchmark" against which CMAQ is evaluated. One might argue that this is closer to the usual concept of model evaluation. My proposed CMAQ evaluation framework is designed not to have a preconceived notion that the observation is the benchmark, or even the truth. I regard them as individual ozone processes whose behaviour is being compared.

Guttorp and Walden (1987) discussed using bootstrapping to analyze the variation, thus the statistical significance of data differences. Compared to the method proposed here, their approach is non-parametric and thus more general. However, bootstrapping results can be difficult to interpret

when the sample is not independently and identically distributed. My proposed method, based on the comparison of GP model predictions, explicitly takes the process correlation structures into account. It also has the advantage of identifying specific conditions behind statistically significant feature differences.

Model Covariates

In the previous chapter, I modelled the statistical associations between the CMAQ model covariates and the ozone feature differences. In this chapter, the ozone feature models of observations are fitted using the data from physical measurements, which as mentioned in Chapter 2, only contain data on wind speed, temperature and ambient (antecedent) NOx concentration. This evaluation requires that both the CMAQ and observation ozone feature models have the same types of covariates, e.g., temperature, wind speed and ambient NOx concentrations. Hence, I am constrained to use fewer model covariates than those listed in Table 4.1.

The regression covariates of the GP models are the same as I described in Section 4.5. The stochastic-term covariates are the ones in Table 4.1 that are both available from CMAQ outputs and observation data, they are summarized in Table 6.1 for each ozone feature model.

6.2 Comparing the Space-time Ozone Processes

One intuitive way of CMAQ evaluation is to compare the space-time ozone fields produced by the CMAQ-based and the observation-based statistical ozone models. The detailed procedure is as follows:

1. Fit statistical ozone feature models $\mathbf{E}_j^c$, $\mathbf{E}_j^o$, $\mathbf{P}_j^c$ and $\mathbf{P}_j^o$, $j = 1, \ldots, 3$. That is, build GP models for the spatial and temporal ozone features of CMAQ and observations.

   The training data is the 2006 ozone episode. The statistical ozone models for CMAQ and observations are fitted using corresponding CMAQ and observed data, and the covariates of both models are those

	Model Covariate
$\mathbf{E}_1^c$ and $\mathbf{E}_1^o$	Longitude, Latitude, Elevation, $\text{Temp}_{E,1}$, $\text{Temp}_{E,2}$, $\text{Wind}_{E,1}$, $\text{Wind}_{E,2}$, $\text{NOx-lag}_{E,1}$
$\mathbf{E}_2^c$ and $\mathbf{E}_2^o$	Longitude, Latitude, Elevation, $\text{Temp}_{E,1}$, $\text{Temp}_{E,2}$, $\text{Wind}_{E,1}$, $\text{NOx-lag}_{E,1}$
$\mathbf{E}_3^c$ and $\mathbf{E}_3^o$	Longitude, Latitude, Elevation, $\text{Temp}_{E,1}$, $\text{Temp}_{E,3}$, $\text{Wind}_{E,1}$, $\text{NOx-lag}_{E,1}$, $\text{NOx-lag}_{E,2}$, $\text{NOx-lag}_{E,3}$
$\mathbf{P}_1^c$ and $\mathbf{P}_1^o$	$\text{Temp}_{P,1}$, $\text{Temp}_{P,3}$, $\text{Wind}_{P,2}$
$\mathbf{P}_2^c$ and $\mathbf{P}_2^o$	$\text{Temp}_{P,1}$, $\text{Temp}_{P,3}$, $\text{Wind}_{P,3}$
$\mathbf{P}_3^c$ and $\mathbf{P}_3^o$	$\text{Temp}_{P,1}$, $\text{Temp}_{P,3}$

Table 6.1: Covariates used for CMAQ evaluation within the stochastic component of the ozone feature (GP) models. The covariate sets are shortened (as compared to Table 4.1) due to the constraint imposed by unavailable observations. The acronym "NOx-lag" indicates antecedent or ambient NOx concentrations.

listed in Table 6.1. I did not perform covariate selection for the observation models due to a dearth of available observation data. Here, I assume that the observed ozone process is driven by the same background variables as the CMAQ model. The difference is *how* these two random processes behave under the same sets of covariates, and this "difference in behaviour" is the focus.

2. Apply a common set of covariate inputs to the fitted models. The input covariates are decomposed from the 96-hour, 2001 CMAQ-WRF-SMOKE outputs. This model input data have dimension 17×96.

3. Combine the GP model outputs $\hat{\mathbf{E}}_j^c$'s, $\hat{\mathbf{P}}_j^c$'s, $\hat{\mathbf{E}}_j^o$'s and $\hat{\mathbf{P}}_j^o$'s into space-time ozone fields $\hat{\mathbf{O}}^c$ and $\hat{\mathbf{O}}^o$ via (6.1).

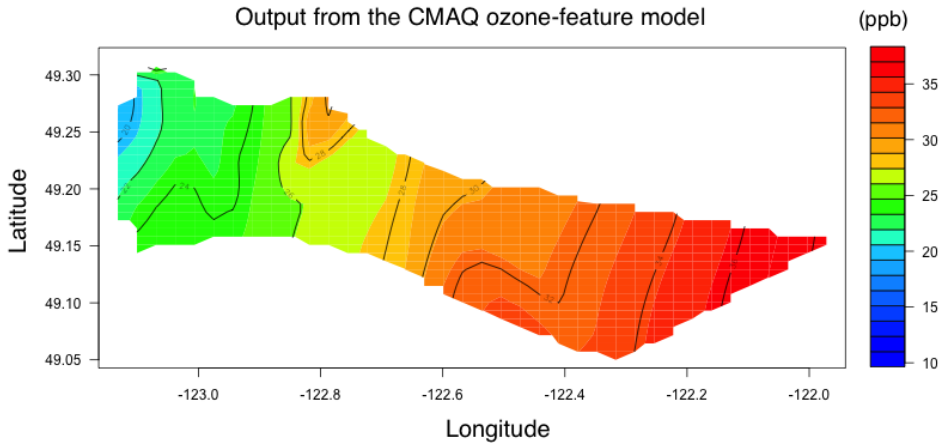
I should reiterate that the purpose here is not to make predictions, but rather to produce outputs from the statistical CMAQ and observation models given the same covariates representing background atmospheric conditions. Since $\hat{\mathbf{O}}^c$ and $\hat{\mathbf{O}}^o$ are ozone fields produced from exactly the same input, the comparisons between $\hat{\mathbf{O}}^c$ and $\hat{\mathbf{O}}^o$ can be viewed as an analysis of the way

the statistical CMAQ model and observation model differ as space-time processes. Furthermore, the input data are associated with an actual ozone episode, hence we are evaluating how CMAQ and observed ozone processes behave under conditions that are conducive to a real-world ozone episode.

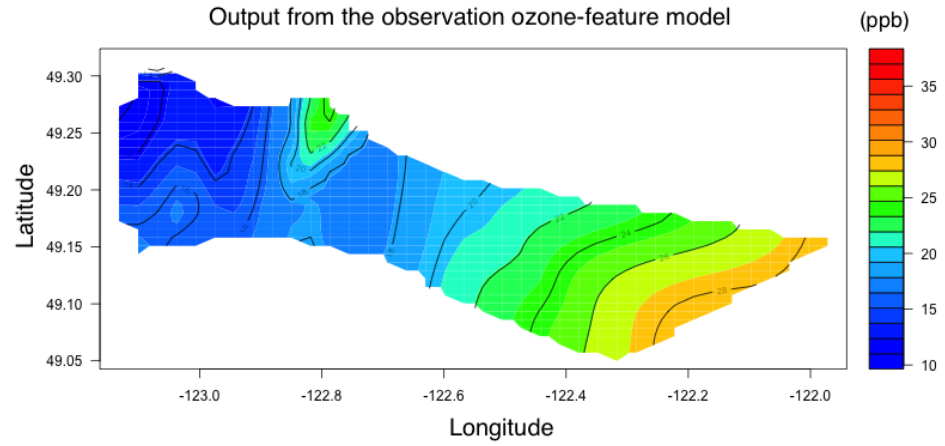
Figures 6.1a and 6.1b show the mean fields of $\hat{\mathbf{O}}^c$ and $\hat{\mathbf{O}}^o$ produced by two statistical ozone models given the same sets of covariate inputs (outputs from the 2001 episode). Both ozone model outputs $\hat{\mathbf{O}}^c$ and $\hat{\mathbf{O}}^o$ have dimension 96×17. The matrices are averaged across the 96 hours to obtain the spatial field of ozone means at the $n = 17$ locations, i.e., the $\hat{\mathbf{O}}_{t\times n}$ data are averaged by the columns. Cubic-spline smoothing is then applied to interpolate the $n = 17$ spatial data within the longitude-latitude boundary of measurement locations. Both ozone fields are plotted over the same colour scale for easy comparison. Figure 6.2 shows the equivalent hourly time-series of mean outputs (averaged across locations) produced by the statistical CMAQ and observation models.

Both the spatial (Figures 6.1a and 6.1b) and temporal (Figure 6.2) plots show that, even under the same background conditions, the CMAQ process tends to produce higher level of ozone than physical observation. Spatially, this result implies that location-by-location in LFV, the episode or temporal ozone means produced by CMAQ are uniformly higher than observations. From the plots of hourly LFV mean ozone (Figure 6.2), we see that using ozone observations as reference, the CMAQ model tends to over-predict the spatial means during the hours between 0000PST and 0800PST, in the second day especially, where the CMAQ spatial means can be more than twice the observed (shaded area in Figure 6.2). CMAQ also tends to produce higher LFV means during a few afternoon peak hours, but the magnitude of over-prediction is not as noticeable as the morning.

In summary, after controlling for differences in background conditions, the CMAQ modelled ozone fields still showed the same space-time patterns of over-prediction noticed from the comparison of original data - refer back to Figures 5.2 and 5.3 for spatial differences, Figures 5.13 and 5.14 for temporal differences. Therefore, the preceding evaluation indicates that the CMAQ is *statistically expected* to produce higher temporal/episode ozone



(a) The mean field (in *ppb*) produced by the statistical ozone model of CMAQ.



(b) The mean field (in *ppb*) produced by the statistical ozone model of observations.

Figure 6.1: Spatial fields of ozone means produced by the statistical ozone models of CMAQ and observation. Statistical ozone feature models are fitted using the middle 3 full days of 2006 CMAQ and observation data (72-hours). The common model inputs are the entire 96 hours of 2001 CMAQ/WRF/SMOKE output.

means throughout LFV, and higher hourly LFV mean ozone during the early morning and afternoon.

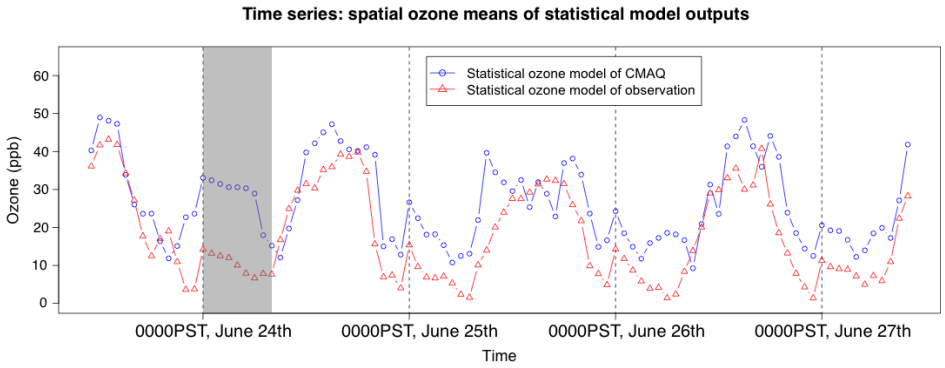


Figure 6.2: Hourly time series of LFV mean ozone (averaged across space by the hour) produced by the CMAQ and observation ozone models. The *feature-based* statistical ozone models are fitted using the 2006 CMAQ and observation data. The common input is the 2001 CMAQ data. The dashed lines indicate hour 0000 of each day, and the shaded region shows the hours when CMAQ over-prediction is the largest during the 96 hours.

In the last chapter, I modelled the covariates for $\mathbf{P}_1^d$: feature that represents the hourly differences in LFV mean ozone between CMAQ and observation. That analysis did not clearly identify one specific input of CMAQ run that is driving the ozone feature difference. I further raised the issue that CMAQ does not model certain nocturnal (or early morning) pollution processes that occur around LFV. The preceding analyses point to a observation-CMAQ difference of $\mathbf{P}_1$ at the *process* level.

6.3 Comparison of $\hat{\mathbf{P}}_1^c$ and $\hat{\mathbf{P}}_1^o$

In this section, I will perform statistical comparison of $\hat{\mathbf{P}}_1^c$ and $\hat{\mathbf{P}}_1^o$, features estimated under the same weather and pollution settings. Before further discussion, it is worth repeating that the difference between $\mathbf{P}_1^c$ and $\mathbf{P}_1^o$ mirrors the scale and temporal pattern of difference between the hourly LFV means of CMAQ and observations. This fact allows one to interpret the following comparison of $\mathbf{P}_1$ as a comparison of hourly LFV mean ozone. One may refer back to Section 5.3 for the discussions of this particular point.

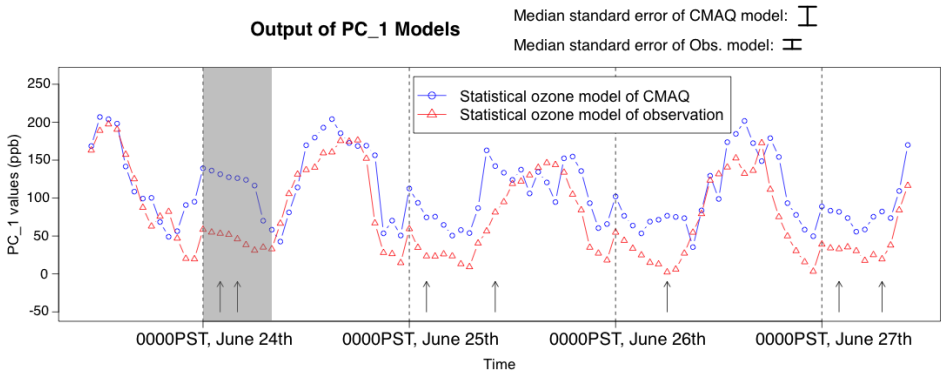
As discussed in Section 4.2, the GP model output from each set of covariate input is the conditional *mean* of the process at that setting, and this estimate of the mean has a standard error. Overall, there are 96 $\hat{\mathbf{P}}_1$'s and associated standard errors.

Figure 6.3a shows the temporal patterns of $\hat{\mathbf{P}}_1^c$ and $\hat{\mathbf{P}}_1^o$ estimated under the same covariate sets (weather and precursor pollution). I also plotted the "error bars" whose magnitudes indicate the median $\hat{\mathbf{P}}_1$ standard errors for the two statistical models, and the arrows indicate the hours where the 95% prediction intervals of $\mathbf{P}_1^c$ and $\mathbf{P}_1^o$ do not overlap. As shown, the 1st-order temporal feature difference between CMAQ and observation (or difference in hourly LFV means) are statistically significant for a few early morning hours and one daytime hour on the 3rd day of the 2006 episode.

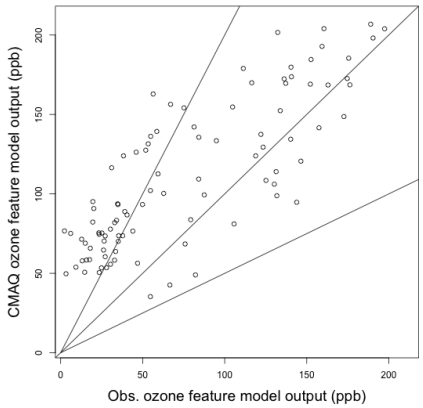
Figure 6.3b further shows the scatter plot of $\hat{\mathbf{P}}_1^o$ vs. $\hat{\mathbf{P}}_1^c$. We see that during early morning hours (when $\hat{\mathbf{P}}_1^o \leq 50$ *ppb*), the CMAQ ozone process tends to over-predict the observed hourly LFV means by more than 100%, whereas the daytime correspondence between CMAQ and observed ozone processes is much better.

Figure 6.4 shows the same $\hat{\mathbf{P}}_1^c$ and $\hat{\mathbf{P}}_1^o$ plotted as functions of Temp_{P_1}, a temperature covariate in both models (Table 6.1). In these plots, $\hat{\mathbf{P}}_1^c$'s and $\hat{\mathbf{P}}_1^o$'s at the same hour are averaged across days of the episode, and the same is done for covariate Temp_{P_1}. This averaging is done to smooth the daily variations of $\hat{\mathbf{P}}_1$ at similar Temp_{P_1} values. Temp_{P_1} captures the temporal features of the hourly mean LFV temperature (Section 4.3 and Appendix C.1). So to enhance interpretation, the x-axis in Figure 6.4 shows the mean temperature values that correspond to Temp_{P_1}.

The CMAQ model produced higher-than observed $\hat{\mathbf{P}}_1$ values across the temperature range typical of LFV ozone episodes. The small exception is shown between 26.0°C-27.0°C, where it alternates between $\hat{\mathbf{P}}_1^c < \hat{\mathbf{P}}_1^o$ and $\hat{\mathbf{P}}_1^c > \hat{\mathbf{P}}_1^o$. This pattern translates to CMAQ over-predictions of hourly LFV means during most of the day, but especially during periods of low temperature such as the hours between the late-night and the morning.



(a) Time series plots of $\hat{\mathbf{P}}_1^c$ and $\hat{\mathbf{P}}_1^o$.



(b) Scatter plot of $\hat{\mathbf{P}}_1^o$ vs. $\hat{\mathbf{P}}_1^c$.

Figure 6.3: Time series plots of $\hat{\mathbf{P}}_1^c$ and $\hat{\mathbf{P}}_1^o$ and scatter plot of $\hat{\mathbf{P}}_1^o$ vs. $\hat{\mathbf{P}}_1^c$. The time-series plot at the top shows the "error bars" whose magnitude indicate the median $\hat{\mathbf{P}}_1$ standard errors from the two ozone feature models, and the arrow indicate the hour where the difference between $\hat{\mathbf{P}}_1^c$ and $\hat{\mathbf{P}}_1^o$ is significant at type-I error = 0.05. The shaded area shows the hours when differences between $\hat{\mathbf{P}}_1^c$ and $\hat{\mathbf{P}}_1^o$ are the largest. The lines in the scatter plot are $y = x$, $y = 2x$ and $y = 1/2x$.

6.4 Chapter Conclusion

During CMAQ evaluation in Chapter 5, I found that the 1st-order temporal feature of CMAQ $\mathbf{P}_1^c$ tends to have noticeably higher values than the observed feature $\mathbf{P}_1^o$ during the morning hours between 0000PST to 0800PST,

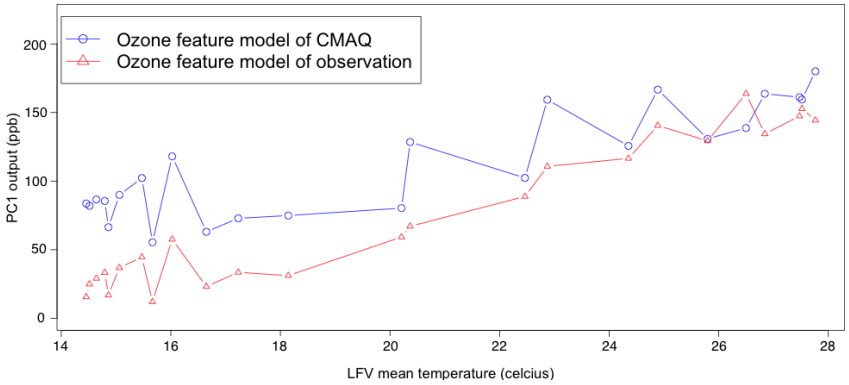


Figure 6.4: Univariate covariate-effect of $\hat{\mathbf{P}}_1^c$ and $\hat{\mathbf{P}}_1^o$ against temperature input Temp_{P_1} - a feature that represents mean LFV temperature. The models for $\mathbf{P}_1^c$ and $\mathbf{P}_1^o$ are fitted using the 2006 CMAQ and observation data. The input is processed from the CMAQ data of 2001.

and a few hours during afternoon peak. Compared to observations, this result corresponds to CMAQ's over-prediction of hourly LFV mean ozone during the morning and afternoon. The feature-based evaluations in this chapter revealed certain process-level differences between CMAQ ozone process and the reality (physical observation).

I applied the ozone feature models from Chapter 4 and implemented statistical analyses that answer the question: "would CMAQ produce higher-than observed ozone under the same atmospheric condition?" The statistical comparisons are based on the assumption that the CMAQ and observation features follow Gaussian Processes in the specific forms estimated in Chapter 4. The reasonableness of this assumption was extensively analyzed in Chapter 4.

The analyses have shown that, given the same background conditions in temperature, wind and ozone precursor concentrations, CMAQ is *statistically expected* to produce aforementioned temporal patterns of ozone over-prediction. During some morning hours, these hourly over-predictions are significant in the sense that the prediction intervals of $\mathbf{P}_1^c$ and $\mathbf{P}_1^o$ do not overlap. Plots of $\mathbf{P}_1^c$ and $\mathbf{P}_1^o$ against the temperature covariate further revealed that CMAQ is expected to produce higher-than-observed hourly LFV

mean ozone at temperatures below 25°C, i.e., outside of afternoon peaks.

Similar analyses of spatial features also showed that CMAQ will also produce higher temporal/episode ozone means than the physical observations throughout LFV. In other words, the spatial CMAQ over-predictions showed in Section 5.2 are expected to present under the same general weather conditions and precursor pollution.

Chapter 7

Conclusion

Traditional method of AQM evaluation directly compares the model outputs against observation data and summarize the deviation values into error statistics such as RMSE and MBE. However, as Dennis et al. (2010) pointed out (summarized in Section 1.2), AQM outputs and observation data are generated by discrepant physical processes, and without a deeper understanding on the complexities of air pollution system at hand, any direct observation-model comparisons are "fortuitous".

More informative and "big picture" approaches to AQM evaluation have been proposed over the years. These evaluations compare the data features obtained from the decompositions of AQM outputs and observation data. However, the differences in the data features are still summarized into statistical measures like correlations, angle between vectors, etc. The systematic observation-AQM differences are visually interpreted using authors' prior knowledge, as atmospheric scientists, about the inner workings of AQMs and physical processes.

The goal of this research is to develop novel statistical methods of AQM evaluations that (1) are more informative than the point-to-point data comparison, and (2) further the existing methods of feature-based AQM evaluation.

This chapter summarizes the novel contributions made in this thesis, both in the fields of statistical AQM evaluation and modelling of space-time ozone process. I will then finish the conclusion by proposing future works. Since the evaluations in this thesis are done for CMAQ modelling of LFV ozone, the following discussion will be written mainly under this context.

7.1 Main Contributions and the Novelty of the Evaluation Methods

Using combined methods of PCA and Gaussian Process modelling, I proposed and implemented means of AQM evaluations that provide a process-level understanding of the way AQM simulated ozone differ from physical observations. Here, the "process-level" evaluation refers to either the feature comparison of AQM ozone and observations as stochastic processes, or the comparison of their dynamic features, e.g., the dominant pattern of ozone advection, and the region and magnitude of ozone formation. A more detailed process-level evaluation, such as the chemical kinetics of certain reactions, is beyond the scope of this thesis.

In addition to the structural comparison of ozone features, I proposed two approaches to AQM evaluation that incorporate the methods of non-linear spatial and temporal modelling. They are:

1. Statistical modelling of ozone feature differences as Gaussian Processes driven by AQM inputs representing atmospheric conditions, ozone precursor emission rates and antecedent pollution. This method evaluates the statistical associations between the inputs and particular conditions of AQM run to its modelling capability.

2. Comparison of the statistical properties of AQM ozone and physical processes. This is done by estimating the ozone feature models for AQM and observations, then comparing their ozone features predicted under the *same* sets of background weather condition and precursor pollution. This method also assesses the statistical significance of any feature difference in both space and time.

The developed statistical methods are implemented to model and evaluate CMAQ ozone. I will now provide a simple recap of the evaluation results; a more detailed summaries are at the ends of Chapters 5 and 6. These results serve to demonstrate the usefulness of my proposed evaluation methods, thereby highlighting the contribution of this thesis.

7.1.1 Recap of Evaluation Results

The ozone feature analyses and CMAQ evaluations are done for five LFV ozone episodes spanning two decades. We found that LFV ozone process is dominated by a few recurring spatial-temporal ozone features, and the episode-by-episode CMAQ evaluation resulted in similar, i.e., systematic sets of conclusions. The following is an evaluation recap:

- Comparison of the 1st-order spatial features showed that CMAQ tends to over-estimate the observed local ozone means (averaged across time) almost uniformly across LFV. The over-estimate is observed for 4 out of 5 episodes. However, the east-west patterns of spatial ozone variation are similar between CMAQ outputs and observations.

- GP modelling of feature differences showed that above mentioned differences in the mean field are statistically associated with the *episode mean* emission rates or the antecedent concentrations of either NOx or VOC.

  A detailed feature difference modelling was performed for the 2001 episode. I found that the main sources of feature difference lie in the areas of LFV where the VOC emission rates are around 0.82 and 1.18 moles·sec^{-1}. These correspond to the middle of LFV in general, and eastern Metro Vancouver in particular. This region is where much of the daily ozone plume forms, and where the ozone process transitions from being VOC-sensitive to NOx-sensitive.

- Certain ozone features capture the most dominant pattern of ozone advection across LFV, as well as the area and the magnitude (in *ppb*) of ozone formation/destruction. Comparison of these features have shown that CMAQ tends to produce higher-than-observed level of ozone pollution around eastern Metro Vancouver during the ozone formation stage of a diurnal cycle. Subsequently, CMAQ ozone process transports a "bigger" ozone plume eastward across LFV.

  However, the same feature comparison also showed that WRF (weather

component of CMAQ) is able to simulate a close-to-observed general patterns of eastward ozone advection.

The above results suggest that the source of observation-CMAQ deviations in spatial ozone lies mainly in the CMAQ/SMOKE deficiencies in simulating LFV's spatial precursor emissions and their subsequent atmospheric reactions. Specifically, the suburbs of eastern Metro Vancouver are identified as areas where SMOKE modelling showed its deficiency. Hence, any future effort in air quality modelling should pay close attention to the accuracy of emission modelling at aforementioned locations.

The evaluation results relating to SMOKE modelling are especially useful. Due to the scarce availability of detailed emission observations (Steyn et al., 2013), it is difficult in practice to evaluate SMOKE outputs and make meaningful associations to CMAQ deficiency. My proposed evaluation method provided a way of addressing this concern.

The second proposed evaluation method analyzes, under the same atmospheric conditions, whether the space and time differences in ozone feature are significantly different. This is a method that uses the statistical ozone feature models to compare the stochastic structures of CMAQ ozone and the physical process. The results showed that, even under the same background conditions, CMAQ is expected to produce significantly higher-than-observed hourly LFV mean ozone (averaged across space) during the morning. Analyses also showed that CMAQ tends to over-estimate the hourly LFV mean at temperatures below 25°C, which are the temperatures outside of the afternoon peak hours. In addition, CMAQ is also statistically expected to produce a spatial ozone field with values that are uniformly higher than observed.

7.1.2 Conclusion on AQM Evaluation

In the end, air-quality model evaluation is not a well-defined science; there is no "one right way" of evaluation. What separates different model evaluation techniques and approaches is the different levels of informativeness, judged by the amount of insight and knowledge into the model behaviour that an

evaluation method can provide.

My proposed AQM evaluation methods provided the types of insight into the modelling capability of CMAQ not possible with direct observation-model data comparison. As the preceding evaluation recap and the detailed analyses in Chapter 5 and 6 have shown, the proposed AQM evaluations delivered an informative and coherent set of results that highlighted both the "big picture" and detailed input-level capability of the CMAQ-SMOKE-WRF system.

My proposed AQM evaluation methods, which are based on the frame work of statistical analysis and modelling of ozone features, constitute novel contributions to the existing works on AQM evaluation (AQMEII, Section 1.3). In particular, the methods proposed in this thesis add to the knowledge of features-based AQM evaluation (existing works summarized in Section 1.4).

7.2 Additional Contributions

Although the analyses in Chapters 3 and 4 are designed to provide the necessary tools for subsequent CMAQ evaluation, these works should also be considered useful contributions on their own. Combined analyses in Chapter 3 and 4 drew an important conclusion that a complex space-time pollution process can be conceptually understood and modelled statistically using a few leading features. The ozone feature models developed in Chapter 4 is especially a novel and efficient means of modelling a complex space-time air pollution process.

7.2.1 Understanding the Features of LFV Ozone

In Chapter 3, I formulated a detailed understanding of LFV ozone features during a summer-time ozone episode. I identified spatial-temporal ozone features that consistently appeared and dominated the ozone processes during the years 1985-2006. The effect of different wind flow patterns on the resultant ozone features are also studied.

During a LFV ozone episode dominated by wind regime types I, II and III, the most important dynamic process is the eastward advection of ozone plume. This transport is driven by a prevailing westerly wind flow. Under a type IV wind regime, there are two defining ozone circulation patterns: the daytime northwest-to-southwest ozone advection and nighttime west-to-east advection. These dynamic processes, in addition to space-time structures of ozone mean, account for over 90% of space-time ozone variation. Further analysis of eigenspectra indicates that these few features are statistically separable from other features.

7.2.2 Ozone Feature Models

I further developed statistical models for individual ozone features, and a framework where a complete space-time ozone field can be modelled through its features. Individual features are modelled as GPs driven by a set of variables describing background meteorology, ozone precursor emission rates and antecedent concentrations. Each ozone feature model is estimated through a forward selection algorithm based on statistical goodness-of-fit measures.

Forecasts of ozone features and resultant space-time ozone fields are made for the 4th day of the 2006 CMAQ output across a complex rectangular domain including LFV and surrounding mountains. The ozone feature models displayed good capability in emulating the complex non-linear structures of respective features. The predicted spatial features captured both the regional-scale patterns and localized details of the *true* spatial features with good numerical accuracy. By combining the predicted features, forecast was made for the hourly spatial ozone fields. The proposed feature-based ozone model is able to forecast the LFV's ozone fields at great spatial resolution, where the hourly forecasts captured the spatial details of local ozone both in the lower-valley region and across north shore mountains. The forecasting accuracy was especially good during the important daily ozone peak hours, with low RMSEs and near 0 prediction bias.

Given the complexities of running CMAQ (Sections 1.1 and 2.1), the developed ozone feature model can be useful as a statistical CMAQ emulator.

Furthermore, compared to the traditional statistical approach of directly modelling the "raw" space-time data, it is more computationally efficient to model the data through its features.

7.3 Future Work

Proposed future works mainly include applying the presented statistical methods on other data and refining the ozone feature models.

7.3.1 Application for Other Air Pollution Data

For the regional ozone fields analyzed in this thesis, one would perform AQM evaluation based on very few (sometimes only one) spatial-temporal features. This is because the observed ozone suffers from noticeable feature degeneracy, making high order feature-by-feature comparison difficult to justify. The existing works on PCA-based AQM evaluation analyzed data over a large spatial domain and longer time period. Fiore et al. (2003) and Eder et al. (2014) performed model evaluations based on air pollution field over eastern United States. Orsolini and Doblas-Reyes (2003) and Camp et al. (2003) performed ozone PCAs (not for AQM evaluation) over the Euro-Atlantic sector (20^o to 90^o latitude and 60^o to -90^o longitude) and the entire global tropic region. These works showed that for a large spatial air pollution field, the are at least 3 clearly structured and interpretable data features in addition to the mean.

The relative simplicity (compared to continental air pollution field) of the LFV ozone shows that, the space-time ozone means and dynamic contrasts (features capturing patterns of advection) are the only important features. Moreover, these features consistently dominated all episodes during the two decades between 1985-2006. Hence, one could justifiably evaluate CMAQ based on very few (sometimes only one) leading ozone features and construct a systematic view on the capability of CMAQ-WRF-SMOKE to interactively model the LFV's air pollution.

The disadvantage of the relative simplicity of LFV ozone is that I could

not, as evident in Chapters 5 and 6, compare and analyze many ozone features in order to demonstrate to fuller extent the utility of my AQM evaluation method. The aforementioned works on continental/global scale PCA did not consider the problem of feature degeneracy, so I do not know whether such large scale AQM evaluation allows for more orders of feature comparisons. Therefore, my foremost future works will focus on the applications of the developed evaluation methods to a much larger air pollution field: to study whether more non-degenerate features can be extracted and put through feature-based AQM evaluation and statistical modelling.

This thesis analyzed hourly spatial ozone data during episode days. Alternatively, one may apply the presented statistical methods to (1) study other space-time pollution process, such as hourly PM2.5 exposure, and (2) analyze long-term ozone data and/or daily maximum data.

7.3.2 Further Works on Ozone Feature Models

My preliminary analysis has shown that an LFV ozone process may be separated into three sub-regions, i.e., the LFV ozone field can also be modelled as spatially heterogeneous processes. This result was revealed when I built "local" ozone models through spatial sampling, where the models are based on the "raw" ozone data; they are *not* feature models. Application of k-means clustering identified three LFV areas with "similar" estimated model parameters: area across the foot of mountains, western and eastern LFV separated by Surrey. The ozone feature models developed in Chapter 4 did not account for possible spatial heterogeneity. Hence, another future work would be to improve upon the existing ozone feature models where the GP parameters are allowed to vary in space to account for spatial heterogeneity.

Besides AQM evaluation, the ozone feature models developed here may serve as a computationally efficient emulator of an AQM. Hence, it may be practical to summarize all relevant code for PCA and GP modelling into an `R` package for fast forecasting of AQM output.

Bibliography

Ainslie, B., Reuten, C., Steyn, D. G., Le, N. D., and Zidek, J. V. (2009). Application of an entropy-based bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants. *Journal of Environmental Management*, 90:2715–2729.

Ainslie, B. and Steyn, D. G. (2007). Spatiotemporal trends in episodic ozone pollution in the Lower Fraser Valley, British Columbia, in relation to mesoscale atmospheric circulation patterns and emissions. *Journal of Applied Meteorology and Climatology*, 46:1631–1644.

Ainslie, B., Steyn, D. G., Reuten, C., and Jackson, P. L. (2013). A retrospective analysis of ozone formation in the Lower Fraser Valley, BC, Canada. part ii: Influence of emission reduction on ozone formation. *Atmosphere Ocean*, 51:170–186.

Allen, M. R. and Tett, S. F. B. (1999). Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, 15:419–434.

Aslett, R., Buck, R. J., Duvall, S. G., Sacks, J., and Welch, W. J. (1998). Circuit optimization via sequential computer experiments: design of an output buffer. *Journal of the Royal Statistical Society, Series C*, 47:31–48.

Bastos, L. S. and O'Hagan, A. (2009). Diagnostics for gaussian process emulators. *Technometrics*, 51:425–438.

Beaver, S., Tanrikulu, S., Palazoglu, A., Singh, A., Soong, S. T., Jia, Y., Tran, C., Ainslie, B., and Steyn, D. G. (2010). Pattern-based evaluation of coupled meteorological and air quality models. *Journal of Applied Meteorology and Climatology*, 49:2077–2091.

Berrocal, V. J., Craigmile, P. F., and Guttorp, P. (2012). Regional climate model assessment using statistical upscaling and downscaling techniques. *Environmetrics*, 23:482–492.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2009). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, 15:176–197.

Bjornsson, H. and Venegas, S. A. (1997). A manual for eof and svd analyses of climate data. Technical Report CCGCR No. 97-1, McGill University.

Bloomfield, P., Royle, A. J., Steinberg, L. J., and Yang, Q. (1996). Accounting for meteorological effects in measuring urban ozone levels and trends. *Atmospheric Environment*, 30:3067–3077.

Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling: theory and applications*. Springer-Verlag.

Boubel, R. W., Fox, D. L., Turner, D. B., and Stern, A. C. (1994). *Fundamentals of Air Pollution*. Academic press.

Byun, D. and Schere, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (cmaq) modeling system. *Applied Mechanics Review*, 59:51–77.

Camp, C. D., Roulston, M. S., and Yung, Y. L. (2003). Temporal and spatial patterns of the interannual variability of total ozone in the tropics. *Journal of Geophysical Research*, 108:4643–4660.

CCME (2000). Canada wide standard for particulate matter (pm) and ozone. Technical report, Canadian Council of Ministries of the Environment. Available at `http://www.ccme.ca/ourwork/air.html?category_id=99`, accessed 2013-09-12.

Cohn, R. D. and Dennis, R. L. (1994). Evaluation of acid-deposition model using principal component spaces. *Atmospheric Environment*, 28:2531–2543.

Conti, S. and O'Hagan, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, 140:640–651.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of American Statistical Association*, 102:824–840.

Craigmile, P. F. and Guttorp, P. (2011). Space-time modelling of trends in temperature series. *Journal of Time-series Analysis*, 32:378–395.

Cressie, N. (1990). The origin of kriging. *Mathematical Geology*, 22:239–252.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of American Statistical Association*, 86:953–963.

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S. T., Scheffe, R., Schere, K., Steyn, D. G., and Venkatram, A. (2010). A framework for evaluation of regional-scale numerical photochemical modeling system. *Environmental Fluid Mechanics*, 10:471–489.

Dou, Y., Le, N. D., and Zidek, J. M. (2010). Modelling hourly ozone concentration fields. *The Annals of Applied Statistics*, 4:1183–1213.

Eder, B., Bash, J., Foley, K., and Pleim, J. (2014). Incorporating principal component analysis into air quality model evaluation. *Atmospheric Environment*, 82:307–315.

Finlayson-Pitts, B. J. and Pitts Jr, J. N. (1999). *Upper and Lower Atmosphere*. Academic Press.

Fiore, A. M., Jacob, D. J., Mathur, R., and Martin, R. V. (2003). Application of empirical orthogonal functions to evaluate ozone simulations with regional and global models. *Journal of Geophysical Research*, 108:4431–4445.

Fuentes, M. and Raftery, A. E. (2005). Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics*, 61:36–45.

Galmarini, S. and Steyn, D. G. (2010). Advancing approaches to the evaluation of regional scale air quality modeling system. Technical report, Air Quality Model Evaluation International Initiative. Available at `http://publications.jrc.ec.europa.eu/repository/handle/111111111/13563`, accessed 2010-06-12.

Gao, F., Sacks, J., and Welch, W. J. (1996). Predicting urban ozone levels and trends with semiparametric modeling. *Journal of Agricultural, Biological and Environmental Statistics*, 1:404–425.

Gotway, C. A., Ferguson, R. B., Herbert, G. W., and Peterson, T. A. (1996). Comparison of kriging and inverse-distance methods for mapping soil parameters. *Soil Science Society of America Journal*, 60:1237–1247.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C. (2006). Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions and Gaxes and Aaerosols from Nature). *Atmospheric Chemistry and Physics*, 6:3181–3210.

Guttorp, P. and Walden, A. (1987). On the evaluation of geophysical models. *Geophysical Journal of the Royal Astronomical Society*, 91:201–210.

Handcock, M. S. and Stein, M. L. (1993). A bayesian analysis of kriging. *Technometrics*, 35:403–410.

Hannachi, A., Jolliffe, I. T., and Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, 27:1119–1152.

Hardle, W. K. and Simar, L. (2012). *Applied Multivariate Statistical Analysis*, chapter 9. Springer.

Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, 6:1957–1971.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer calibration using high-dimensional output. *Journal of American Statistical Association*, 103:570–583.

Hobbs, W. R., Bindoff, N. L., and Raphael, M. N. (2015). New perspectives on the observed and simulated Antarctica sea ice extent trend using optimal fingerprinting techniques. *Journal of Climate*, 28:1543–1560.

Hogrefe, C., Rao, S. T., Zurbenko, I. G., and Porter, P. S. (2000). Interpreting information in time series of ozone observations and model predictions relevant to regulatory policies in the eastern united states. *Bulletin of the American Meteorological Society*, 81:2083–2106.

JAICC (2005). A report to CCME: An update in the support of canada wide standard for particulate matter (pm) and ozone. Technical report, Joint Action Implementation Coordinating Committee. Available at `http://www.ccme.ca/ourwork/air.html?category_id=99`, accessed 2013-09-12.

Jin, L., Harley, R. A., and Brown, N. J. (2011). Ozone pollution regimes modeled for a summer season in California's San Joaquin Valley: A cluster analysis. *Atmospheric Environment*, 45:4707–4718.

Jolliffe, L. (2002). *Principal Component Analysis, 2nd ed.* Springer.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492.

Jrrar, A., Braesicke, P., Hadjinicolaou, P., and Pyle, J, A. (2006). Trend analysis of ctm-derived northern hemisphere winter total ozone using self-consistent proxies: How well can we explain dynamically induced trends? *Quarterly Journal of Royal Meteorological Society*, 132:1969–1983.

Kalenderski, S. and Steyn, D. G. (2011). Mixed deterministic statistical modelling of regional ozone air pollution. *Environmetrics*, 22:572–586.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B*, 63:425–464.

Kleiber, W., Sain, S., Heaton, M., Wiltberger, M., Reese, C., and Bingham, D. (2014). Parameter tuning for a multi-fidelity dynamical model of the magnetosphere. *Annals of Applied Statistics*, 7:1286–1310.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78:1464–1480.

Krzanowski, W. J. (1979). Between group comparison of principal components. *Journal of American Statistical Association*, 74:703–707.

Le, N. D. and Zidek, J. M. (2006). *Statistical Analysis of Environmental Space-Time Process*. Springer.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.

Li, S., Anlauf, K., Weibe, H., Bottenheim, J., and Pucket, K. (1994). Evaluation of a comprehensive euclidean air-quality model with multiple chemical species measurement using principal component analysis. *Atmospheric Environment*, 28:3449–3461.

Lindstrom, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., and Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and Ecological Statistics*, 21:411–433.

Lippmann, M. (1989). Health effects of ozone. A critical review. *Journal of Air Pollution Control Association*, 39:672–695.

Liu, L., Hawkins, D. M., Ghosh, S., and Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100:13167–13172.

Liu, Z. (2007). *Combining Deterministic and Statistical Methods in Modeling Environmental Processes*. PhD thesis, UBC.

Lorenz, E. D. (1956). Empirical orthogonal functions and statistical weather prediction. Technical report, Statistical Forecast Project Report 1, Dept. of Meteorology, M.I.T.

Marmur, A., Liu, W., Wang, Y. H., Russell, A., and Edgerton, E. S. (2009). Evaluation of model simulated atmospheric constituents with observations in the factor projected space: CMAQ simulations of SEARCH measurements. *Atmospheric Environment*, 43:1839–1849.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246–1266.

Metro Vancouver (2012). Station information: Lowed Fracer Valley air-quality monitoring network. Technical report, metrovancouver. Available at `http://www.metrovancouver.org/services/air-quality/emissions-monitoring/monitoring/network/Pages/default.aspx`, accessed 2013-06-23.

Metro Vancouver (2013). Lowed Fracer Valley air-quality report. Technical report, metrovancouver. Available at `http://www.metrovancouver.org/services/air-quality/emissions-monitoring/monitoring/reports/Pages/default.aspx`, accessed 2013-06-23.

Monahan, A. H., Fyfe, J. C., Ambaum, M. H. P., Stephenson, D. B., and North, G. R. (2009). Empirical orthogonal functions: the medium is the message. *Journal of Climate*, 22:6501–6514.

North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J. (1982). Sampling errors in the estimations of empirical orthogonal functions. *Monthly Weather Review*, 110:699–706.

Nychka, D., Wikle, C., and Royle, A. J. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling: an International Journal*, 2:315–331.

Orsolini, Y. J. and Doblas-Reyes, F. J. (2003). Ozone signatures of climate patterns over the euro-atlantic sector in the spring. *Quarterly Journal of Royal Meteorological Society*, 129:3251–3263.

Pearson, K. (1902). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572.

Porter, P. S., Hogrefe, C., Gego, E., Foley, K., Goodwitch, J. M., and Rao, S. T. (2010). Application of wavelet filters in an evaluation of photochemical model performance. In Steyn, D. G. and Rao, S. T., editors, *Air Pollution Modeling and its Application XX*, chapter 4, pages 415–420. Springer.

Preisendorfer, R. W. (1988). *Principal Component Analysis in Meteorology and Oceanography*. Elsevier.

Preisendorfer, R. W. and Barnett, T. P. (1983). Numerical model-reality inter comparison tests using small-sample statistics. *Journal of the Atmospheric Science*, 40:1884–1896.

Reuten, C., Ainslie, B., Steyn, D. G., Jackson, P. L., and McKendry, I. (2012). The impact of climate change on ozone pollution in the Lower Fraser Valley, BC. *Atmosphere Ocean*, 50:42–53.

Richman, M. B. (1986). Review article: Rotation of principal components. *Journal of Climatology*, 6:293–335.

Robeson, S. M. and Steyn, D. G. (1990). Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmospheric Environment*, 24B:303–312.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. (1989). Design and analysis of computer experiments (with discussion). *Statistical Science*, 4:409–423.

Salmond, J. A. and McKendry, I. G. (2002). Secondary ozone maxima in a very stable nocturnal boundary layer: observations from the Lower Fraser Valley, BC. *Atmospheric Environment*, 36:5771–5782.

Schonlau, M. and Welch, W. J. (2006). *Methods for Experimentation in Industry, Drug Discovery, and Genetics*, chapter 14. Springer. Book edited by Dean, A. and Lewis, S.

Seagram, A., Steyn, D. G., and Ainslie, B. (2013). Modelled recirculation of pollutants during ozone episodes in the Lower Fracer Valley, B.C. In Steyn, D. G. and Timmermans, R., editors, *Air Pollution Modeling and its Application XXII*. Springer.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., and Powers, J. G. (2008). Description of advanced research WRF version 3. (technical report NCAR/TN-475+STR). Technical report, National Centre for Atmospheric Research.

Steyn, D. G., Ainslie, B., Reuten, C., and Jackson, P. L. (2011). A retrospective analysis of ozone formation in the Lower Fraser Valley, BC, Canada. Available at `https://circle.ubc.ca/handle/2429/36587`, accessed 2014-10-15.

Steyn, D. G., Ainslie, B., Reuten, C., and Jackson, P. L. (2013). A retrospective analysis of ozone formation in the Lower Fraser Valley, BC, Canada. part i: Dynamical model evaluation. *Atmosphere Ocean*, 51:153–169.

Steyn, D. G., Bottenheim, J. W., and Thomson, R. B. (1997). Overview of the tropospheric ozone in the Lower Fraser Valley, and the Pacific '93 field study. *Atmospheric Environment*, 31:2025–2035.

Storch, H. V. and Zwiers, F. W. (1999). *Statistical Analysis of Climate Research*. Cambridge University Press.

Stull, R. B. (1988). *An Introduction to Boundary Layer Meteorology*. Springer Books.

Taylor, B. (1992). The relationship between ground-level ozone concentrations, surface pressure gradients, and 850mb temperatures in the Lower Fraser Valley of British Columbia. Technical Report PAES-92-3, Atmospheric Issues and Service Branch, Pacific Region, Environment Canada.

Taylor, E. (1991). Forecasting ground-level ozone in vancouver and the Lower Fraser Valley of British Columbia. Technical Report PAES-91-3, Scientific Service Division, Pacific Region, Environment Canada.

Thiebaux, H. J. and Zwiers, F. W. (1984). The interpretation and estimation of effective sample size. *Journal of Climate and Applied Meteorology*, 23:800–811.

Thompson, M. L., Reynolds, J., Cox, L. H., Guttorp, P., and Sampson, P. D. (2001). A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment*, 35:617–630.

US Environmental Protection Agency (2010). MOBILE6 Vvehicle Emission Modelling Software. Technical report, U.S.E.P.A. Available at `http://www.epa.gov/otaq/m6.htm`, accessed 2013-11-09.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics*, 34:15–25.

WHO (2003). Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide. Technical report, WHO Regional Office for Europe. Available at `http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/pre2009/health-aspects-of-air-pollution-with-particulate-matter,-ozone-and-nitrogen-dioxide`, accessed 2013-09-13.

Willmot, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. S., Klink, K. E., Legates, D. R., O'Donnell, J., and Rowe, C. M. (1985). Statistics for the evaluation and comparison of models. *Journal of Geographical Research*, 90:8995–9005.

Yarwood, G., Rao, S., Yocke, M., and Whitten, G. Z. (2005). The carbon bond mechanism: CB05. Technical report, U.S. Environmental Protection Agency.

Zidek, J., Le, N. D., and Liu, Z. (2012). Combining data and simulated data for space-time fields: application to ozone. *Environmental and Ecological Statistics*, 19:37–56.

Appendix A

Appendix Related to Chapter 2

A.1 Details on Simulated Ozone Data

The aim is to emulate, in a simplified form, the space-time feature of west-to-east ozone advection in LFV (Sections 3.1 and 3.4). This simulation does not include the ozone process over the mountains. Simulation provides space-time ozone data whose underlying structures and statistical properties are known exactly; it also allows for repeated realizations of the *same* ozone process. The usefulness of simulated data is demonstrated in Appendix B.1.

I simulate 3 types of space-time ozone data:

1. Simulated *true* ozone: this is the underlying *true* space-time process, which in reality, is never known.

2. Simulated CMAQ output. This is created based on simulated *true* ozone. Moreover, I simulate CMAQ output on a complete and regular spatial grid as well as an irregular grid at the real-life observation locations shown in Figure 2.1.

3. Simulated physical observations. These are simulated at locations where real-life ozone monitoring stations are situated.

Simulated CMAQ output and physical observations are both generated by adding error functions to the simulated *true* ozone. Temporally, the simulations are created for a 24-hour period of 0000-2300 during an episode.

In this section, I will first present the method of generating the *true* LFV ozone field, followed by the methods of generating synthetic CMAQ outputs

and physical observations.

A.1.1 Simulated *True* Ozone Data

Before developing simulated data, I first define a few key variables:

- The usual geographical coordinate system of longitude-latitude is not used. Although frequently used to identify locations on a 2-D plane, they actually measure the angle of a location in reference to the meridian and equator, and it is only necessary in large domain studies where the curvature of the earth makes the horizontal and vertical coordinates non-Cartesian. To avoid misunderstanding, self-defined LFV is mapped onto a 2-D Cartesian coordinate system. The Southwest corner of the LFV is $(x, y) = (0, 0)$ and the Northeast corner is $(x, y) = (100, 100)$. My formulas for generating simulated data are based on this (x, y) Cartesian coordinate, $x = 0, \ldots, 100$ and $y = 0, \ldots, 100$.

- The ozone generating functions are also time-dependent. Since the simulation runs between 0000 to 2300, I define an hourly time variable h, $h = 0, \ldots, 23$, and transform functions $f(h)$ and $f_T(h)$:

$$f(h) = \begin{cases} 1.714 \cdot h - 10.286 & h = 6, \ldots, 20 \\ 0 & h = 0, \ldots, 5 \text{ and } 21, \ldots, 23, \end{cases}$$

$$f_T(h) = \begin{cases} 0.333 \cdot h - 1.667 & h = 5, \ldots, 20 \\ 0 & h = 0, \ldots, 4 \text{ and } 21, \ldots, 23. \end{cases}$$

  The units of $f(h)$ and $f_T(h)$ are both *hour*. The above formulation is by no means necessary, these step-wise linear functions $f(h)$ and $f_T(h)$ are used to transform daily hour values $h = 0, \ldots, 23$ into a series of values (found by trial and error) convenient for simulating daily ozone and temperature.

Weather Variables

There are two weather variables built into the simulation: wind and temperature. Wind represents the driving force behind the phenomenon of ozone transportation. Here, I define a westerly wind system that transports the ozone plume eastward (Taylor, 1991; Ainslie and Steyn, 2007) with a constant speed of 3 $m \cdot s^{-1}$, below which is considered a light wind (Ainslie and Steyn, 2007).

Temperature is an important factor controlling the rate of photochemical reactions (Robeson and Steyn, 1990; Taylor, 1992; Reuten et al., 2012). During a summer-time ozone episode, it has a diurnal profile similar to Figure A.1. This curve is created using the equation

$$T(h) = \{-[f_T(h) \cdot U_h]^2 + 5 \cdot f_T(h) \cdot U_h + 17\} \cdot U_T, \quad t = 0, \ldots, 23,$$

where $U_h = 1 \cdot hour^{-1}$ to make the hourly function terms unitless and $U_T = 1°\text{C}$ to give $T(\cdot)$ an appropriate temperature unit. This function allows for a season-appropriate minimum daily temperature of 17°C during the early-morning and the evening. The 2nd order term in turn creates a concave downward function during the daytime with daily maximum of 23°C occurring between noon and 1300.

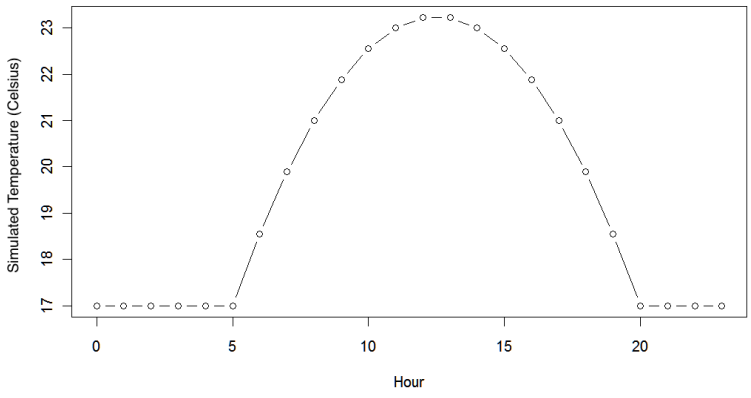


Figure A.1: Simulated diurnal temperature profile.

The Ozone Field

Let x and y range from 0 to 100, and h vary from 0 to 23. Equation (A.1) produces a eastward-moving "Gaussian hill" that simulates daytime ozone,

$$^{t}O(x,y,h) = \{c(h) \cdot \exp\left[-\left(\frac{x-\mu_x(h)}{\sigma_x(h)}\right)^2 - \left(\frac{y-\mu_y(h)}{\sigma_y(h)}\right)^2\right] + 15\text{ppb}\}. \tag{A.1}$$

As I will show, the term $c(h)$ has unit *ppb*, while $(x-\mu_x(h))/\sigma_x(h)$ and $(y-\mu_y(h))/\sigma_y(h)$ are unit less. Together with the addition of the constant 15ppb, (A.1) produces ozone data with unit *ppb*.

It should be noted that Gaussian function (A.1) is used for the purpose of generating simulation. It is a simplistic description of the more complex real-life ozone process, and the use of gaussian spatial profile is a convenient approximation.

Equation (A.1) contains a collection of time-dependent functions: $c(h)$, $\mu_x(h)$, $\mu_y(h)$, $\sigma_x(h)$ and $\sigma_y(h)$. The aforementioned wind and temperature variables are incorporated into these functions, thus influencing the Gaussian ozone function. Here are the details needed to construct (A.1).

- $c(h)$ incorporates the diurnal temperature and ozone pattern, and acts as a time-dependent scaling function in (A.1). It is described by the function:

$$c(h) = \begin{cases} c_0(h) & T(h) \leq 20^\circ\text{C} \\ \frac{1}{20^\circ\text{C}} \cdot T(h) \cdot c_0(h) & T(h) > 20^\circ\text{C} \end{cases}$$

  $T(h)$ is the aforementioned diurnal temperature profile. The stepwise function scales the daily temperature values to 1 at $T(h) \leq 20^\circ$C and to values above 1 at $T(h) > 20^\circ$C using 20°C^{-1}. The scaled hourly temperatures are then applied to (A.1).

  $c_0(h)$ depends on time as:

$$c_0(h) = 24\text{ppb} \cdot \left\{1 + \frac{-[f(h) \cdot U_h]^2 + 24 f(h) \cdot U_h - 48}{48}\right\},$$

where $f(h)$ and U_h are already defined hourly function and its unitless scalar. The constant 24ppb gives the scaling function $c(h)$ a desired diurnal profile and a unit of *ppb*. By definition, $c_0(h)$ is 0 when $f(h) = 0$. In turn, $c(h)$ becomes 0 and the equation (A.1) will take a minimum background ozone of 15ppb. This takes place between late evening and early morning. $c(h)$ takes positive values during 0600-2000. The diurnal profile of $c(h)$, $h = 0, \ldots, 23$, is shown in Figure A.2.

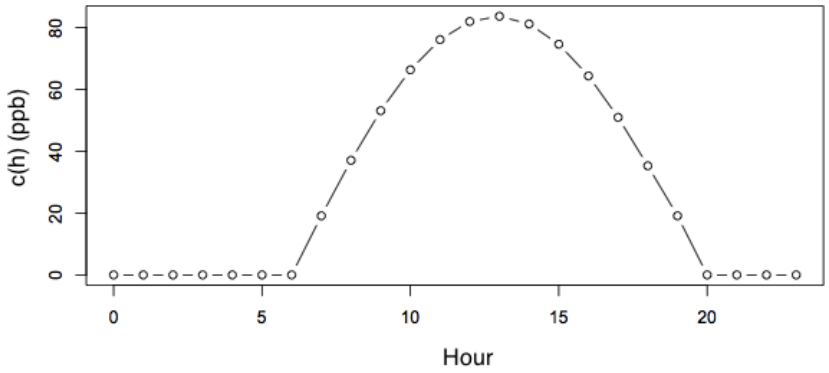


Figure A.2: Simulated diurnal profile of $c(h)$, the time-dependent scaling factor in (A.1).

- $\mu_x(h)$ and $\mu_y(h)$ are the locations of the maximum of the Gaussian hill along the x-axis and y-axis, i.e., the hourly location of the highest ozone concentration. Making these vary by the hour will enable the ozone field to travel in any direction as the day goes by. In order to mimic the commonly observed ozone movement in the LFV, $\mu_y(h)$ is fixed at the middle of the y-range: $\mu_y = 50$ for $h = 0, \ldots, 23$. As the map in Figure 2.1 shown, the LFV's vertical extent is less than the horizontal extent. Because I specified both location variables x and y to range from 0 to 100, it is not sensible to let both have the same unit. Here, I let x have unit *kilometre* (km) and y have an arbitrary distance unit U_y. The units of x and y are not critical since both $(x - \mu_x(h))/\sigma_x(h)$ and $(y - \mu_y(h))/\sigma_y(h)$ terms are unitless.

Moreover, letting $\mu_x(h)$ increase during daytime would simulate an *eastward* ozone movement. Conversely, a decreasing $\mu_x(h)$ translates to a westward ozone movement. $\mu_x(h)$ has form

$$\mu_x(h) = Wind \cdot U_w \cdot f(h) + 24 \cdot km.$$

As discussed, the parameter $Wind$ is fixed at 3m/s, together with a positive sign, it corresponds to an eastward ozone movement at a constant speed of 3m/s. The constant $U_w = 3.6(km \cdot s)/(hour \cdot m)$ transforms the wind parameter into unit *kilometres per hour* ($km \cdot hour^{-1}$). The larger the parameter $Wind$, the faster the simulated ozone plume travels. Multiplied by $f(h)$, an hourly variable with unit $hour$, I obtain a distance unit km for the $\mu_x(h)$.

An additive term of $24 \cdot km$ results in $\mu_x(7) \approx 30km$, i.e., the centre of ozone formation at 0700 takes place near $x = 30km$, the approximate centre of Vancouver city in this simulated LFV.

- $\sigma_x(h)$ and $\sigma_y(h)$ are time-varying spatial values that control the "spread" of the Gaussian field in x and y orientations. Observations and modellings show that the spatial variation is larger from East to West than North to South, so $\sigma_x(h) > \sigma_y(h)$ for $h = 0, \ldots, 23$. I define them as time-dependent 2nd order functions:

$$\begin{aligned} \sigma_y(h) &= 24 \cdot \left\{ 1 + \frac{24 f(h) U_h - [f(h) U_h]^2}{144} \right\} \cdot U_y \text{ and} \\ \sigma_x(h) &= (1.5km) \cdot (\sigma_y U_y^{-1}). \end{aligned}$$

A quick explanation of unit constants: U_y is the aforementioned distance unit of y, the constant $1.5km$ in the second equation gives σ_x a unit of km.

In summary, the defining characteristics of the simulated *true* ozone data are easily summarized by understanding the temporal functions $c(h)$, $\mu_x(h)$, $\mu_y(h)$, $\sigma_x(h)$ and $\sigma_y(h)$. The scaling function $c(h)$ incorporates diurnal temperature and ozone trend to increase or decrease the spatial ozone levels at

the appropriate hours. The means $\mu_x(h)$ and $\mu_y(h)$ incorporate the wind information to transport the simulated ozone field with the appropriate direction and speed (ozone advection). The mean functions also define the hourly locations of high ozone concentrations. Finally, the spatial deviation functions $\sigma_x(h)$ and $\sigma_y(h)$ control how widely the hourly ozone field is spread along the East-West and North-South orientations.

All functional coefficients and scale values were determined following extensive fine-tuning. Figures A.3 and A.4 show for selected hours, the spatial ozone field of the simulated true data. The simulated data are produced over a spatial grid of 51×51 (horizontal$\times$vertical) cells, hence the ozone data shown have dimension 24×2601.

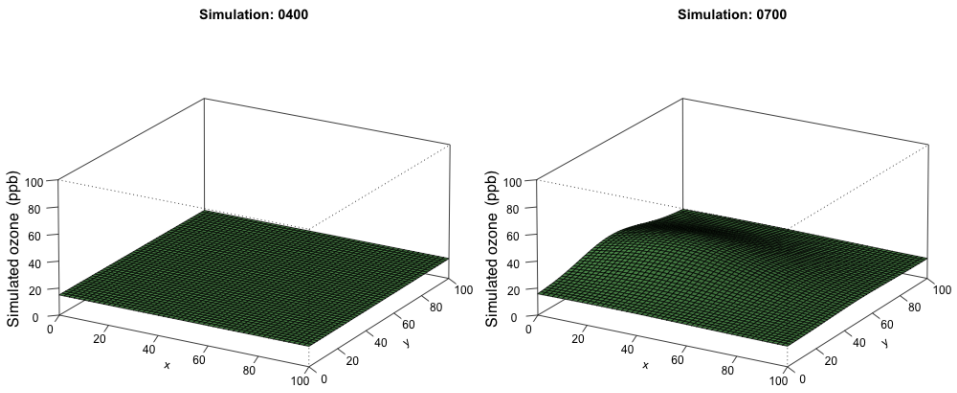
A.1.2 Simulated CMAQ Output and Observation

The CMAQ model output and physical observation are commonly regarded as functions of the *true* underlying ozone level ${}^tO(x, y, h)$ with an additive random error (Kennedy and O'Hagan, 2001; Fuentes and Raftery, 2005). If ${}^cO(x, y, h)$ and ${}^oO(x, y, h)$ are the simulated CMAQ output and observation respectively at location (x, y) and time h, they are expressed by the formulas:

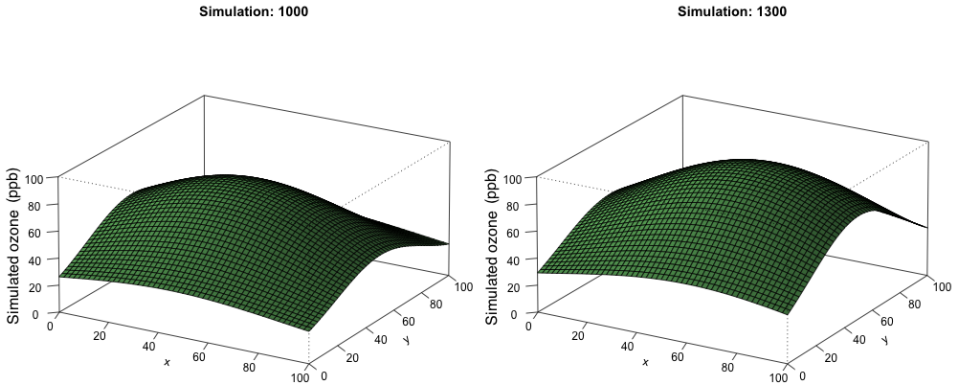
$$\begin{aligned} {}^cO(x, y, h) &= f_{cmaq}[{}^tO(x, y, h)] \text{ and} \\ {}^oO(x, y, h) &= {}^tO(x, y, h) + \varepsilon_o, \end{aligned} \tag{A.2}$$

where ε_o is independent random error of observation. Although the observed value is often treated as the "true" ozone level when judging against air-quality model output, random measurement error (ε_o) is in reality unavoidable. The task of ozone monitoring may further be complicated by a host of factors. However, in general it is sufficient to regard physical measurement as the sum of true value and an additive error (Dennis et al., 2010). Observations are simulated at the actual observation locations: I transformed the longitude-latitude of ozone monitoring sites onto the 2-D (x, y) coordinate system of simulated *true* ozone field.

The procedure for simulating a space-time CMAQ output is more involved. First, one needs to conceptualize the *real-life* relationship between



(a) Hours 0400 and 0700.



(b) Hours 1000 and 1300.

Figure A.3: Simulated *true* ozone fields at hours 0400, 0700, 1000 and 1300.

observation and corresponding (in space and time) CMAQ model output. Figure A.5 plots observations against CMAQ using data from June 26th, 2006, the 4th day of the 2006 ozone episode at all observation locations. The middle diagonal line is $x = y$, the other two lines are $x = 0.5 \cdot y$ and $x = 2 \cdot y$. CMAQ output are interpolated on the locations of the physical measurements as described in Section 2.2.

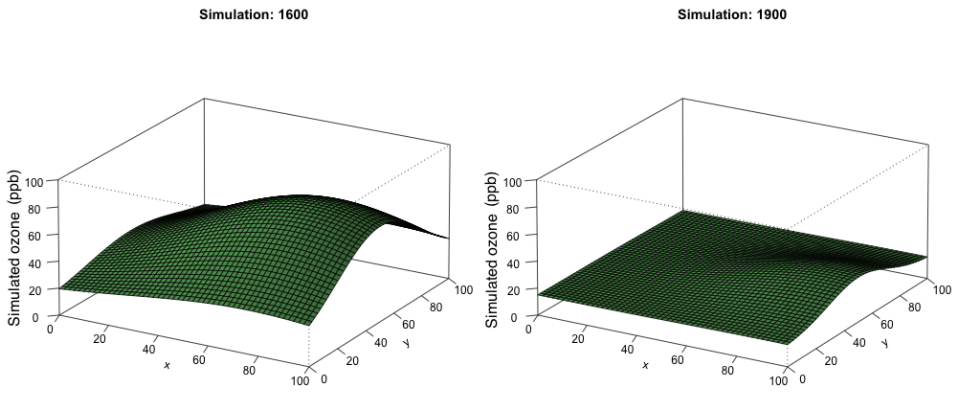


Figure A.4: Simulated *true* ozone fields at hours 1600 and 1900.

The goal is to simulate the slightly downward concave pattern seen in Figure A.5. After some trial and error, the detailed formulation of ${}^cO(x,y,h)$ in (A.2) is determined to be

$$\begin{aligned} {}^cO(x,y,h) &= f_s[\delta(x,y,h)] \cdot \delta(x,y,h), \quad \text{where} \\ \delta(x,y,h) &= |{}^tO(x,y,h) + \varepsilon_c|, \quad \varepsilon_c \sim N(\mu_c, \sigma_c^2), \quad \text{and} \\ f_s[\delta(x,y,h)] &= a_{f\_s}\delta(x,y,h)^2 + b_{f\_s}\delta(x,y,h) + 1.02, \\ \text{where} \qquad & a_{f\_s} = -1.25 \cdot 10^{-4} \frac{1}{\text{ppb}^2}, b_{f\_s} = 6.74 \cdot 10^{-3} \frac{1}{\text{ppb}}. \end{aligned} \tag{A.3}$$

The function $\delta(x,y,h) \geq 0$ ppb is the absolute value of the sum: *true* simulated ozone plus random error. The ε_c is the random additive error of CMAQ, it has unit *ppb*.

In order to capture aforementioned CMAQ behaviour into the simulated data, I use multiplicative scaling function $f_s[\delta(x,y,h)]$ in (A.3), which produces scaling factors for ${}^cO(x,y,h)$. A plot of scaling factors against a range of $\delta(x,y,h)$ is shown in Figure A.6. Values of $f_s[\delta(x,y,h)]$ start above 1 at $\delta(x,y,h) = 0$ and drop below 1 when $\delta(x,y,h)$ reaches higher ozone levels. Such scaling profile helps to simulate the concave pattern in Figure A.5. Note that the constants a_{f_s} and b_{f_s} in the scaling function have units

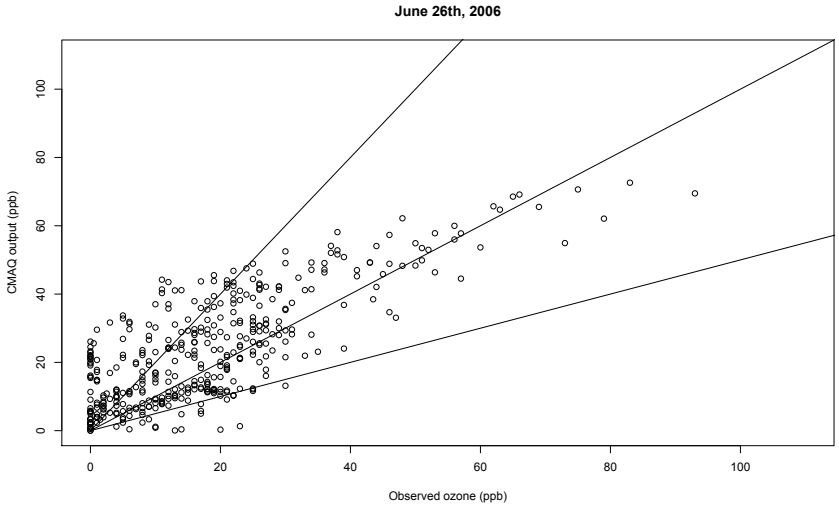


Figure A.5: CMAQ output vs. observation for June 26th, 2006, the 4th day of 2006 ozone episode.

that make $f_s[\delta(x, y, h)]$ unit less, which gives ${}^cO(x, y, h)$ a unit ppb when multiplied by $\delta(x, y, h)$.

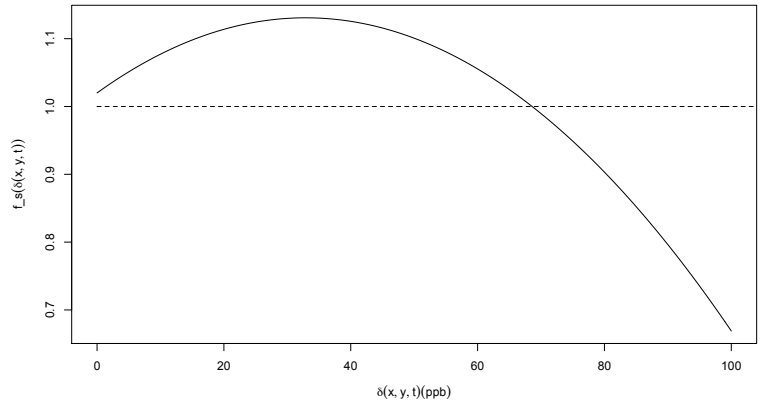


Figure A.6: $f_s[\delta(x, y, h)]$ (scaling term in (A.3)) plotted against $\delta(x, y, h)$.

Lastly, one need define the parameters of the random errors: $\varepsilon_o \sim N(0, \sigma_o^2)$ and $\varepsilon_c \sim N(\mu_c, \sigma_c^2)$. Table A.1 lists the parameter values, which

are based on the grand mean of the simulated true data:

$$\bar{Z} = \frac{1}{24 \cdot 100 \cdot 100} \sum_{h=0}^{23} \sum_{x=1}^{100} \sum_{y=1}^{100} {}^{t}O(x, y, h) = 31.5\text{ppb}.$$

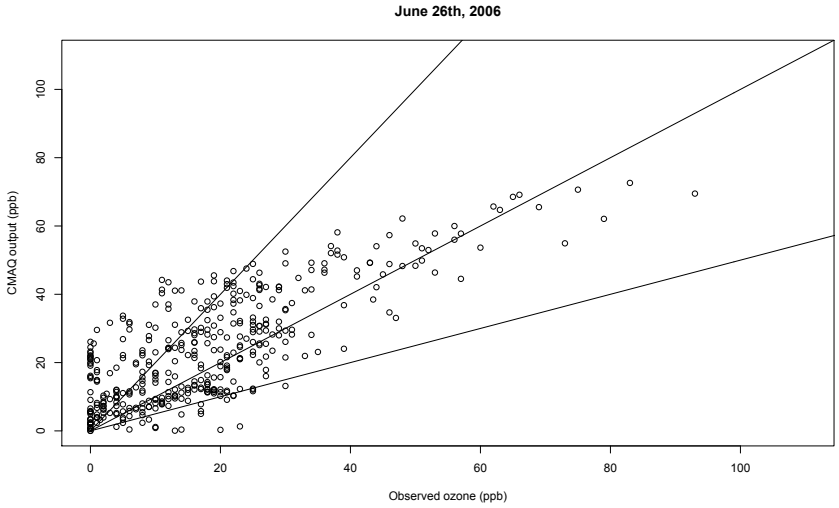
In keeping with the knowledge of CMAQ deficiency in low-concentration ozone modelling, the mean and variance of random errors increase when ${}^{t}O(x, y, h) \leq 30\text{ppb}$.

	σ_o^2	$\mu_c({}^{t}O(x,y,h) > 30ppb)$	$\sigma_c^2({}^{t}O(x,y,h) > 30ppb)$
values	$0.20 \cdot \bar{Z}$	$0.1 \cdot \bar{Z}$	$0.25 \cdot \bar{Z}$
		$\mu_c({}^{t}O(x,y,h) \leq 30ppb)$	$\sigma_c^2({}^{t}O(x,y,h) \leq 30ppb)$
values		$0.1 \cdot \bar{Z} + 2 \cdot ppb$	$0.25 \cdot \bar{Z} + 3 \cdot ppb$

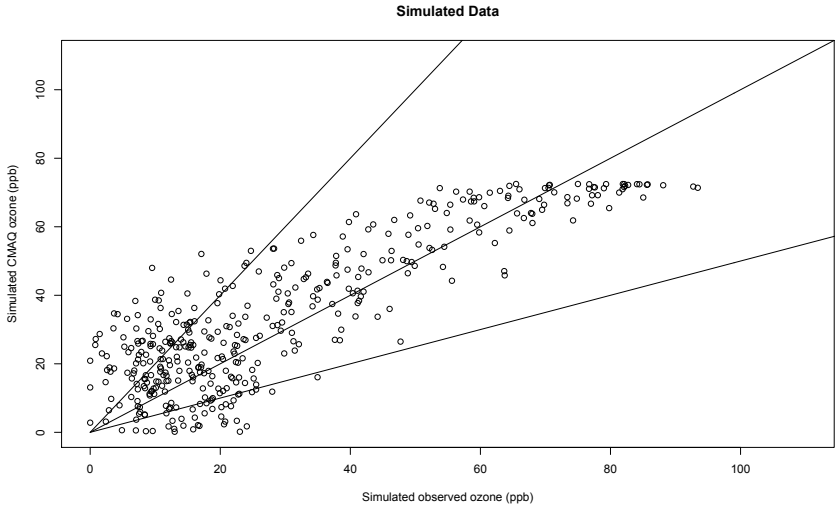
Table A.1: Parameter values used for generating additive errors in simulated CMAQ and observation.

Figure A.7b plots the simulated CMAQ against the simulated observations, the same plot using the real data is shown again for convenience in Figure A.7a. The percentage of points lying within the bounds are around 70% for both simulation and real data. As seen from the "scatter patterns", the simulations captured the main characteristics of the "real" CMAQ outputs and observations. Figure A.8 shows the same scatter plot of the simulated CMAQ against observations, where the CMAQ is simulated without the use of scaling function $f_s[\delta(x, y, h)]$ in (A.3). It is noticeable that without the scaling factors, the simulated CMAQ output no longer capture the important "concave relationship" with the observations, indicating the importance of $f_s[\delta(x, y, h)]$ as a part of simulation procedure.

The simulated ozone observations are not used in ozone PCA. Due its straightforward relationship with true ozone, observations are simulated to provide reference points for the formulation of (A.3) (method of simulating CMAQ output).



(a) CMAQ output vs. observation for June 26th, 2006, the 4th day of 2006 ozone episode (repeated from Figure A.5).



(b) Simulated CMAQ vs. simulated observation.

Figure A.7: Scatter plots for assessing "similarities" between the real data and simulated data.

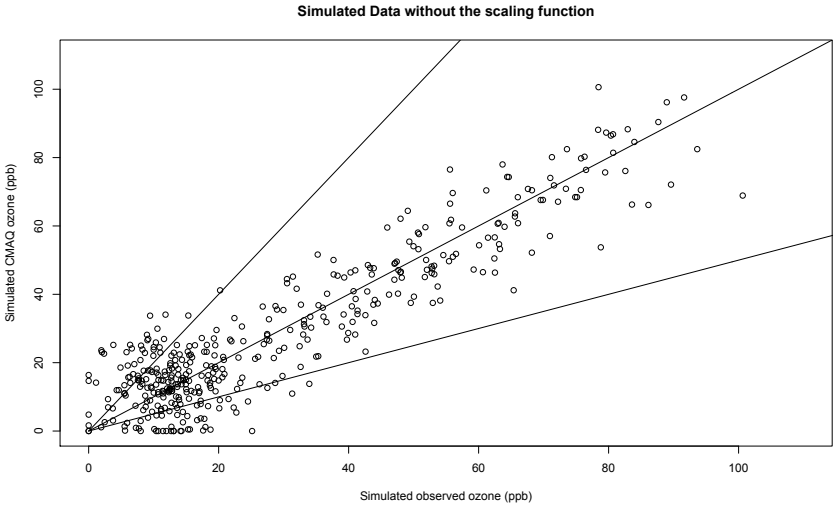


Figure A.8: Plot of simulated CMAQ vs. simulated observation, where the CMAQ data are simulated without the use of scaling function $f_s[\delta(x, y, h)]$.

Appendix B

Appendix Related to Chapter 3

B.1 Analyses of Simulated (Synthetic) Ozone Field

This appendix section presents the simulation-based analysis that determines the number of *meaningful* ozone features, this analysis was quickly summarized in Section 3.3. This simulation analysis is based on the repeated generations of simulated (synthetic) CMAQ data described in Appendix A.1, and it proceeds as follow:

1. Generate two sets of synthetic space-time CMAQ outputs: one dataset is used as the "modelling set" and the other the "testing set", which I denote as $\mathbf{O}_{model}$ and $\mathbf{O}_{test}$.

2. The "modelling set" is subjected to PCA, and the output EOFs and PCs are used to build ozone predictions for $\mathbf{O}_{test}$ via (3.3): $\sum_{j=1}^{p} \mathbf{P}_j * \mathbf{E}_j^{\mathrm{T}}$ for $p = 1$, $p = 2$ and so forth. That is, make predictions for $\mathbf{O}_{test}$ using increasing number of $\mathbf{E}_j$'s and $\mathbf{P}_j$'s decomposed form $\mathbf{O}_{model}$. Moreover, these simulations are generated with dimension $n \times t$. This step also shortens the simulation time: $\mathbf{O}_{model}^{\mathrm{T}}\mathbf{O}_{model}$ has dimension $t \times t$ instead of $n \times n$, where it is usually $n > t$ in a synthetic data.

3. Predictions with $p = 1, \ldots, 6$ are evaluated against the simulated "testing set" $\mathbf{O}_{test}$ and their *Root Mean Squared Errors* (RMSEs) are calculated.

4. Repeat the above 3 steps. I chose a repetition size of 500.

This simulation exercise is designed around the idea that ozone data simulated using exactly the same simulation parameters are multiple realizations of the same process. In other words, these simulated CMAQ fields are driven by a common ozone formation-circulation mechanism, and any differences between simulations are purely due to random noise. Adding more EOFs to (3.3) increases the prediction quality for the *modelling set* as more underlying ozone feature and patterns (data components) are used. However, notice that the EOFs from one simulation are used in (3.3) to predict another simulation, and it can be argued that after a certain point, the act of adding EOFs into (3.3) will cease to be beneficial and the prediction quality for the *testing set* will deteriorate. Since EOFs are noise after a few spatially or temporally "meaningful" ones, adding additional EOFs is tantamount to using one set of noise to predict another set of noise, and the modelling quality subsequently suffers. The transition between "beneficial" to "detrimental" is the p value to choose in (3.3), as it signals that additional EOFs no longer represent useful ozone features.

Each simulation produces 6 RMSEs, denoted as RMSE_j, $j = 1, \ldots, 6$. With a simulation size of 500, there are 500 sets of $(\text{RMSE}_1, \ldots, \text{RMSE}_6)$, and subsequently the differences $(\text{RMSE}_1 - \text{RMSE}_2, \ldots, \text{RMSE}_5 - \text{RMSE}_6)$. I use these $RMSE_j - RMSE_{j+1}$ samples to determine the number of "useful" EOFs/PCs. Figure B.1 shows the histograms of each $RMSE_j - RMSE_{j+1}$ and Table B.1 shows the 95% confidence intervals of the means of $RMSE_j - RMSE_{j+1}$'s, which are calculated as

$$\text{Sample Mean}(RMSE_j - RMSE_{j+1}) \pm 1.96 \cdot \frac{\text{Sample Std.deviation}(RMSE_j - RMSE_{j+1})}{\sqrt{n}}.$$

The histograms in Figure B.1 are approximately *Normally distributed*, hence I used the multiplier ± 1.96 that defines a 95% confidence region of a Normal distribution. As shown, the $RMSE_j - RMSE_{j+1}$ start to become negative at $j = 3$. The range of $RMSE_j - RMSE_{j+1}$ is completely negative at $p = 4$, implying that a model with 4 EOFs is worse than that with 3 EOFs.

Using our simulated CMAQ fields as the reference data, analyses in this subsection show that at $p = 3$, EOFs/PCs begin to capture less-structured patterns. From $p = 4$ onward, the EOFs/PCs extracted from our simulated CMAQ fields begin to be dominated by random noise.

	Lower bound	Upper bound
$RMSE_1 - RMSE_2$	0.09	0.24
$RMSE_2 - RMSE_3$	-0.03	0.11
$RMSE_3 - RMSE_4$	-0.32	-0.19
$RMSE_4 - RMSE_5$	-0.31	-0.18
$RMSE_5 - RMSE_6$	-0.28	-0.18

Table B.1: Table of estimated 95% confidence intervals for the means of $RMSE_j - RMSE_{j+1}$'s.

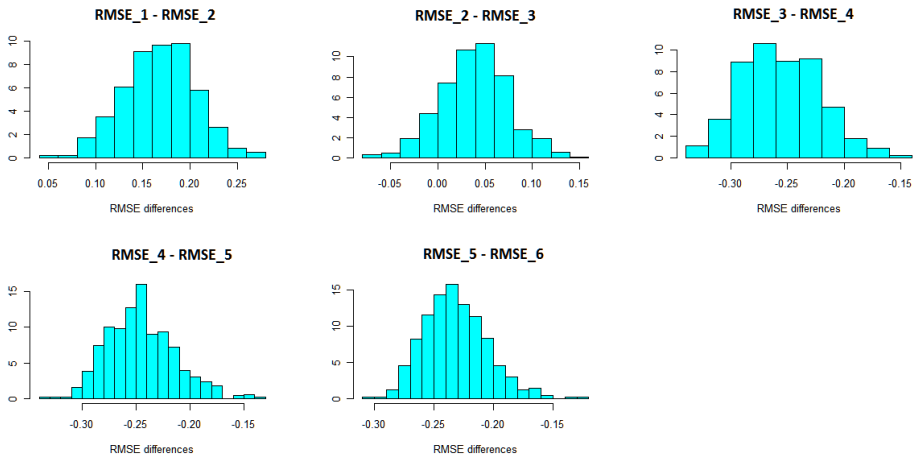


Figure B.1: Histograms of $RMSE_i - RMSE_{i+1}$ from simulation.

B.2 Plots of $\mathbf{E}_j$ from Column-centered Ozone Data and Rotated-$\mathbf{E}_j$

Figure B.2 compares the $\mathbf{E}_j$ (left), $j = 2, 3, 4$, decomposed from $\mathbf{O}_{t\times n}$ to $\mathbf{E}_j^{centre}$ from column-entered ozone data of orders $j = 1, 2, 3$. Figure 3.9 in

Section 3.4 showed that $\mathbf{E}_1$ from original data captures the spatial variation of LFV's temporal ozone mean. Subtracting the mean from each location results in $\mathbf{E}_1^{centre}$ no longer capturing said mean feature. As Figure B.2 shown, spatial features captured by $\mathbf{E}_j^{centre}$ is similar to $\mathbf{E}_{j-1}$. Hence, the PCA of column centred data extracted similar feature as PCA of original data, albeit the feature rank j is one order lower.

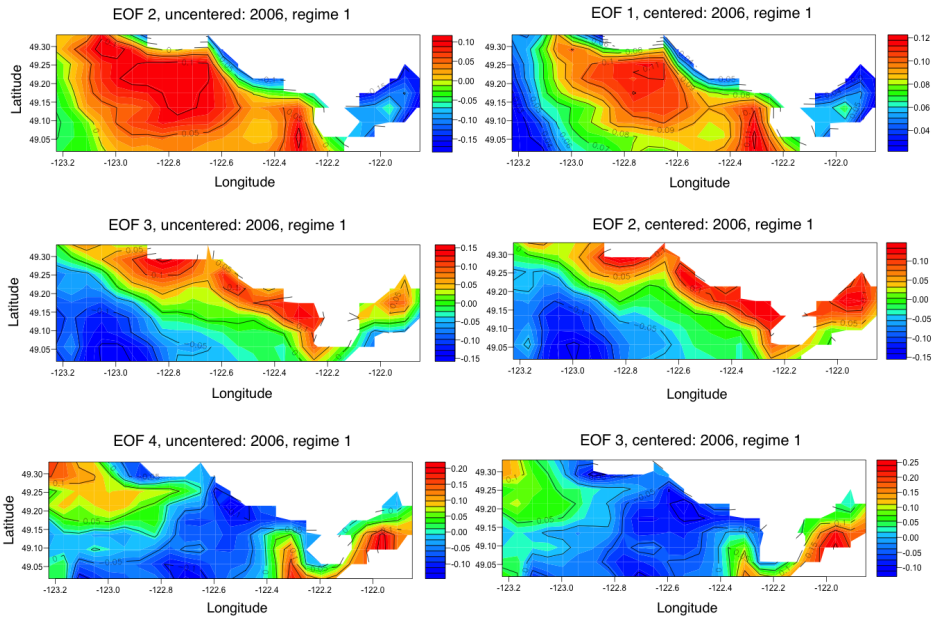


Figure B.2: Comparison plots between the $\mathbf{E}_j$ from the PCA of original $\mathbf{O}_{t\times n}$ (left) and column-centered ozone data (right). The EOF orders are $j = 2, 3, 4$ for the original data and $j = 1, 2, 3$ for the centered data. The ozone data is the CMAQ output for the entire 96 hours of the 2006 episode.

Figures B.3 shows the time-series of hourly LFV ozone means and standard deviations of the 2006 CMAQ output, along with $\mathbf{P}_j^{centre}$, $j = 1, \ldots, 4$, obtained from the PCA of column-centered data (the 2006 CMAQ output). Figure B.4 shows the same-ordered $\mathbf{P}_j$ from the original data. The 1st-order temporal feature captures the time-series pattern of the hourly LFV mean with PC values ranging from negative to positive. This shows that the subtraction of the mean field will retain the structure of the hourly spatial

ozone mean. The space-time feature $\mathbf{P}_1^{centre}(\mathbf{E}_1^{centre})^{\mathrm{T}}$ is not the dynamic east-west contrast captured by $\mathbf{P}_2\mathbf{E}_2^{\mathrm{T}}$ (Section 3.4): it shows the pattern of $\mathbf{E}_1^{centre}$ scaled positive during the daytime and negative at night, where the magnitude of ozone values (in both negative and positive direction) are higher for the western LFV. The dynamic patterns of $\mathbf{P}_j^{centre}(\mathbf{E}^{centre})_j^{\mathrm{T}}$, $j \geq 2$, reflects those of $\mathbf{P}_j\mathbf{E}_j^{\mathrm{T}}$ at $j \geq 3$.

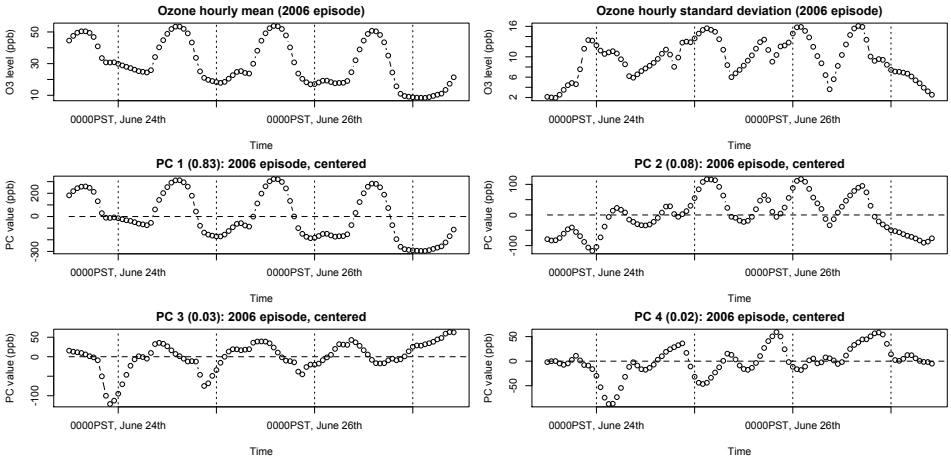


Figure B.3: From the PCA of centered ozone data: plots of hourly LFV ozone mean, standard deviation and $\mathbf{P}_j$, $j = 1, \ldots, 4$. The number in PC plot headings indicate the proportion of data variation each feature recovers. The ozone data is the CMAQ output for the entire 96 hours of the 2006 episode.

Figure B.5 compares the original $\mathbf{E}_j$, $j = 1, \ldots, 4$, to VARIMAX rotated EOFs of the same order. The EOF rotation is done for $\mathbf{E}_j$, $j = 1, \ldots, 4$, simultaneously; the $\mathbf{E}_j$'s of orders $j \geq 5$ remain unrotated. As shown, the rotated-$\mathbf{E}_1$ captures a spatial ozone pattern with region of positive spatial weights around the middle LFV with maximum at Maple Ridge, and areas of negative spatial weights at two edges of LFV. When multiplied by corresponding $\mathbf{P}_1$, which is strictly positive, the dynamic spatial pattern (not shown) captures a daily peak around middle of LFV during afternoon (1400PST to 1600PST) and negative peak at edges of LFV during the same afternoon hours. After rotations, EOFs of orders $j = 2, 3, 4$ revealed a sim-

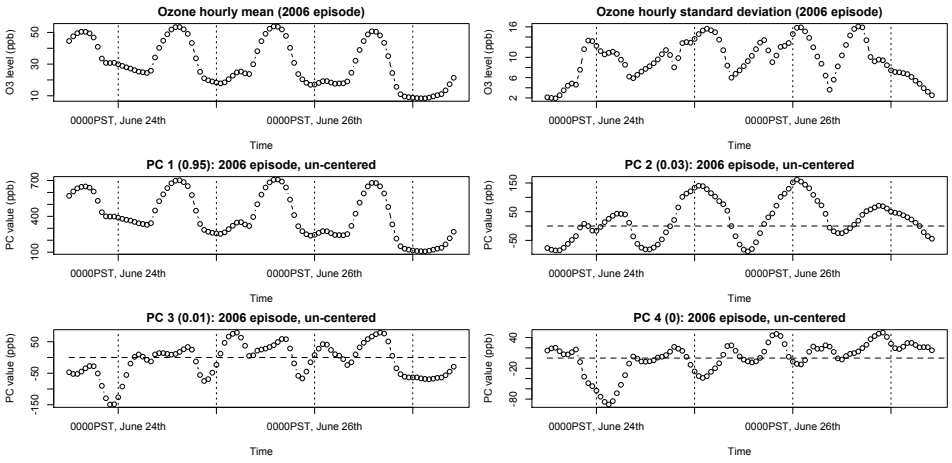


Figure B.4: From PCA of original ozone data: plots of hourly LFV ozone mean, standard deviation and $\mathbf{P}_j$, $j = 1, \ldots, 4$. The number in PC plot headings indicate the proportion of data variation each feature recovers. The ozone data is the CMAQ output for the entire 96 hours of the 2006 episode.

ilar spatial features as un-rotated $\mathbf{E}_j$'s (Figure B.5). We also experimented with data from other episodes, data with different sized spatial domains and various orders of VARIMAX rotation (how many $\mathbf{E}_j$ to rotate). It was found that for the particular LFV ozone data under analysis, rotated EOF did not provide clear advantage in interpretability compared to regular EOFs.

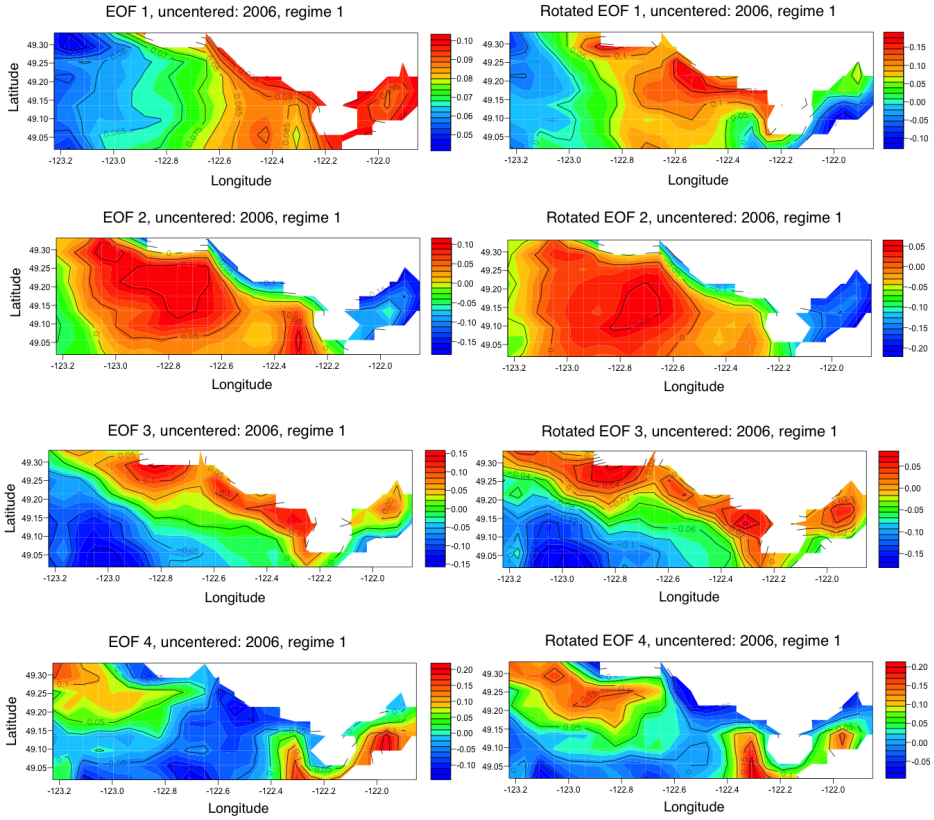


Figure B.5: Comparison plots between the original (left) and VARIMAX rotated (right) $\mathbf{E}_j$, $j = 1, \ldots, 4$. The VARIMAX rotation is done for the 1st 4 $\mathbf{E}_j$'s only. The dataset is the CMAQ output for the entire 96 hours of the 2006 episode.

Appendix C

Appendix Related to Chapter 4

C.1 PCA of Meteorological and Chemical Precursor Variables

This section shows the spatial-temporal ozone means and standard deviations, as well as PCA outputs $\mathbf{E}_j$'s and $\mathbf{P}_j$'s of ozone model variables. They are space-time data of: temperature, wind speed, planetary boundary layer height, NOx and VOC emission rates and antecedent concentrations.

- **NOx**: The 1st EOF captures the structures and variations of both the field of means and standard deviations of NOx emission (Figure C.1a). The 3rd EOF shows a well-defined spatial structure, albeit with less interpretability. The 2nd and 4th EOFs both have spatial fields that show little variation. The 1st PC reflects the temporal patterns of both hourly spatial means and standard deviations (Figure C.1b). Note the double peak of NOx production at morning and afternoon, illustrating how emissions are distributed inside SMOKE. The first PC alone accounts for about 98% of data variation, although the higher-order PCs still represent noticeable temporal patterns.

- **Temperature**: The 1st and 2nd EOF capture the spatial patterns of the episodic means and standard deviations (averaged across 96-hours of the episode, Figure C.2a). The 3rd EOF shows a spatial pattern reflecting the topography of my self-defined LFV: it captures the spatial

contrast in temperature between the "valley floor" and the mountainous region. The first 3 PCs show distinctive diurnal hourly patterns. It seems that the 2nd and 3rd PCs each capture a specific feature of the hourly spatial standard deviation of temperature (Figure C.2b).

- **Wind speed**: The 1st EOF and 2nd EOF have near identical spatial distributions to the data means and standard deviations (Figure C.3a). Like temperature, the 3rd EOF shows a spatial pattern akin to the topography of the LFV. The 1st PC as usual, reflects the time-series of the hourly wind speed averaged across space (Figure C.3b). When examined closely, the 2nd PC is inversely related to the hourly pattern of spatial standard deviation. The later PCs also capture clear temporal structures. Together, the first 3 EOFs/PCs are responsible for over 95% of data variation.

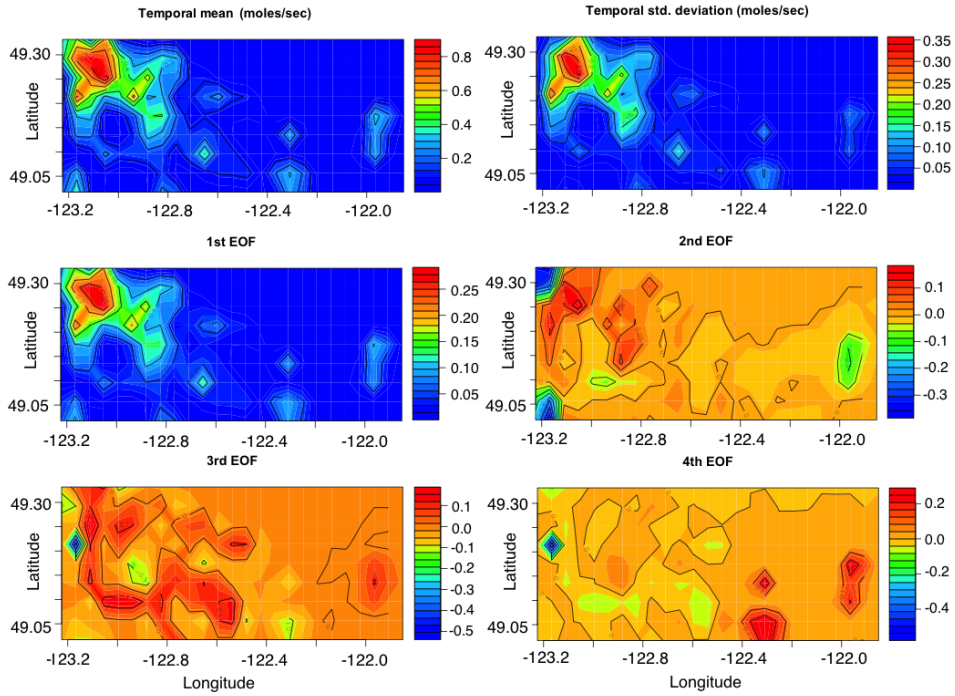
- **Boundary layer height**: The spatial fields of means and standard deviations have patterns that are very similar, both of which are captured by the 1st EOF (Figure C.4a). The diurnal patterns of the 1st PC and the hourly spatial means (Figure C.4b) closely resemble the 1st-order temperature feature (Figure C.2b). Higher order EOFs and PCs also exhibit discernible spatial and temporal patterns.

- **Antecedent NOx**: As with NOx emission, the spatial fields of episodic means and standard deviations have similar pattern, and it is captured by the 1st EOF (Figure C.5a). The diurnal patterns of hourly spatial means and standard deviations are similar: they peak early in the morning, then decreases significantly during the daytime before recovering late in the afternoon (Figure C.5b). This indicates that the peak of photochemical reaction, thus NOx consumption, takes place during daytime, whereas morning and night-time are times for NOx deposition. As with all aforementioned model variables, the higher-

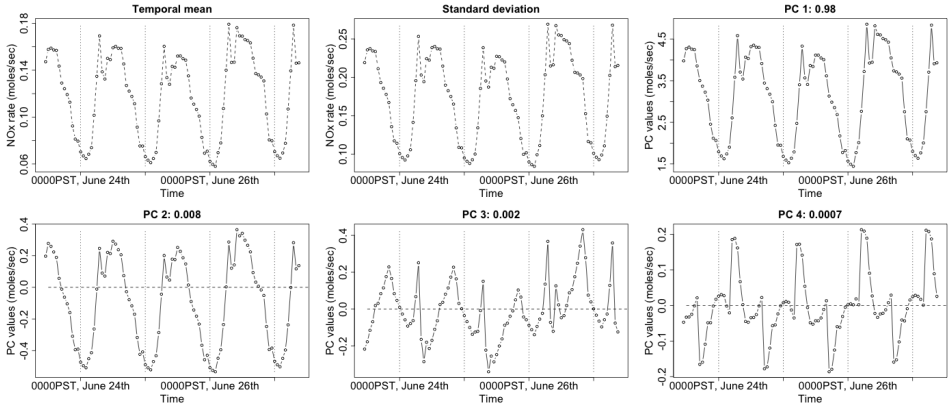
order data features decomposed from the antecedent NOx data display strong spatial and temporal structures.

- **VOC emission rate and antecedent VOC**: The conclusions are similar to those for NOx emission and antecedent NOx concentration (Figures not shown).

As one might expect, without centering (or standardizing) the space-time data, the spatial/temporal mean and standard deviation become the dominant features. Once again, I do not perform centering or standardization on the data because my goal is to analyze the most important data features. For all model variables, the first 3 EOF-PC pairs capture over 90%, or in certain cases, close to 100% of data variation. Hence the *full covariate set* of ozone feature models contain the first 3 EOFs/PCs of all 7 variables.

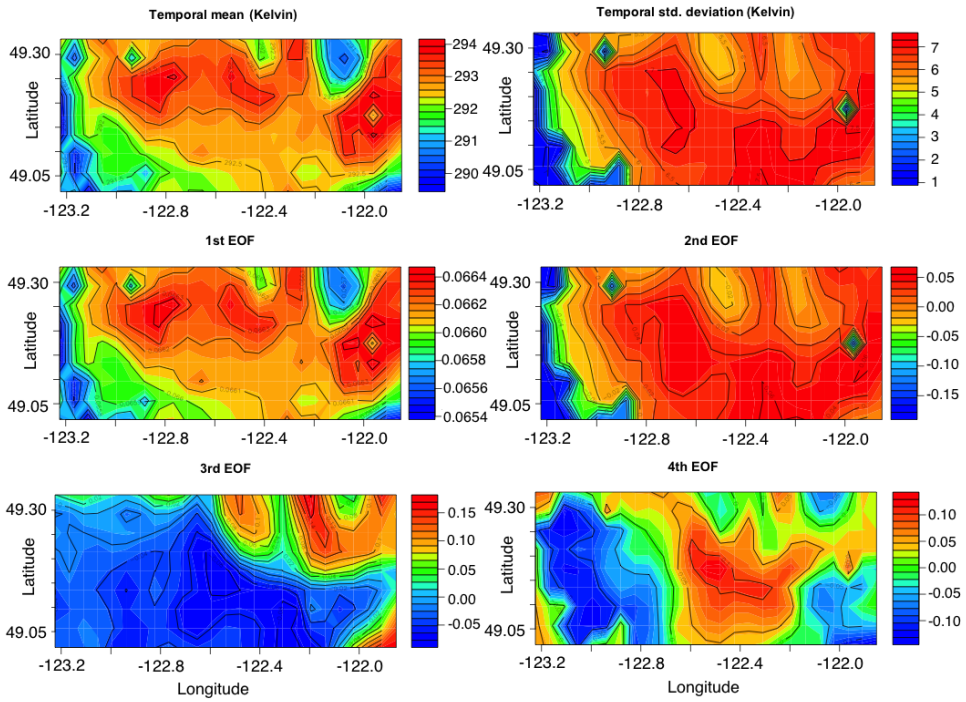


(a) From the NOx emission data for the 2006 CMAQ ozone (SMOKE output): plots of the spatial fields of means, standard deviations and the first four EOFs.

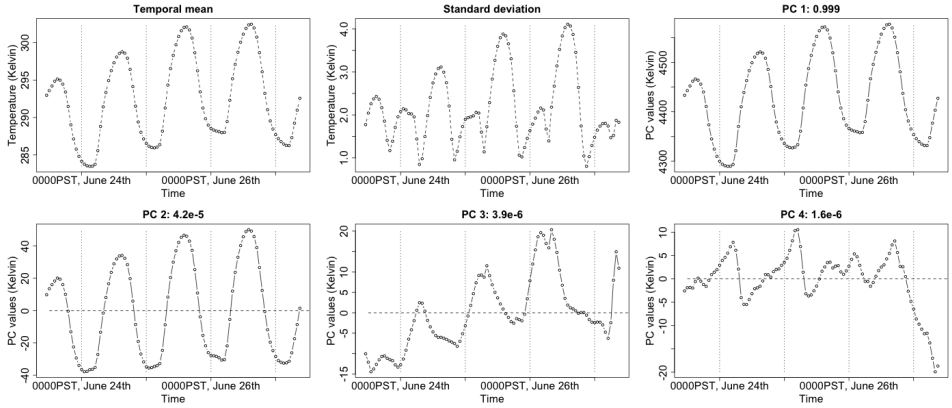


(b) From the NOx emission data for the 2006 CMAQ ozone (SMOKE output): plots of hourly spatial means, standard deviations and the first four PCs. The value in the PC plot heading indicate the proportion of data variation explained. The dashed line indicate the hour 0000 of each day.

Figure C.1: Spatial and temporal feature plots of NOx emission rates associated with the 2006 CMAQ output

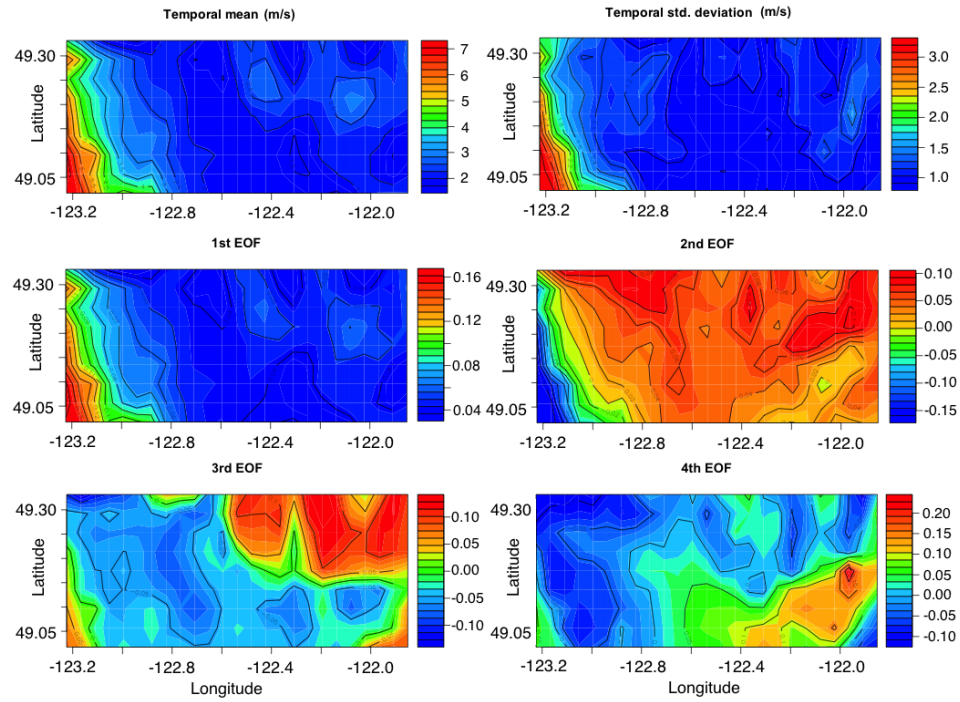


(a) From the temperature data for the 2006 CMAQ ozone (WRF output): plots of the spatial fields of means, standard deviations and the first four EOFs.

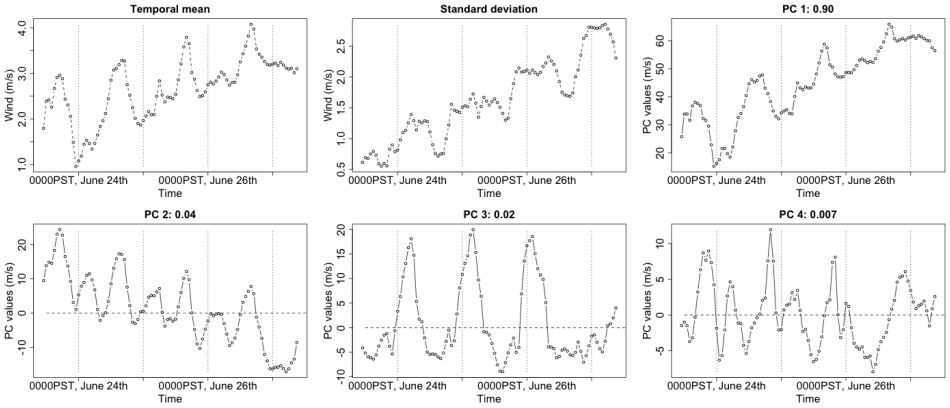


(b) From the temperature data for the 2006 CMAQ ozone (WRF output): plots of hourly spatial means, standard deviations and the first four PCs. The value in the PC plot heading indicate the proportion of data variation explained. The dashed line indicate the hour 0000 of each day.

Figure C.2: Spatial and temporal feature plots of temperature associated with the 2006 CMAQ output

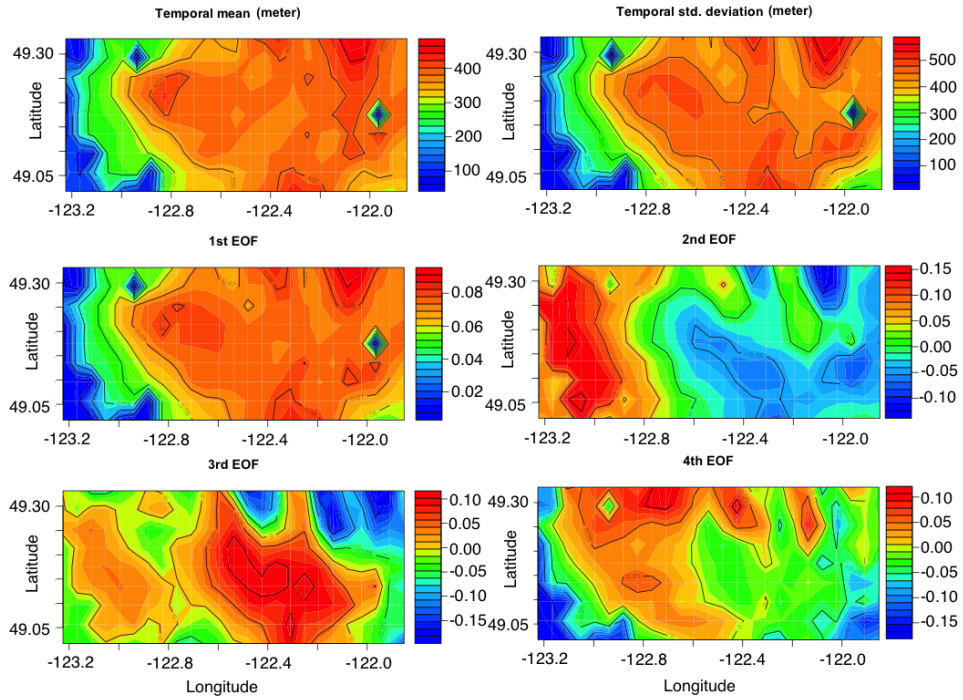


(a) From the wind speed data for the 2006 CMAQ ozone (WRF output): plots of the spatial fields of means, standard deviations and the first four EOFs.

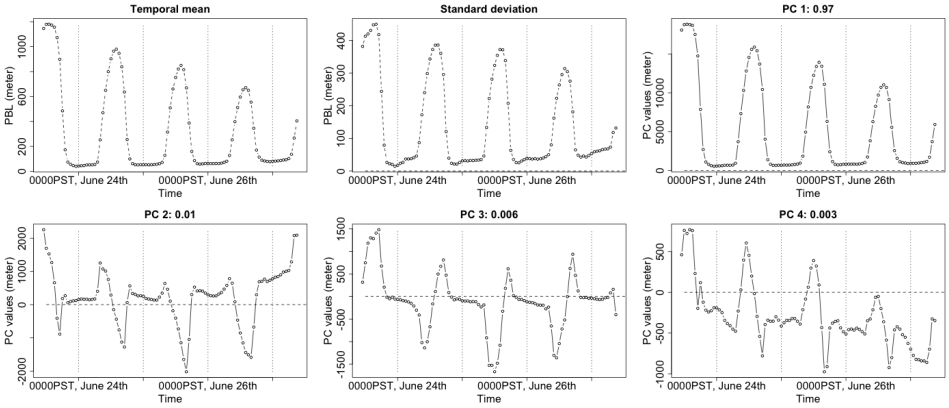


(b) From the wind speed data for the 2006 CMAQ ozone (WRF output): plots of hourly spatial means, standard deviations and the first four PCs. The value in the PC plot heading indicate the proportion of data variation explained. The dashed line indicate the hour 0000 of each day.

Figure C.3: Spatial and temporal feature plots of the wind speed associated with the 2006 CMAQ output

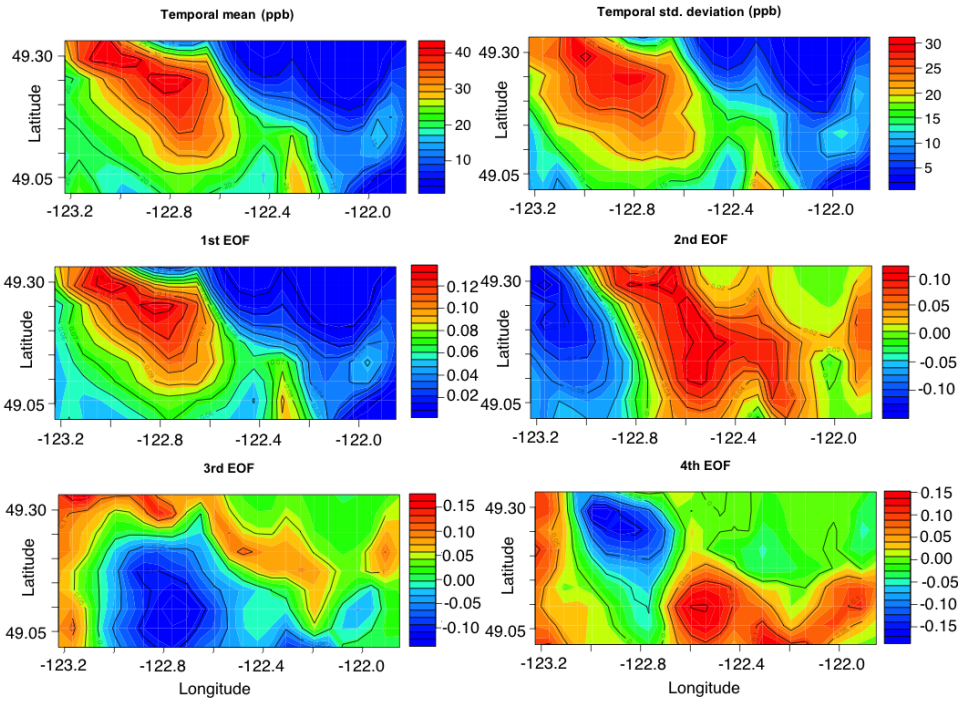


(a) From the data of BL height for the 2006 CMAQ ozone (WRF output): plots of the spatial fields of means, standard deviations and the first four EOFs.

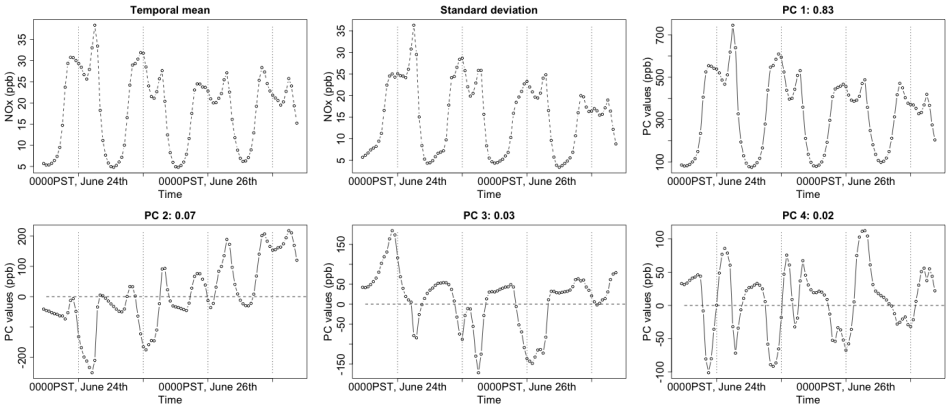


(b) From the data of BL height for the 2006 CMAQ ozone (WRF output): plots of hourly spatial means, standard deviations and the first four PCs. The value in the PC plot heading indicate the proportion of data variation explained. The dashed line indicate the hour 0000 of each day.

Figure C.4: Spatial and temporal feature plots of the boundary-layer (BL) height associated with the 2006 CMAQ output.



(a) The antecedent NOx concentration data associated with the 2006 CMAQ output: plots of the spatial fields of means, standard deviations and the first four EOFs.



(b) The antecedent NOx concentration data associated with the 2006 CMAQ output: plots of hourly spatial means, standard deviations and the first four PCs. The value in the PC plot heading indicate the proportion of data variation explained. The dashed line indicate the hour 0000 of each day.

Figure C.5: Spatial and temporal feature plots of the antecedent NOx concentration data associated with the 2006 CMAQ output.

C.2 Prediction Bias of Feature-Based Ozone Model

Table 4.5 showed *Mean Percentage Errors*. Due to not taking the absolute values of prediction residuals, positive errors offset the negative errors. Hence MPE can be viewed as a rough measurement of prediction *bias*. The MPE results show that the ozone predictions are close to unbiased during the day-time, but the bias becomes more pronounced towards the night-time. At 1900PST, the *mean prediction residual* is -5 *ppb* for the CII model and -7 *ppb* for the VM model. The residual means for both models gradually increase as the night progresses, while they are close to 0 during earlier hours.

In addition to the discussed "lack-of-accuracy in night-time PC predictions", I suspect there is also the issue of *ozone prediction bias*. As discussed, the predictions of each *ozone features* are *unbiased* owing to the application of BLUP. However, as I will now show, the use of equation (4.13) entails that the *ozone field* prediction is biased.

Let $E_{i,j}$ denote i-th element of $\mathbf{E}_j$ and $P_{h,j}$ be h-th element of $\mathbf{P}_j$, hence the ozone prediction at the i-th location and h-th hour of my self-defined modelling region is

$$\hat{O}_{h,i} = \sum_{j=1}^{p} \hat{P}_{h,j}\hat{E}_{i,j},$$

$$\text{where } p = 3,\ h = 1, \ldots, 24 \text{ and } i = 1, \ldots, 229.$$

Here, the "hat" notation indicates an estimate. The issue is that $\hat{O}_{h,i}$ is a *statistically biased* prediction of $O_{h,i}$. This conclusion can be verified by

expanding the covariance equation between $E_{i,j}$ and $P_{h,j}$:

$$\begin{aligned}
\mathrm{Cov}\,(P_{h,j}, E_{i,j}) &= \mathrm{E}\,[(P_{h,j} - \mathrm{E}\,(P_{h,j}))(E_{i,j} - \mathrm{E}\,(E_{i,j}))] \\
&= \mathrm{E}\,(P_{h,j}E_{i,j}) - \mathrm{E}\,(P_{h,j})\mathrm{E}\,(E_{i,j}) \;\Rightarrow \\
\mathrm{E}\,(P_{h,j}E_{i,j}) &= \mathrm{E}\,(P_{h,j})\mathrm{E}\,(E_{i,j}) + \mathrm{Cov}\,(P_{h,j}, E_{i,j}) \text{ and similarly} \\
\mathrm{E}\,(\hat{P}_{h,j}\hat{E}_{i,j}) &= \mathrm{E}\,(\hat{P}_{h,j})\mathrm{E}\,(\hat{E}_{i,j}) + \mathrm{Cov}\,(\hat{P}_{h,j}, \hat{E}_{i,j})
\end{aligned}$$

I have established unbiasedness of ozone feature predictions, $\mathrm{E}\,(\hat{E}_{i,j}) = \mathrm{E}\,(E_{i,j})$ and $\mathrm{E}\,(\hat{P}_{h,j}) = \mathrm{E}\,(P_{h,j})$. Therefore, space-time ozone predictions $\hat{\mathbf{P}}_j\hat{\mathbf{E}}_j^{\mathrm{T}}$'s become unbiased predictors by adding a correction term:

$$\mathrm{Cov}\,(P_{h,j}, E_{i,j}) - \mathrm{Cov}\,(\hat{P}_{h,j}, \hat{E}_{i,j}).$$

That is, the covariances of all EOF-PC pairs between elements of $\mathbf{E}_j$'s and $\mathbf{P}_j$'s *subtract* by covariances of their corresponding predictions. The resultant unbiased ozone prediction function is

$$\{\hat{O}_{h,i}\}_{\text{unbiased}} = \sum_{j=1}^{p}\{\hat{P}_{h,j}\hat{E}_{i,j} - \mathrm{Cov}\,(\hat{P}_{h,j}, \hat{E}_{i,j}) + \mathrm{Cov}\,(P_{h,j}, E_{i,j})\}, \quad \text{(C.1)}$$

and it can be shown that for each j,

$$\begin{aligned}
\mathrm{E}\,[(\hat{P}_{h,j}\hat{E}_{i,j})_{\text{unbiased}}] &= \mathrm{E}\,(\hat{P}_{h,j}\hat{E}_{i,j}) - \mathrm{Cov}\,(\hat{P}_{h,j}, \hat{E}_{i,j}) + \mathrm{Cov}\,(P_{h,j}, E_{i,j}) \\
&= \mathrm{E}\,(\hat{P}_{h,j})\mathrm{E}\,(\hat{E}_{i,j}) + \mathrm{Cov}\,(P_{h,j}, E_{i,j}) \\
&= \mathrm{E}\,(P_{h,j}E_{i,j}).
\end{aligned}$$

For any j-th EOF/PC, $\mathrm{Cov}\,(P_h, E_i)$ is not the covariance between location i and time h. Rather, it is the covariance between the i-th element of an EOF and h-th element of the corresponding PC: it measures the covariance between the spatial feature at a particular location $E_{i,j}$ and the temporal feature at a particulate hour $P_{h,j}$. Similar interpretation can be made for the covariance $\mathrm{Cov}\,(\hat{P}_{h,j}, \hat{E}_{i,j})$. Using the data at hand, I devised a way of estimating the covariances between all $24 \times 229 \times 3 = 16488$ pairs

of $\{P_{h,j}, E_{i,j}\}$ and corresponding $\{\hat{P}_{h,j}, \hat{E}_{i,j}\}$.

Re-sampling Method

Given ozone data of dimension $t \times n$, one can only extract one pair of $\{P_{h,j}, E_{i,j}\}$. To obtain a sample for any pair of $\{P_{h,j}, E_{i,j}\}$, one may repeatedly sample from the complete $t \times n$ data to construct subsamples. For each subsample, decompose and extract $\{P_{h,j}, E_{i,j}\}$. The detailed implementation preceeds as follow:

1. Create a *subsample* from the complete $t \times n$ ozone data. To create a sample of any particular pair of $\{P_{h,j}, E_{i,j}\}$, *include dataset on location i at hour h in every subsample.* Remember that after the EOF-decomposition of $\mathbf{O}_{t\times n}$, row i in $\mathbf{E}$ corresponds to the EOFs (summarized spatial data) for location i, and row h of $\mathbf{P}$ is the PC at hour h. By decomposing a subsample, I can acquire $E_{i,j}$ and $P_{h,j}$, for all $j = 1, \ldots, 3$, from the appropriate rows of $\mathbf{E}_{\text{subsample}}$ and $\mathbf{P}_{\text{subsample}}$. Hence from one subsample, I obtain exactly one EOF-PC pair $\{P_{h,j}, E_{i,j}\}$ between location i and h at $j = 1, 2, 3$.

   Repeating the above for multiple subsamples that include the data point at location i and hour h, I obtain a *sample* of $\{P_{h,j}, E_{i,j}\}$. Sample covariance is then used for estimating all corresponding Cov $(P_{h,j}, E_{i,j})$'s at $j = 1, 2, 3$.

2. Repeat step 1 for every $i = 1, \ldots, n$ and $h = 1, \ldots, t$. In the end, I have the sample covariances of all $(P_{h,j}, E_{i,j})$'s , $i = 1, \ldots, n$, $h = 1 \ldots, t$ and $j = 1, 2, 3$. Here, I let $n = 229$ (defining the the LFV modelling region), and $t = 24$, based on the assumption that EOF-PC correlation is a type of diurnal behaviour.

   The dimension of each resampled-data and the number of repetitions are determined through experimentation.

The 24×229 sampled ozone data are obtained by separating the 48-hour ozone *training set* into two daily 24×229 data, then average them across

the *same* hour of the day h and location i.

Following the procedure, I can collect all $3 \times 229 \times 24$ covariance estimates into $j = 3$ number of 24×229 "EOF-PC covariance matrix", where element (h, i) in j-th matrix contain the sample estimation of Cov $(P_{h,j}, E_{i,j})$.

For estimating the covariance Cov $(\hat{P}_{h,j}, \hat{E}_{i,j})$, I propose to implement the aforementioned sampling method on the 24×229 data of ozone forecasts: $\sum \hat{\mathbf{P}}_j \hat{\mathbf{E}}_j^{\mathrm{T}}$. Subsequently, one may collect the estimated covariances into corresponding $j = 3$ covariance matrices of Cov $(\hat{P}_{h,j}, \hat{E}_{i,j})$, where $i = 1, \ldots, 229$ and $h = 1, \ldots, 24$.

In the end, by adding the two types of covariance matrices to matching $\mathbf{P}_j \mathbf{E}_j^{\mathrm{T}}$ via (C.1), one obtains sample based *covariance-corrected* ozone predictions.

Results of Correcting Ozone Prediction Bias

As the analyses below show, biased prediction is *not* an issue during the all-important day-time ozone modelling. During night-time modelling, covariance-corrected predictions do induce a level of "localized spatial variations" at the right locations. However, at a regional level, these fine spatial variations introduce a fair amount of noise into the modelled ozone field.

First, I made space-time ozone forecasts using only the first two EOFs and PCs: $\hat{\mathbf{O}}_{t \times n} = \hat{\mathbf{P}}_1 \hat{\mathbf{E}}_1^{\mathrm{T}} + \hat{\mathbf{P}}_2 \hat{\mathbf{E}}_2^{\mathrm{T}}$. As the results at 1400PST in Figure C.6 shown, with only 2 predicted space-time features and no bias correction, the daytime ozone fields are still well-forecasted. The bottom plots in Figure C.6 show that a noticeable lack of prediction quality emerges at night-time. This is because the $\mathbf{P}_1 \mathbf{E}_1^{\mathrm{T}}$ and $\mathbf{P}_2 \mathbf{E}_2^{\mathrm{T}}$ capture the underlying spatial-temporal mean structures and "daytime" ozone features.

Also shown in Figure C.6, are predictions from two types of improvement scheme: (1) adding the 3rd space-time component $\hat{\mathbf{P}}_3 \hat{\mathbf{E}}_3^{\mathrm{T}}$, and (2) apply the bias-correction covariance matrices. The first method gives the model I had in Section 4.7. As for the bias-correction method, there is evidence that by correcting for prediction bias, the night-time prediction quality is improved by the addition of detailed spatial ozone variations at appropriate

locations. However, the improvement is not as drastic as the method 1: using additional ozone feature forecasts $\hat{\mathbf{P}}_3\hat{\mathbf{E}}_3^{\mathrm{T}}$ (Figure C.6). Furthermore, I am inclined to conclude that, since little prediction bias exists during the daytime, there is no practical reason for the application of covariance-based correction terms.

In summary, the addition of $\hat{\mathbf{P}}_3\hat{\mathbf{E}}_3^{\mathrm{T}}$ into the space-time ozone forecast $\hat{\mathbf{O}}_{t\times n}$ improves the quality of night-time prediction *without* inducing unnecessary "local spatial noise" into the modelled ozone fields. These results indicates that, the statistical ozone model's issue with night-time modelling is attributed more to not modelling higher-order ozone features; less to the presence of prediction bias. In the end, I believe the issue of prediction bias should be approached given one's particular research focus. In my case, the importance of day-time ozone (or daily 8-hour maximum) coupled with relevant modelling results inform my decision to forgo the application of EOF-PC covariances for bias-correction.

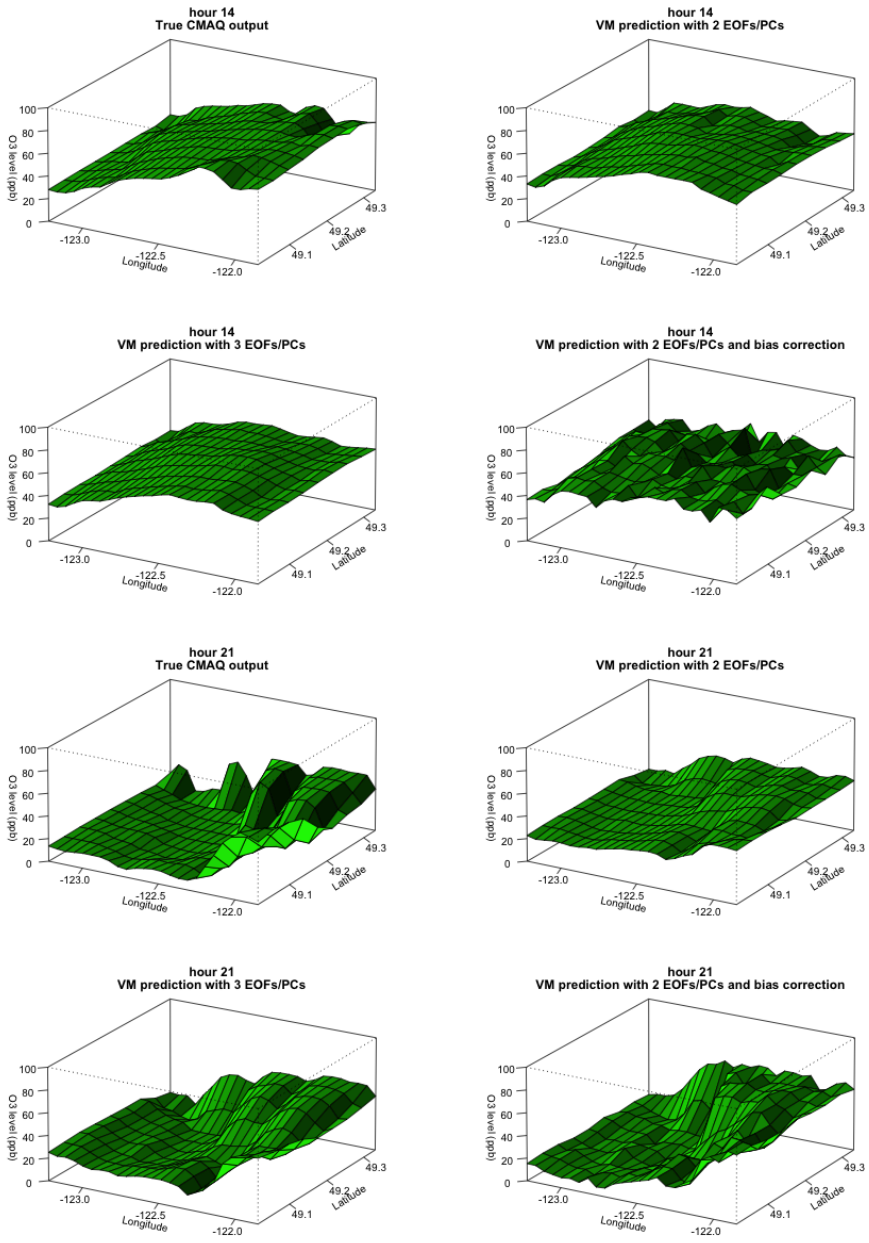


Figure C.6: Hours 1400 and 2100 of June 26th, 2006 (the predictive set): 3-D spatial ozone fields of *true* CMAQ output (upper-left), prediction made by the VM model with 2 and 3 sets of predicted EOFs/PCs (upper-right and lower-left), the lower-right plots shows the *bias-corrected* version of the VM model prediction made with 2 sets of EOFs/PCs.