ENSEMBLE-AVERAGED, PROBABILISTIC, AND KALMAN-FILTERED REGIONAL OZONE FORECASTS

by

LUCA DELLE MONACHE

Laurea in Mathematics, Universitá degli Studi di Roma "La Sapienza", Italy, 1997
M.S., The San José State University, California, 2002

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

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Atmospheric Science)

THE UNIVERSITY OF BRITISH COLUMBIA

November 2005

Abstract

This thesis investigates the hypothesis that ensemble methods and Kalman-filter (KF) post-processing can be utilized to improve near-surface real-time ozone forecasts.

The ensemble approach combines multiple forecasts to yield ensemble-averaged and probabilistic predictions. In non-linear systems such as the atmosphere, it is well established that the ensemble approach provides a better estimate of future evolution than a deterministic forecast. This approach is extended here for ozone forecasts.

KF post-processing is applied to remove ozone-forecast bias; i.e., systematic errors. In this dissertation, the filter is applied in a predictor mode to the raw ozone forecasts from the Community Multiscale Air Quality (CMAQ) 3-D numerical model.

An ozone ensemble-forecast system based on a multi-model approach has been analyzed. Moreover, a new ensemble design for air-quality forecasts has been proposed, based on both meteorology and emission perturbations. Ozone ensemble-averaged and probabilistic forecasts resulting from these ensemble methods have been realized and tested (introducing a new reliability index).

The following are the main findings of this thesis. An ozone ensemble-forecast system based on a multi-model approach produces an ensemble-averaged prediction more skillful than a single-model approach. Ensemble-averaging is able to compensate for some of the predictive-skill deficiencies in deterministic ozone forecasts, and for part of the initial-condition inaccuracy. In the new ensemble air-quality forecast system proposed, the meteorology perturbation is important to capture ozone temporal and spatial distributions. The emission perturbation is needed to accurately predict the ozone concentration magnitude. The emission perturbations are more important than the meteorology ones to capture high (and rarely measured) ozone concentrations.

The KF successfully removes part of the ozone-forecast bias caused by errors in the model. The combination of ensemble averaging (unsystematic-error removal) and Kalman filtering (systematic-error removal) results in the best ozone forecast.

Ensemble and KF methods can indeed significantly improve near-surface ozone forecasts, even in the complex coastal mountain setting of the Lower Fraser Valley. There are no intrinsic limitations to these methods that would prevent their application in real time to other pollutants in other geographic settings.

Table of Contents

List of Tables

List of Figures

Notation, Symbols, and Abbreviations

Symbols and abbreviations

Symbols and abbreviations are given in the following order:

1. lowercase Greek letters
2. uppercase Greek letters
3. Roman letters in alphabetical order
    (a) lowercase letters
    (b) uppercase letters
    (c) calligraphic letters
    (d) double-struck letters

List of symbols and abbreviations:

	Lowercase Greek
β	Kalman gain
ϵ	white noise of forecast error
λ_i	Lyapunov exponent of the i-th axis
η	white noise of true forecast bias
σ	Lorenz model parameter
σ_ϵ^2	ϵ variance
σ_η^2	η variance
τ	forecast initialization time in the lagged-averaged forecast
	Upper-case Greek
Δt	time lag in Kalman filter algorithm
	Roman
	A
ABL	atmospheric boundary layer
AC	additive bias correction
$ANATEX$	across North America tracer experiment

APL	agreement in percentile level
AQ	air quality
ATL	agreement in threshold level
	B
b	Lorenz model parameter
BC	British Columbia
BCs	boundary conditions
BIA	bias of FCT
	C
$C_{AC}(t, sta.)$	additive bias-corrected concentration at a monitoring station for hour t
$C_{MC}(t, sta.)$	multiplicative bias-corrected concentration at a monitoring station for hour t
$C_o(t, sta.)$	1-h average observed concentration at a monitoring station for hour t
$C_o(x, t_i)$	observed value at the monitoring station located at x for hour t_i
$\overline{C_o(sta.)}$	average of 1-h average observed concentrations at a monitoring station
$C_o(d., sta.)_{max}$	maximum 1-h average observed concentration at a monitoring station over one day
$C_o(x, t^{'})_{max}$	maximum 1-h observed concentration at a monitoring station located at x over one day
$C_p(t, sta.)$	1-h average predicted concentration at a monitoring station for hour t
$C_p(x, t_i)$	predicted value at the monitoring station located at x for hour t_i
$\overline{C_p(sta.)}$	average of 1-h average predicted concentrations at a monitoring station
$C_p(d., sta.)_{max}$	maximum 1-h average predicted concentration at a monitoring station over one day
$C_p(x, t^{'})_{max}$	maximum 1-h predicted concentration at a monitoring station located at x over one day
$CALGRID$	California photochemical grid model
CART	classification and regression tree schemes
$CBM-IV$	carbon bond mechanism
CCME	Canadian Council of Ministers of the Environment
$CMAQ$	models-3/community multiscale air quality model
$CRMSE$	centered root mean square error
CRT	control forecast member
$CTRL$	run with base emission scenario
CTM	Chemistry Transport Model
$CTMs$	Chemistry Transport Models
CWS	Canada wide standard
$CYVR$	Vancouver International Airport station
	E
E	forecasts ensemble average
$ECMWF$	European Center for Medium-range Weather Forecasts
EK	ensemble-mean of the Kalman filter bias-corrected forecasts

EMEP	European monitoring and evaluation programme
EPA	Environmental Protection Agency
ETEX	European tracer experiment
EURAD	European air pollution dispersion model
	F
FCT	operational forecast data
	G
GE	gross error
GEM	generalized environmental multiscale model
GVRD	Greater Vancouver Regional District
	H
H_2O_2	hydrogen peroxide
HIRLAM	high resolution limited area model
HNO_3	nitric acid
HYSPLIT	hybrid single-particle Lagrangian integrated trajectory model
	I
ICs	initial conditions
	K
K	Kalman-corrected forecast
KEK	Kalman filter bias-corrected ensemble-mean of the Kalman filter bias-corrected forecasts
KF	Kalman filter
KFP	Kalman filter predictor
	L
LFV	Lower Fraser Valley
LLLV	local leading Lyapunov vectors
LOTOS	long-term ozone simulation model
	M
MC	multiplicative bias correction
*MC*2	Canadian mesoscale compressible model
MCF	meteorological complexity factor
MEBI	modified Euler backward iterative chemistry solver
*MM*5	Penn State-NCAR mesoscale model
MMT	middle member of the ensemble trajectories
MOS	model output statistics
	N
N	number of hourly concentrations

N	number of ensemble members in the lagged-averaged forecast
N_{day}	number of days
N_{hour}	number of 1-h average concentrations
NAAQOs	national ambient air quality objectives
NAM	North American mesoscale model
$NCAR$	National Center for Atmospheric Research
$NCEP$	National Centers for Environmental Prediction
NMC	National Meteorological Center
NO	nitric oxide
NO_2	nitrogen dioxide
NO_y	total reactive nitrogen
$NOXN$	run with minus 50 % NO_x
$NOXP$	run with plus 50 % NO_x
NO_x	nitrogen oxides
NWP	numerical weather prediction
	O
O	atomic oxygen
O_2	molecular oxygen
O_3	ozone
$OEFS$	ozone enseble forecast system
	P
p	expected mean square error
PDT	Local Pacific Daylight Time
ppb	part per billion
ppbv	part per billion by volume
PDF	probability density function
PFS	probabilistic forecast system
	R
r	Lorenz model parameter
$rmsd$	root-mean-square deviation
$RMSE$	root mean square error
$REM3$	regional Eulerian model
ROC	Relative Operating Characteristics
ROGs	reactive organic gases
RPN	Recherche en Prévision Numérique
$RTMOD$	real time model evaluation
	S
$SCIPUFF$	second order closure integrated puff
$SEAREX$	sea-air exchange experiments
SEF	global spectral model
$SMOKE$	Sparse Matrix Operator Kernel Emission

$SNAP$	severe nuclear accident program
SO	space overlap
SSE	system simulation experiment
SV	singular-vectors
	T
t	time
	U
UBC	University of British Columbia
$UPPA$	unpaired peak predicition accuracy
US	United States
UTC	Coordinated Universal Time
$UW-NMS$	University of Wisconsin nonhydrostatic modeling system
	V
V	phase space volume
V_0	phase space initial volume
VOCs	volatile organic compounds
	W
WMO	World Meteorological Organization
	X
x	cartesian coordinate
$\hat{x}$	Kalman filter bias estimate
x_t	true forecast bias at time t
	Y
y	Cartesian coordinate
y_t	forecast error at time t
	Z
z	Cartesian coordinate
z_t	time series

Preface

The goal of this dissertation is to improve our ability to predict the spatial and temporal distribution of ozone concentration. This goal has been achieved by applying ensemble and Kalman-filter methods to air-quality (AQ) forecasting.

Some *dynamical systems* are called *chaotic* if they show *divergent* behavior, meaning that two different solutions starting from similar but not identical initial states would eventually diverge nonlinearly in solution space. The atmosphere exhibits this behavior, and is thus a chaotic system. As a consequence there is an upper limit in time on the predictive skill of weather forecasts. The ensemble approach is one method to represent the time evolution of the probability density function (PDF) describing the atmosphere's initial state and its uncertainty. This PDF can be represented by a limited set of points. The evolution of each of those points would be a member of the ensemble. Each of those members should ideally represent an equally likely evolution of the dynamical system.

It has been found for numerical weather prediction (NWP) that the ensemble-mean is more accurate that an individual model realization, when verified for many cases. The ensemble technique yields similar benefits to AQ prediction, because there are similar model complexities and constraints. Different AQ models can be better for different air-pollution episodes, in ways that cannot always be anticipated. Similar to NWP ensembles, AQ ensemble members can be created with different meteorological and/or emission inputs, parameterizations within a single model, numerics within a single model, and multiple models. Moreover, NWP ensembles have been very useful by providing information about the likelihood of possible future evolutions of the atmosphere. Similarly, AQ ensembles may be able to provide reliable probabilistic information about possible AQ scenarios. Given the nonlinear nature of photochemical reactions, an Ozone Ensemble Forecast System (OEFS), and the differences among the ensemble members, may rapidly account for the uncertainties associated with each component of the modeling process.

The first chapter introduces chaos theory and reviews the state of research relevant to this dissertation. The remaining chapters except the last chapter consist of journal papers resulting from this dissertation research. Thus, these chapters have their own introduction, conclusions and references. These journal papers are cited on the first page of each chapter. Chapter 2 investigates a multi-model approach to realize an OEFS. Chapter 3 introduces a new AQ ensemble design, combining meteorology and emission (NO_x) perturbations. These successful experiments prompted the work described in Chapter 4, where also a VOC perturbation is tested. Also the effects of different horizontal spatial resolutions, emission perturbations, and driving NWP models on the ensemble performance are investigated. Chapter 5 explores the application of Kalman-filter postprocessing to AQ forecasts to remove their systematic ozone errors. Finally, conclusions and recommendations for future work are the subject of Chapter 6.

Acknowledgments

First, I would like to thank my supervisor, Dr. R. B. Stull, for his tireless support and guidance, that allowed my research to develop with creativity and at the same time not to diverge toward outer spaces. I also would like to thank my examining committee, Drs. I. G. McKendry and D. G. Steyn who were always available to share with me their knowledge, and to provide timely advice.

This dissertation could not be completed without the support of the Weather Research and Forecast Team: for this reason I am grateful to all the team members. In particular, Dr. Xingxiu Deng and Yongmei Zhou deserve a special thanks for providing the Numerical Weather Prediction simulations used in this research. Also, I greatly appreciate the support that Colin di Cenzo (Environment Canada) showed for my research since I started my Ph.D.

I am grateful to Heinz Hass and his colleagues for providing the data on which the analysis in Chapter 2 is based. I am also grateful to RWDI for providing the emission inventory and the scripts to run SMOKE. Ken Stubbs and John Swalby (Greater Vancouver Regional District) graciously provided the ozone observation data used in Chapters 3, 4 and 5. Furthermore, I would like to thank Todd Plessel (EPA) for providing very useful tools to handle Models-3 formatted data.

I deeply appreciate all the time Bruce Thomson, Dr. Bruce Ainslie and Dr. Alberto Martilli (CIEMAT) spent with me discussing a variety of topics related to air quality (AQ) in the Lower Fraser Valley and to weather and AQ modeling. Also, I gratefully acknowledge several discussions and emails I exchanged with Drs. Stefano Galmarini (JRC) and Josh Hacker (NCAR), that helped me to gain insight into the ensemble-forecasting topic. Drs. Steven Hanna (Hanna Consultants), S. T. Rao (EPA) and John Irwin (EPA) dedicated part of their precious time to discuss with me different aspects of AQ modeling uncertainty studies, and I would like to express all my gratitude to them for doing this.

Finally, I want to thank my wife Serena and daughter Elisa for inspiring every bit of my research.

Support for this research was provided by the Natural Science and Engineering Research Council of Canada, the BC Forest Investment Account, the British Columbia Ministry of Water Land and Air Protection (formerly the Ministry of Environment), Environment Canada, and the Canadian Foundation for Climate and Atmospheric Science. Geophysical Disaster Computational Fluid Dynamics Center computers were used, funded by the Canadian Foundation for Innovation, the BC Knowledge Development Fund, and the University of British Columbia.

A Papá (To My Father)

Co-authorship Statement

Chapter (2):

Delle Monache, L., and R. B. Stull, 2003: An ensemble air-quality forecast over western Europe during an ozone episode, *Atmospheric Environment*, **37**, 3469-3474.
Published as "Fast Track", i.e., "...for papers that contain important and topical results whose significance merits fast publication.".

Co-author contributions:

R. B. Stull - suggested the application of the ensemble approach to AQ forecasts; revised the original manuscript; was lead researcher for the project, and provided funding for L. Delle Monache.

For this manuscript, I first searched the literature for previous work involving Chemistry Transport Models comparison studies. I then contacted H. Hass, who kindly provided me with the model forecasts and the observations used here. I designed and performed the analysis, and I prepared the manuscript.

Chapter (3):

Delle Monache, L., X. Deng, Y. Zhou, and R. B. Stull, 2005: Ozone ensemble forecasts. Part I: a new ensemble design, *accepted in November 2005 to be published in the Journal of Geophysical Research.*

Co-author contributions:

X. Deng - ran the MC2 model.

Y. Zhou - ran the MM5 model.

R. B. Stull - revised the original manuscript; was lead researcher for the project, and provided funding and computer-cluster access for L. Delle Monache.

For this manuscript I designed the experiment and ran the CMAQ model. I performed the analysis presented and I prepared the manuscript.

Chapter (4):

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2005: Probabilistic and ensemble-averaged ozone forecasts, *manuscript submitted in November 2005 to the Journal of Geophysical Research.*

Co-author contributions:

J. P. Hacker - revised the original manuscript; suggested the RI normalization; contributed to the design of the analysis and presentation.

X. Deng - ran the MC2 model.

Y. Zhou - ran the MM5 model.

R. B. Stull - revised the original manuscript; was lead researcher for the project, and provided funding and computer-cluster access for L. Delle Monache.

For this manuscript I designed the experiment and ran the CMAQ model. I performed the analysis presented and I prepared the manuscript.

Chapter (5):

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. B. Stull, 2005: Ozone ensemble forecasts. Part II: a Kalman-filter predictor bias correction, *accepted in August 2005 to be published in the Journal of Geophysical Research.*

Co-author contributions:

T. Nipen - wrote the Matlab code to compute Kalman-filtered forecasts based on the earlier work by Miranda Holmes. Wrote the section describing the Kalman-filter algorithm.

X. Deng - ran the MC2 model.

Y. Zhou - ran the MM5 model.

R. B. Stull - suggested the application of the Kalman-filter bias correction to AQ forecasts; revised the original manuscript; was lead researcher for the project, and provided funding and computer-cluster access for L. Delle Monache.

For this manuscript I designed the experiment and ran the CMAQ model. I performed the analysis presented and I prepared the manuscript.

Chapter 1

Background

To understand the potential for ensemble air-quality (AQ) forecasts, one must first understand the factors that limit atmospheric predictability. Those factors are introduced and discussed in the next three sections, along with a description of the ensemble approach and its implementation in weather and AQ forecasts. The fourth section describes the Kalman filter (KF) algorithm and discusses how it can be used to remove systematic errors in AQ forecasts.

1.1 Chaos; Atmospheric Predictability

1.1.1 Dynamical Systems

A *dynamical system* is a system evolving from an initial to a future state following physical laws that can be expressed with mathematical equations. To predict the evolution of such system, one can integrate the equations, starting from an observed initial state. Unfortunately, often this initial state cannot be determined precisely, as for the atmosphere.

In some cases, the evolution of a system can be anticipated without describing its initial state. This is the case of a dynamical system whose evolution is known to have a periodic behavior. For example ocean tides can accurately be predicted using the motions of sun, earth, and moon, and without knowing the initial spatial distributions of tide height. In other cases the evolution of a dynamical system cannot be predicted because there is an inexact or incomplete knowledge of its present state.

Some systems show *divergent* behavior, meaning that two different solutions starting from close (but different) initial states would eventually diverge in the solution space. In those cases we don't know a priori which of the two solutions is closest to the true evolution of the system.

Poincaré (1897, as described in (Alligood et al., 1997)) first discovered that the motion of a three-body system is "sensitively dependent on the initial conditions", introducing for the first time a so-called *chaotic* systems. Lorenz (1963) studied the behavior of such systems. The numerical solution of a system of equations coupled with each other, and/or involving non-linear terms, can be very sensitive to the initial values of the independent variables describing the present state of the system.

Regardless of how small is the distance in solution space between the initial states of two solutions of the same system, these solutions would eventually differ from each other as if they were randomly chosen. This implies that, for such a system of equations, there is an upper limit to their predictability; i.e., a limit in time after which one solution (the forecast) does not possess any useful information about the evolution of the other solution (the real weather).

The *phase-space* of a dynamical system is the multi-dimensional space of independent variables. The number of those variables is the *dimension* of the phase-space. A point in the phase space is called *state.* The evolution of the system in time is a set of states that is called the *trajectory*, or *orbit.* A dynamical system shows a chaotic behavior if most trajectories in the phase-space exhibit *sensitive dependence* to initial conditions (Lorenz, 1993). A trajectory is characterized by sensitive dependence if most other trajectories that pass close to it at some point do not remain close to it later.

The atmosphere shows this behavior, and is thus a chaotic system. We are not able to accurately measure the initial state of the atmosphere, due to instrumental errors and large gaps between observations sites. Moreover, we are able to solve only a simplified version of the equations describing the atmosphere, and those solutions are usually numerical approximations; i.e., they are sources of error as well. As a consequence there is an upper limit in time to the predictability of the weather. As Lorenz (1963) concluded in his milestone paper, "When our results concerning the instability of non-periodic flow are applied to the atmosphere, which is ostensibly non-periodic, they indicate that prediction of the sufficiently distant future is impossible by any method, unless the present conditions are known exactly. In view of the inevitable inaccuracy and incompleteness of weather observations, precise very-long-range forecasting would seem to be non-existent".

1.1.2 The Lorenz Model

Lorenz (1963) introduced a three-variable simple model representing finite amplitude convection, analogous to a tank of water with a heated bottom. Here a brief description of the model is given, and more details can be found in Lorenz (1993). The following equations represent this chaotic system:

$$
\begin{aligned}
\frac{dx}{dt} &= \sigma(y-x) \\
\frac{dy}{dt} &= rx - y - xz \\
\frac{dz}{dt} &= xy - bz
\end{aligned}
\tag{1.1}
$$

Where σ, r and b are parameters, and where x is proportional to the intensity of the convection motion, y is proportional to the temperature difference between the updraft and downdraft, and z is proportional to the distortion of the vertical temperature profile from linearity (positive if the strongest gradient occurs nearby the boundaries). The parameter values (chosen from Lorenz, 1963) were $\sigma = 10$, $r = 28$, and $b = 8/3$.

A *family of solutions* can be defined by varying those parameter values. Here only one set of values are considered to give a classical example of a chaotic system.

The system (1.1) is *dissipative*; i.e., the phase space volume contracts along a trajectory. This can be seen from the divergence of the flow:

$$
\frac{\partial \dot{x}}{\partial x} + \frac{\partial \dot{y}}{\partial y} + \frac{\partial \dot{z}}{\partial z} = -(\sigma + b + 1) \tag{1.2}
$$

An original volume V contracts with time to the value $Ve^{-(\sigma+b+1)t}$. For example the atmosphere (with friction) is a dissipative system. The fact that the system is dissipative as shown by (1.2) implies the existence of a bounded globally attracting set of zero volume, or more generally, an attractor of dimension smaller than the dimension of the phase-space.

An *attractor* is a set of states (points in the phase-space), invariant under the dynamics.

A *basin of attraction* is a set of points in the phase-space such that initial conditions chosen in this set dynamically evolve to the attractor. Neighboring states in a given basin of attraction asymptotically approach the attractor in the course of the system evolution. The first portion of the trajectory (away from the attractor set) is called the *transient.*

Lorenz (1965) extended his work by considering a 28 variable model. For the following discussion, Szunyogh et al. (1997) and Pu et al. (1997) are also valuable references. Lorenz noticed that the errors tend to grow along selected directions in the phase-space. He observed how a hyper-sphere around a state, representing a small perturbation of the state in the phase-space, evolves in time following the system evolution. If the system is chaotic (two close trajectories will eventually diverge completely) the hyper-sphere will initially evolve in a hyper-ellipsoid, because at the beginning linear effects dominate the system evolution. Once non-linear effects start to become important, the hyper-ellipsoid becomes a "banana" shaped surface, and a few axes of the hyper-ellipsoid will start to grow more rapidly than others. While those axes will keep growing, the banana shaped surface will keep elongating and stretching along those axes directions. If the system is bounded, this surface will fold on itself several times, and eventually will converge into the zero volume attractor with an infinitely foliated structure (Kalnay, 2003).

More precisely, along each axis the long-term average of stretching or contraction of the solution space is given by $e^{\lambda_i t}$ where λ_i are called the Lyapunov exponents of the i axis. The volume of the hyper-ellipsoid is proportional to $V_0 e^{-(\lambda_1+\lambda_2+\cdots+\lambda_n)t}$, where V_0 is the volume of the initial hyper-sphere and n is the dimension of the phase-space. If the sum of the Lyapunov exponents is zero the system is *Hamiltonian* and it conserves its volume. For a dissipative system, the sum is positive.

In a chaotic system, one or more exponents are positive. The axis associated with them will grow indefinitely, and will allow the separation of two trajectories initially close to each other. Moreover, if the system is bounded at least one of the Lyapunov exponents must be equal to zero.

Lyapunov exponents give us information about global properties of flow evolution, and for the atmosphere its attractor is climatology. However for weather predictability, the main focus is on local space and short time stability properties of the flow. Thereby, *local leading Lyapunov vectors* (LLLV) or *finite Lyapunov vectors* can be defined (Trevisan and Legnani, 1995), to indicate the direction in which the maximum error growth occurs, locally. LLLV can be viewed as the results of defining the Lyapunov exponents for a particular limited space region, and for a finite period of time.

The concepts introduced by Lorenz (1965) lead to the development of the *singular-vectors* (SV) approach (Molteni and Palmer, 1993). An error term is added to the linearized version of the system equations, and the solution of the resulting system can be seen as an *error-propagator matrix.* The eigenvectors of this matrix multiplied by its transpose are called the singular-vectors, and their eigenvalues equal the square of the singular values of the matrix. Those vectors point to the directions of greater error growth, and the adjoint of the linear model will project the errors back onto the initial state. The singular vectors are extremely sensitive to the choice of norm and the time period over which they are applied (Errico and Vukicevic, 1992).

An advantage of LLLV over SV is the fact that they do not depend on the norm used to define them. Moreover SVs initially do not point to the attractor, but point to subspaces of the phase-space where solutions do not usually occur. A disadvantage is that LLLVs grow much slower than the SV, and initially they do not closely resemble the true error growth.

The Lorenz's work had important implications for predictability of chaotic systems and ensemble forecasts as will be illustrated in the next section. The error growth of weather forecasts are highly dependent on the flow of the day, and is bigger along a few directions in phase space, while along others errors diminish. After a short time, there are only few dominant directions important to describe the error growth of a dynamical system. Since the error growth can be characterized with few Lyapunov exponents, only a few ensemble members with appropriate perturbations are needed to potentially estimate the dynamical system evolution better than a single deterministic prediction.

1.2 Numerical Weather Prediction (NWP) Ensembles

The first part of this section is a brief history of the evolution of NWP ensembles; namely, the use of multiple NWP forecasts to better estimate the future weather and the confidence of the prediction. This is followed by a description of the ensemble approaches used in the main operational forecast centers around the world. Last, a closer look at the more recent ensemble research efforts is given.

1.2.1 NWP Ensemble History

After Lorenz (1963, 1965), the scientific community started to consider the issue of limited predictability of any non-linear dynamical system with instabilities, such as the atmosphere. The smallest approximation in the forecast model or the tiniest error in initial conditions will lead to a total loss of skill in the weather forecast after a finite time. He estimated the weather predictability limit is two weeks, on average, and that the limit for any given day strongly depends on the instabilities associated with the flow of the day.

The Lorenz work inevitably lead researchers to consider the stochastic nature of the atmosphere. Namely, one could follow the evolution in phase-space of a probability density function (PDF) describing the atmosphere's initial state and its uncertainty. Although a mathematical formulation of such system of equations can be formulated through the continuity equation for probability (Liouville equation, Ehrendorfer (1994)), a solution of it would imply an over simplification of the equations themselves, or would require an impossible computational effort.

The *ensemble* approach comes from the necessity of representing the time evolution of the PDF describing atmospheric state. The PDF can reasonably be represented by a limited set of points. The evolution of each of those points would be a member of the ensemble. Each of those members should ideally represent an equally likely evolution of the dynamical system.

Gleeson (1966, 1967) and Epstein (1969a) first clearly stated the necessity of a probabilistic prediction, as oppose to a deterministic one, in simulating atmospheric evolution. Their statement is based on the unquestionable fact that we can estimate the true value of the atmosphere in only a probabilistic fashion.

In *stochastic-dynamic forecasting*, Epstein (1969b) derived a continuity equation for a PDF representing the model solution. He explicitly forecasted the first and second moments of a PDF related to a simplified version of the Navier-Stokes equations. He concluded that a stochastic prediction provides better forecasts than a deterministic one, and also gives useful information about the uncertainty. Unfortunately his approach is completely unfeasible for a model with millions of degrees of freedom.

Leith (1974) proposed *Monte Carlo forecasting*, where a limited number of ensemble members was required to create the ensemble. He randomly created the perturbation of the starting analysis from which each ensemble member was initialized. He assumed a perfect model; i.e., assumed that the model can closely follow the evolution of the atmosphere if initialized correctly. This is a deficiency of any ensemble that takes into account only the errors associated with the initial conditions. Nevertheless, he suggested that a Monte Carlo ensemble behaves similarly to a stochastic ensemble, but is computationally much cheaper. He also noticed that the ensemble average filters out the unpredictable small scales, and improves the forecast skill by leaving the bigger scales virtually unaltered. The loss of small-scale features related to averaging is still an open issue in the weather forecasting community (Kalnay, 2003). He finally suggested that a set of ensemble members as small as eight would lead to adequate accuracy in the forecast.

After Leith (1974), and after the error estimates in Daley and Mayer (1986), many applications of the Monte Carlo ensemble forecasts can be found in the literature (Errico and Baumhefner, 1987; Tribbia and Baumhefner, 1988; Mullen and Baumhefner, 1989, 1994). Those works show that the assumption of initial-state errors are limitations of the Monte Carlo approach, and that the magnitude of ensemble spread is not representative of the error growth, especially for short-range forecasts.

Hoffman and Kalnay (1983) introduced the *lagged-average forecast.* The forecast initialized at the current initial time, $t = 0$, as well as forecasts from the previous times, $t = -\tau, -2\tau, \ldots, (N-1)\tau$ are combined at a common valid time to form an ensemble. They weighted each member with its expected error, based on its "age". They estimated this error by parameterizing the observed error covariance growth. They found the lagged-average forecast to be slightly better than the Monte Carlo forecast, and they found higher correlation between error growth and ensemble spread (i.e., differences between the ensemble members) in their approach. These improvements were achieved because the lagged-average forecast perturbations are not randomly chosen, but better capture the error of the day. In the literature a few other applications of this ensemble approach can be found, as for example in Dalcher et al. (1988).

1.2.2 Operational Ensemble Forecasting and Recent Advances

In the early 90's the US National Centers for Environmental Prediction (NCEP) and the European Center for Medium-Range Weather Forecasts (ECMWF), implemented two new approaches. The common idea is that the perturbations to create the ensemble members should be focused mainly on the fastest-growing modes of the atmosphere. NCEP and ECMWF obtained similar perturbations, even though there are important differences. Similarities and differences of these two approaches are illustrated next.

Toth and Kalnay (1993, 1997) introduced the *breeding method.* The idea is to build a cycle to "breed" the fast growing "errors of the day". A breeding cycle is introduced by random perturbations with a given size, measured with any norm. This random approach must be followed only in the first cycle. The model then is run from the original analysis (*control run*) and from the perturbed analysis for a fixed cycle length. At the end of each cycle, the control run is subtracted from the perturbed forecasts, and the resulting differences are scaled down to the same amplitude as the initial perturbation. Then they are added to the new analysis, and another forecast cycle is performed. The authors argued that those differences resemble the fastest growing errors, a desirable feature. They also suggest that this method creates perturbations only along those modes that dominate the forecast errors, as opposed to a random perturbation that will span mainly onto non-growing modes. Toth and Kalnay (1993) and Kalnay (2003) moreover argue that the breeding method is a nonlinear generalization of the process used to obtain the Lyapunov vectors. The nonlinear aspect of the breeding method filters out the Lyapunov vectors associated with energetically negligible and fast-growing instabilities as for example convection. A drawback is that the bred vectors depend on the initial random seed.

Molteni and Palmer (1993) and Buizza (1997) describe a different approach used at ECMWF, where the perturbations rely on the SV properties. The SVs depend on the particular norm that is utilized, and also on the time over which the operator is applied. The errors can be projected back onto the initial state by applying the adjoint of the linear model. Since the SVs represent the axes of the ellipsoid picturing the initial error evolution in the phase-space, to create the perturbations their values are added and subtracted to the initial conditions. Those perturbations are finally scaled down to the magnitude of the analysis error estimate.

Rabier et al. (1996) showed that the day-2 forecast error growth projects well into the space of dominant SVs. Using the information given by the SVs on the directions of the most rapid error growth in NWP models, localized SVs have been used to construct initial perturbations for the ensemble prediction system of ECMWF (Palmer, 1993; Molteni and Palmer, 1993; Ehrendorfer and Errico, 1995).

Ensemble forecasts also provide information on the reliability of the forecast: if the ensemble members have large spread, this indicates that at least some of them do not represent the true evolution. The standard deviation of the ensemble members about the ensemble mean is called *ensemble spread.* The relationship between the ensemble spread and the forecast error is not yet well defined (Kalnay, 2003). Nevertheless, it often provides very useful information about ensemble skills. Greater spread suggests less confidence in the ensemble mean forecast.

The most promising approach for limited-area (mesoscale) short-range forecasts is the *multi-model ensemble* approach (Krishnamurti et al., 1999; Hou et al., 2001; Toth, 2001; Wandishin et al., 2001), where forecasts from different models form the ensemble members. The idea is not only to capture the uncertainties in the initial and boundary conditions, but to also acknowledge that the models contain many uncertainties in their formulation, numerics, parameterizations, and time and space discretization.

Currently the Canadian ensemble system at the Recherche en Prévision Numérique (RPN) center is based on a *system simulation experiment* (SSE) (Houtekamer et al., 1996). In an SSE, it is considered that all elements of the forecast system, observations, analysis, and model are subject to uncertainty. The elements of the system are perturbed in different ways for different members of the ensemble. By considering uncertainty in both the analysis and the model, the RPN approach, in its current 16 member configuration, is a true multi-model ensemble: two completely different models, the older global spectral model (SEF) and the newer global version of the generalized environmental multiscale (GEM) model are used.

1.3 Ozone

1.3.1 Introduction

Ozone (O_3) is a reactive oxidant gas naturally produced in the atmosphere. Figure 1.1 shows a typical vertical O_3 profile. Stratospheric levels can reach $10,000$ parts per billion (ppb), whereas background levels near the surface are only few tens of ppb.

Stratospheric O_3 absorbs ultraviolet radiation emitted by the sun. In the last 30 years this

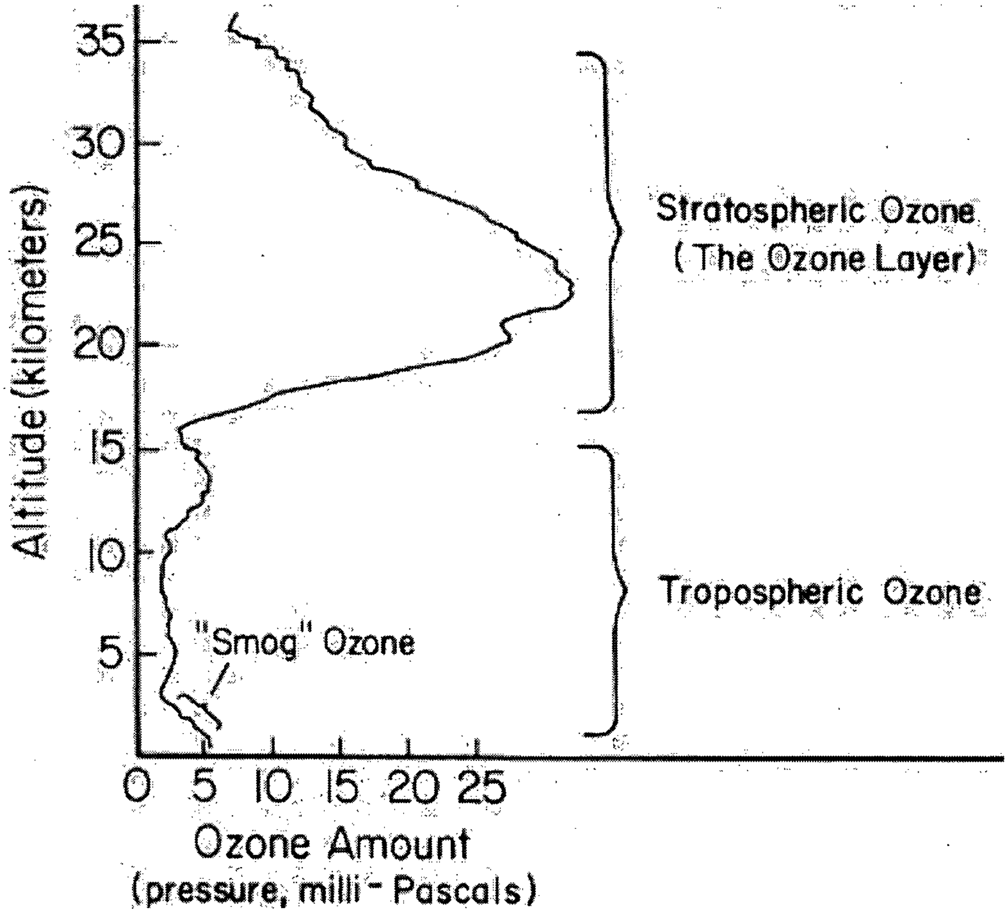

Figure 1.1: Ozone typical vertical profile (source: http://www.al.noaa.gov/WWWHD/pubdocs/Assessment98).

layer has partially depleted, partly caused by anthropogenically produced chlorine compounds (Molina and Roland, 1974; WMO, 1998).

However, this thesis focuses on tropospheric O_3, which is increasing primarily because of increased fossil-fuel combustion by people (WMO, 1986, 1990). Ozone-rich *photochemical smog* is the result of chemical interactions of nitric oxide (NO), nitrogen dioxide (NO_2), and reactive organic gases (ROGs) - also called volatile organic compounds (VOCs), and sunlight. Often NO and NO_2 are classified together as NO_x.

Typically, NO_x and ROGs are emitted from vehicular and stationary combustion sources. ROGs become free radicals via chemical reactions. The radicals or O_3 (via NO_x *titration*) can transform NO into NO_2. Finally molecular oxygen (O_2) reacts with atomic oxygen (O) to form ozone. Pollutants can be divided into *primary* and *secondary*: the former are gases and particles that are directly emitted into the atmosphere from surface or elevated sources (e.g., NO), and the latter are created chemically (e.g., O_3) or physically within the atmosphere. A detailed description of the chemical transformations involved in tropospheric O_3 formation can be found in Jacobson (1999) and Seinfeld and Pandis (1998).

The chemical pathway summarized above can be described with a set of nonlinear equations representing the chemical reactions. The O_3 production depends on the concentration of primary pollutants that lead to its formation. Those pollutants have lifetimes that may differ significantly from one another. Ozone, once formed, can reside in the atmosphere a month or longer, but is often titrated by contact with the earth's surface. This leads to seasonal, synoptic, diurnal, and subdiurnal variations of ozone concentration at the surface and aloft (Hogrefe et al., 2001).

Meteorology is an important factor affecting photochemical pollution creation, transport, and deposition. Sunlight allows photochemical reactions, and its intensity directly governs photolysis rates. Local and mesoscale flows mainly determine the distribution of pollutants: sea and land breezes, katabatic and anabatic winds, and valley flows. They all can play an important role in the development of air quality. Moreover specific synoptic conditions usually are necessary for photochemical pollution episodes to happen in different locations. High-pressure systems at the surface and aloft, surface thermal lows, subsidence and entrainment in the Atmospheric Boundary Layer (ABL), and stagnation conditions all may affect the composition of the air we breathe.

Tropospheric O_3 has been recognized as an harmful gaseous pollutant for many years. Oxidant pollutants can affect negatively the human respiratory system (for example, Horvath and McKee, 1994; Brauer and Brook, 1995). O_3 exposure reduces lung function, and aggravates existing respiratory diseases, such as asthma. The degree of adverse respiratory effects produced by O_3 depends on several factors, including concentration and duration of exposure, climate characteristics, individual sensitivity, and preexisting respiratory diseases.

O_3 is one of the most damaging air pollutants to plants. O_3 can be advected by the wind across great distances to cause damage to plants far from its origin. The extent of plant damage depends on the concentration of O_3, the duration of exposure, and plant sensitivity. Acute damage has been observed to both deciduous trees and conifers (Runeckles, 2002). Finally, ozone can also damage materials. For example rubber and plastic products deteriorate quicker if exposed to high ozone concentrations (Brown et al., 2001).

1.3.2 Ozone Forecasts

Tropospheric O_3 has been designated a "criteria pollutant" since 1970, and health standards have been set since then in many countries. Those standards try to account for the natural variability of O_3, and for rare events. For example in Canada, the National Ambient Air Quality Objectives (NAAQOs) set the maximum 1-hour average concentrations to 82 ppb. From 2010, a new Canada Wide Standard (CWS) will be set to 65 ppb for the 4^{th} highest 8-hour averaged concentration during a span of three consecutive years (CCME, 2000). In the US, the 1997 Clear Air Act Revision (EPA, 1997) set the O_3 8-hour averaged standards to 85 ppb for the 3^{rd} highest reading over four years.

The discovery of ozone's harmful effects on humans and vegetation led to two outcomes: the necessity of issuing AQ forecasts; and the need to limit and control adverse anthropogenic emissions. Although O_3 formation is extremely complex, its maximum concentration is well correlated to weather parameters, and its variations can be described with fewer meteorological predictors. For these reasons, different attempts have been made to design simple ways to predict O_3 maxima at a specific location or over a prescribed spatial domain. Statistical approaches include multiple-regression analysis (Ryan, 1994), nonlinear regression (Hubbard and Cobourn, 1997), neural networks (Ruiz-Suarez and Mayora-Ibarra, 1995), classification and regression tree schemes (CART) (Burrows et al., 1995), and hybrid approaches (Liu and Johnson, 2002). A comprehensive discussion of these techniques and their forecast skills can be found in EPA (2003). The statistical approaches have limited, if any, description of physical and chemical processes; they usually predict only the maximum concentration, and they have difficulties in anticipating rare events. Moreover they can be applied only over areas with large data availability, and this limits their applicability mainly over metropolitan areas (EPA, 2003).

Ainslie (2004) proposed a scaling-level model for ozone photochemistry, where a dimensional analysis was used to categorize the relevant variables in different dimensionless groups. The relationship between the groups can be parameterized with a simple expression. The model appeared to capture the ozone dependency on meteorological conditions and precursor concentrations, resulting in a useful screening tool.

To better account for all the processes and variables involved in O_3 formation, a complex 3-D modeling system is needed. For regional AQ forecasts, such a system should include a mesoscale model to produce the meteorological fields, an emission inventory processor, and a chemistry and transport model. With such a system the population can be alerted about impending air-quality degradation in urban, rural, and remote areas. Such forecasts provide much more detailed spatial and temporal information, allowing better decisions regarding daily activities. Daily AQ forecasts can give insights into peculiarities of pollutant behavior in specific regions, such as winter valley particulate matter down-transport, tropopause folding, and gravity-wave breaking over the Fraser Valley and South West British Columbia (Hacker et al., 2001). AQ forecasts can be useful for prescribed forest fires and agricultural field burning to minimize smoke impact on the local population and on regional haze (e.g., Achtemeier et al., 2005). Moreover, 3-D AQ models can be used to plan long-term emission controls to reduce the impact of pollution on population (e.g., Jonson et al., 2001).

Dabberdt and Miller (2000) confirm the need for an operational AQ forecast system, and recommend the use of probabilistic approaches as is already used in weather forecasts. The first experiences in this direction are described in Delle Monache et al. (2004), McHenry et al. (2004) and Vaughan et al. (2004). Finally the need for AQ probabilistic forecasts, which are the subject of this thesis, have been promoted also by the U.S. Weather and Research Program and its Prospectus Development Team on Air Quality Forecasting (Dabberdt et al., 2003).

1.3.3 Ensemble Trajectory Modeling

This section reviews ensemble dispersion studies, where air parcel paths, also called *trajectory* (not to be confused with the definition given in Section 1.1), are modeled. Most studies estimate trajectory errors by simulations that verify for the same period. They share the basic premise that deterministic model prediction cannot reliably represent pollutant trajectories.

In early work, Merrill et al. (1985) computed isentropic trajectories using data from the 1979 Pacific Sea-Air Exchange (SEAREX) experiments. They computed trajectories kinematically using grid-point values of geopotential height and wind provided by ECMWF. The authors assert that single trajectories are of limited usefulness, because of the uncertainties in their calculations and in the data. Therefore, they computed an ensemble of trajectories with nine to 19 ensemble members by perturbing the initial conditions. They recognized that trajectory-calculation precision was affected by: the assumption of adiabatic flow, the exclusion of precipitation and particle gravitational setting, and the data void in remote areas of the Pacific Ocean. Those factors forced them to consider a probabilistic approach.

They also tested the trajectory sensitivity to the meteorological analysis. For a few cases, they used both ECMWF and National Meteorological Center (NMC) global analyses, where the latter had coarser resolution. They realized that trajectories can be sensitive to differences in meteorological inputs. Moreover, for different trajectories that verify for the same period and domain, they associated the ensemble spread to information on the intrinsic predictability of the flow.

Stohl et al. (1995) identified interpolation errors as a major problem in trajectory computation. To estimate those errors, they suggested creating an ensemble of trajectories by adding random errors at each time step to a reference trajectory.

Similarly Kahl (1996) studied the relationship between errors in predicted trajectories and the instability associated with the meteorology. He defined a *meteorological complexity factor* (MCF) to forecast model-trajectory errors. MCF is the average distance between the trajectories and a reference trajectory. He assumed that trajectory uncertainty could be predicted as a function of the MCF. The weakness of this approach is that the magnitude of the MCF depends critically on the integration time step.

He computed MCF using results from 22 published studies. He also used a Monte Carlo simulation to compute 144,000 different trajectories by superimposing random perturbations upon the wind field used to compute the reference trajectory. The author found that the error growth may be unstable with respect to small perturbations in the wind field. This behavior closely resembles the description of a chaotic system. The author concluded by encouraging as future development "...a methodology for predicting the confidence which one may place in individual trajectory calculations...". One methodology could certainly be the ensemble approach.

Baumann and Stohl (1997) analyzed a 4-day record of gas balloon tracks during an international long-distance ballooning competition. They compared the balloon trajectories using ECMWF meteorological analyses, and they ran a modified version of the model FLEXTRA (Stohl et al., 1995), taking into account balloon ascent and descent. In addition to a reference trajectory they calculated 100 ensemble trajectories. They started the ensemble trajectories from a 100 km radius circle around the reference starting position, which was the grid resolution of their meteorological data. To consider interpolation errors they also perturbed the horizontal wind field, by adding normally distributed random errors to the reference field. The authors recognized that those ensembles did not account for the errors embedded in the wind analysis. Moreover, they did not account for uncertainties in the vertical wind. Nevertheless, they concluded that their ensemble usually enveloped the balloon tracks, indicating that the errors neglected from their ensemble approach were small. The ensemble provided useful information about the computed trajectory uncertainties. The authors noticed a good qualitative correlation between the uncertainties and the ensemble-member spread.

Stull et al. (1997) considered the potential benefit of ensemble AQ dispersion modeling, analogous to the benefit for weather ensembles. They perturbated the weather analysis for the Global Spectral Model of NCEP to generate a set of equally-likely initial conditions to initialize two weather mesoscale models, the Canadian Mesoscale Compressible Model (MC2) and University of Wisconsin Nonhydrostatic Modeling system (UW-NMS). The authors speculated about trajectory behavior, forecast confidence and predictability.

Straume et al. (1998) used a Lagrangian dispersion model, the Severe Nuclear Accident Program (SNAP). Ensemble meteorological forecasts produced by ECMWF (where the ensemble perturbations are calculated using singular vectors) were used as input to study starting-analysis error growth associated with atmospheric instabilities (Section (1.2.1)).

The 32 ensemble members plus the control forecast were processed with the High Resolution Limited Area Model (HIRLAM). The 33 ECMWF forecasts output data every 12 hours for five vertical layers. HIRLAM transformed this data into a data set with values every six hours onto 32 vertical layers.

Straume et al. (1998) compared their simulation results with the European Tracer Experiment (ETEX), which includes measurements of two tracers released in France during southwesterly flow in October and November 1994 (Nodop et al., 1998). The tracers were measured over a period of 72 hours for both releases.

The ensemble members and the control forecast were used as inputs to the SNAP model, to realize 33 dispersion simulations for the same domain and time period. To estimate the weather predictability, Straume et al. (1998) computed the *root-mean-square deviation* (rmsd) of the computed concentrations from the control forecast. Because of the strong dependence of this value upon the geographical area (due to some grid points containing zero tracer concentrations) the authors argue that those deviations are qualitative measurements of uncertainties of the meteorological input. The rmsd grows from 0 to 4 % or less for the first three days, and reaches an average value of 7 % after 72 hours from the release. The authors also computed the centroid position for both the modeled and measured dispersion, and found an uncertainty of the model between 10 and 20 %, with a distance between the two centroids between 20 and 90 km for the first 21 hours, and a maximum of 300 km after 48 hours. Even though the modeled and measured puff arrivals were significantly correlated, the authors found an average of 6 hours delay of the model puff arrival at the stations compared to what was measured. The puff durations were not correlated.

Dabberdt and Miller (2000) simulated an actual three hour accidental release of oleum in the city of Richmond, in the San Francisco Bay Area. They ran a non-steady-state puff-type dispersion model driven by a diagnostic mass-consistent wind field model. They argue about the utility of a probabilistic approach, particularly in cases of accidental releases, when there are scarce meteorological measurements, and scarce background concentration data.

The authors generated 162 ensemble members by perturbating the stability classes, the wind speed and direction, the source strength, and the plume rise. They clearly showed how the information that can be extracted from the ensemble could help the decision makers in taking the most appropriate and feasible actions.

Galmarini et al. (2001) developed a Real Time Model Evaluation (RTMOD) procedure, whose aim is to improve the ability to simulate long-range dispersion processes for nuclear emergency applications (Bellasio et al., 1999). Their ensemble is formed by more than 20 models run by different organizations around the world to predict the transport and deposition of radioactive releases in the atmosphere. They tested ensemble performance by comparing the model prediction against each other and against observations during the ETEX experiment (Nodop et al., 1998). The ensemble is created by perturbating the initial conditions, and by using multiple models, where the uncertainties of all the dispersion modeling process are somewhat taken into account. With the ensemble, the authors could estimate forecast uncertainty, and could indicate which parts of the domain are more likely to be exposed to the dispersed contaminant. Moreover, the ensemble gives clues on the reliability of this information. The authors argued that such a multi-model system could be useful for operational decision-makers, and for modelers to check systematic model errors and general tendencies in their prediction.

Straume (2001) extended the earlier work of Straume et al. (1998), by further evaluating the HIRLAM model. The author found that the ensemble members close to the control forecast, as measured with one statistical parameter, were not necessarily close if a different parameter was used. She computed the *bias*, the *Pearson correlation coefficient*, the *figure of merit in space*, the *absolute horizontal transport deviation*, and the *relative horizontal transport deviation*. She also compared the SNAP results with 34 dispersion models that participated in the ETEX experiment, arguing that the errors in the meteorological input fields and in the model formulation are important throughout the simulation period, whereas the analysis error starts to be important only after the first day of simulation. The ensemble mean was more reliable than the control forecast in predicting the arrival of the contaminant at a given location, but was less reliable in predicting non-arrival events. Moreover, the ensemble mean predicted the puff trajectory better than the control forecast. Finally, the author noted that the selection of ensembles that are based upon singular vectors, which show the greatest growth at longer times, might not be the most appropriate for shorter-range-dispersion forecasts.

Scheele and Siegmund (2001) used the ECMWF wind data for the period 4 to 28 April 1998 to estimate the uncertainty in the trajectory of a transported air parcel, using the ensemble approach. They investigated how the accuracy of the forecasted trajectory is related to the ensemble spread and to other ensemble properties. They defined the *middle member of the ensemble trajectories* (MMT), the *operational forecast data* (FCT), the *control forecast member* (CRT), and the *bias of FCT* (BIA), as the root-mean-square distance of the FCT from the members of the ensemble.

Their results show a "modest but significantly positive" correlation between MMT and BIA, particularly at the beginning of the run. Also the difference between FCT and CRT is large, because of the different resolutions at which they are computed. The authors argue that for this reason the FCT uncertainty can be computed, but its actual position cannot. Nevertheless, the possible positions can be computed by adding the uncertainty, estimated with the ensemble, to the FCT, because a higher spatial and temporal resolution FCT is more accurate than MMT, especially when BIA is large. However, after two days of simulations they found that the contrary is true.

Warner et al. (2002) simulated an hypothetical dispersion of a toxic gas near Al Muthanna, Iraq, during the 1991 Gulf War. They tested an ensemble created by coupling the Penn State-NCAR Mesoscale Model (MM5) with the Second Order Closure integrated Puff (SCIPUFF) Lagrangian dispersion model. The authors created 12 ensemble members by running MM5 with different boundary-layer parameterizations, different surface physics, and different large-scale analyses used as a first guess and for the lateral boundary conditions. They found that the uncertainties in the dynamic meteorological model can be quantified, using the ensemble dosage probabilities, in a much more efficient way than with a single deterministic forecast. Moreover Warner et al. (2002) used the ensemble fields to generate the wind-field variances, which were then used directly in the dispersion model to compute the air concentration probability function.

Draxler (2003) used the same approach as Baumann and Stohl (1997), to study the sensitivity of dispersion to trajectory errors. The dispersion model was a modified version of the Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT). The meteorological input field was provided by the NCEP National Center for Atmospheric Research reanalysis project. The ensemble results were compared to measurements done over a period of three months, during the Across North America Tracer Experiment (ANATEX; Draxler et al. (1991)). In building their ensemble system the author assumed that the errors in the plume position are mainly dependent on the error in the particle trajectories. The 27 ensemble members were calculated by offsetting the meteorological fields by ±1 grid point in the horizontal and ±250 m in the vertical direction. The rationale behind this approach is that the meteorological field depends strongly on the limited spatial and temporal resolution used in the analysis, and that only processes larger than the actual grid size can be described and resolved.

Draxler found that the cumulative distribution of the ensemble probabilities was similar to one of the concentration measurements, but he also found that the distribution of those probabilities was not uniform. This is not a desirable feature of an ensemble, where, ideally, each member should be equally likely. The author argued that this could be attributed to the sensitivity of the release height. He also found that the ensemble accounts for approximately 41 % to 47 % of the variability of the measurement data, and this can be attributed to the fact that not all the errors embedded in this dispersion modeling processes and in the data are accounted for with this specific ensemble design.

Delle Monache and Stull (2003) analyzed for the first time the benefit of the ensemble approach in studies involving not only the pollutant transport, but also the associated photochemical reactions. Their ensemble was composed of four Chemistry Transport Models (CTM). Details on this study are given in Chapter 2.

In Galmarini et al. (2004a) the ensemble approach and its application to long-range transport and dispersion studies is rigorously presented. The authors introduce ad-hoc statistical treatments and parameters that nicely summarize the extensive information provided by an ensemble system. They also prove the superior forecast skills of the ensemble when compared to any single deterministic forecast representing an ensemble member. The parameters they introduce are *space overlap* (SO), *agreement in threshold level* (ATL) and *agreement in percentile level* (APL).

Following this study, Galmarini et al. (2004b) used the data collected during the ETEX experiment (Nodop et al., 1998) to quantitatively estimate the concepts and parameters introduced in Part I of their coupled papers. They tested a multi-model ensemble dispersion system, by considering several operational long-range transport and dispersion models (run in various European centers, in the US, and Canada) used to support decision making in case of accidental releases. The parameters they proposed were shown to be well suited for long-range transport and dispersion models. The median member of the forecast ensemble exhibited the best forecast skill. This differs from most ensemble weather forecasts, where the ensemble average is usually used. Finally the authors speculated that those parameters could also be applied to short-range dispersion and weather fields.

1.4 Systematic-Error Removal

Three-dimensional, coupled, NWP and AQ models do not usually make perfect forecasts in spite of their high level of physical detail and spatial resolution. For NWP models, statistical postprocessing known generally as model output statistics (MOS) had been used for many decades by the large government forecast centers to improve the raw NWP output. One such MOS method is called Kalman filtering (KF), which is a recursive algorithm to estimate a signal from noisy measurements (Homleid, 1995; Roeger et al., 2003).

Details of the KF method are given in Chapter 5. In summary, it uses a predictor-corrector approach to estimate future forecasts biases from past biases. When this future bias is combined with a NWP forecast of future weather, the result removes a large portion of the systematic error of the forecast, and can also remove a small portion of random error. In short, it yields a much more accurate forecast.

It will be shown in this dissertation that the KF is also very effective at improving the accuracy of AQ forecasts.

1.5 Research Goals and Activities

The ultimate goal of this research is to improve real-time short-term forecasts of tropospheric pollutants such as ozone measured at near-surface receptor sites.

This research is based on the hypothesis that the ensemble technique and Kalman-filter postprocessing can be transferred to AQ modeling, and can potentially yield similar benefits as for NWP. The method is 3-D mesoscale NWP modeling coupled with 3-D chemical numerical modeling. The procedure is to run these models using real emission inventories for real ozone episodes, and to calibrate and verify the results against actual near-surface ozone observations.

To accomplish these goals, the following research work is conducted:

- The realization and test of an AQ ensemble built on a previous photochemical model intercomparison study (see Chapter 2). This preliminary work demonstrated the value of ensemble AQ forecasts, and opened the door for the subsequent, more-detailed research that followed.

- The realization and test of a new AQ ensemble design, created by perturbating the input fields that most affect the uncertainty of the AQ photochemical models; i.e., the meteorological and the emissions fields (see Chapters 3 and 4).

- The realization and test of probabilistic forecasts resulting from ensemble methods (see Chapter 4).

- The realization and test of a new way to remove AQ forecasts systematic errors, based on the KF-predictor algorithm (see Chapter 5).

- Investigation of possible generalizations deduced from the results of the AQ-ensembles and KF corrections implemented and tested during this research (see Chapter 6).

1.6 References for Chapter 1

Achtemeier, G. L., S. L. Goodrick, and Y. Liu, 2005: A coupled modeling system for connecting prescribed fire activity data through CMAQ for simulating regional scale air quality. In *NOAA/EPA Golden Jubilee Symposium on Air Quality Modeling and Its Applications*. American Meteorological Society, Durham, North Carolina.

Ainslie, B., 2004: *A photochemical model based on a scaling analysis of ozone photochemistry*. Ph.D. thesis, 311 pp., University of British Columbia, Vancouver, Canada.

Alligood, K. T., T. D. Sauer, and J. A. Yorke, 1997: *CHAOS: an introduction to dynamical systems*. Springer-Verlag, New York.

Baumann, K. and A. Stohl, 1997: Validation of a long-range trajectory model using gas balloon tracks from the Gordon Bennett Cup 95. *Journal of Applied Meteorology*, **36**, 711–720.

Bellasio, R., R. Bianconi, G. Graziani, and S. Mosca, 1999: RTMOD: an internet based system to analyze the predictions of long-range atmospheric dispersion models. *Computers and Geosciences*, **25**, 819–833.

Brauer, M. and J. R. Brook, 1995: Personal and fixed-site ozone measurements with a passive sampler. *Journal of the Air and Waste Management Association*, **45**, 529–537.

Brown, R. P., T. Butler, and S. W. Hawley, 2001: *Ageing of Rubber - Accelerated Weathering and Ozone Test Results*. Rapra, Shawbury, United Kingdom.

Buizza, R., 1997: Potential forecast skill and ensemble prediction, and spread and skill distribution of the ECMWF ensemble prediction system. *Monthly Weather Review*, **125**, 99–119.

Burrows, W. R., M. Benjamin, S. Beauchamp, E. R. Lord, D. McCollor, and B. Thomson, 1995: Cart decision-tree statistical analysis and prediction of summer season maximum surface ozone for the vancouver, montreal, and atlantic regions of canada. *Journal of Applied Meteorology*, **34**, 1848–1862.

CCME, 2000: Canada-wide standards for particulate matter (PM) and Ozone. Technical report, Canadian Council of Ministers of the Environment.

Dabberdt, W. F., M. A. Carroll, D. Baumgardner, G. Carmichael, R. Cohen, T. Dye, J. Ellis, G. Grell, S. Grimmond, S. Hanna, J. Irwin, B. Lamb, S. Madronich, J. McQueen, J. Meagher, T. Odman, J. Pleim, H. P. Schmid, and D. Westphal, 2003: Meteorological research needs for improved air quality forecasting: report of the 11th Prospectus Development Team of the U.S. Weather Research Program. Technical report, National Center for Atmospheric Research.

Dabberdt, W. F. and E. Miller, 2000: Uncertainty, ensembles and air quality dispersion modeling: applications and challenges. *Atmospheric Environment*, **34**, 4667–4673.

Dalcher, A., E. Kalnay, and R. N. Hoffmann, 1988: Medium range lagged average forecasts. *Monthly Weather Review*, **116**, 402–416.

Daley, R. and T. Mayer, 1986: Estimates of global error from the global weather experiment observational network. *Monthly Weather Review*, **114**, 1642–1653.

Delle Monache, L., X. Deng, Y. Zhou, H. Modzelewski, G. Hicks, T. Cannon, R. B. Stull, and C. di Cenzo, 2004: Air quality ensemble forecast over the Lower Fraser Valley, British Columbia, Canada. In *27th NATO/CCMS Conference on Air Pollution Modeling*. Banff, Alberta.

Delle Monache, L. and R. B. Stull, 2003: An ensemble air quality forecast over western Europe during an ozone forecast. *Atmospheric Environment*, **37**, 3469–3474.

Draxler, R. R., 2003: Evaluation of an ensemble dispersion calculation. *Journal of Applied Meteorology*, **42**, 308–317.

Draxler, R. R., R. Dietz, R. J. Lagomarsino, and G. Start, 1991: Across North America Tracer Experiment (ANATEX): sampling and analysis. *Atmospheric Environment*, **25**, 2815–2836.

Ehrendorfer, M., 1994: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: Theory. *Monthly Weather Review*, **122**, 703–713.

Ehrendorfer, M. and R. M. Errico, 1995: Mesoscale predictability and the spectrum of optimal perturbations. *Journal of the Amospheric Sciences*, **52**, 3475–3500.

EPA, 1997: National ambient air quality standards for ozone; final rule. Technical Report Federal Register 62 (138), U.S. Environmental Protection Agency.

—, 2003: Guidelines for developing an air quality (Ozone and PM2.5) forecasting program. Technical Report EPA-456/R-03-002, U.S. Environmental Protection Agency.

Epstein, E. S., 1969a: The role of initial condition uncertainties in prediction. *Journal of Applied Meteorology*, **8**, 190–198.

—, 1969b: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Errico, R. M. and D. Baumhefner, 1987: Predictability experiments using a high-resolution limited-area- model. *Monthly Weather Review*, **113**, 488–504.

Errico, R. M. and T. Vukicevic, 1992: Sensitivity analysis using adjoint of the PSU-NCAR mesoscale model. *Monthly Weather Review*, **120**, 1644–1660.

Galmarini, S., R. Bianconi, R. Bellasio, and G. Graziani, 2001: Forecasting the consequences of accidental releases of radionuclides in the atmosphere from ensemble dispersion modeling. *Journal of Environmental Radioactivity*, **57**, 203–219.

Galmarini, S., R. Bianconi, W. Klug, T. Mikkelsen, R. Addis, S. Andronopoulos, P. Astrup, A. Bklanov, J. Bartniki, J. C. Bartzis, R. Bellasio, F. Bompay, R. Buckley, M. Bouzom, H. Champion, R. D'Amoursn, E. Davakis, H. Eleveld, G. T. Geertsema, H. Glaab, M. Kollax, M. Ilvonen, A. Manning, U. Pechinger, C. Persson, E. Polreich, S. Potemski, M. Prodanova, J. Saltbones, H. Slaper, M. A. Sofiev, D. Syrakov, J. H. Sørensen, L. Van der Auwera, I. Valkama, and R. Zelazny, 2004a: Ensemble dispersion forecasting-Part I: concept, approach and indicators. *Atmospheric Environment*, **38**, 4607–4617.

—, 2004b: Ensemble dispersion forecasting-Part II: application and evaluations. *Atmospheric Environment*, **38**, 4619–4632.

Gleeson, T. A., 1966: A causal relation for probabilities in synoptic meteorology. *Journal of Applied Meteorology*, **5**, 365–368.

—, 1967: On theoretical limits of predic. *Journal of Applied Meteorology*, **6**, 355–359.

Hacker, J., I. McKendry, and R. Stull, 2001: Modeled downward transport of a passive tracer over western North America during an Asian dust event in April 1998. *Journal of Applied Meteorology*, **40**, 1617–1628.

Hoffman, R. N. and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus*, **35**, 100–118.

Hogrefe, C., S. T. Rao, P. Kasibhatla, W. Hao, G. Sistla, R. Mathur, and J. McHenry, 2001: Evaluating the performance of regional-scale photochemical modeling system. part ii: Ozone predictions. *Atmospheric Environment*, **35**, 4175–4188.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using Kalman filter. *Weather and Forecasting*, **10**, 987–707.

Horvath, S. M. and D. J. McKee, 1994: Acute and chronic health effects of Ozone. In *Tropospheric ozone, human health and agricultural aspects*. Lewis Publisher, Boca Raton, Florida, pages 39–84.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX'98 ensemble forecasts. *Monthly Weather Review*, **129**, 73–91.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Monthly Weather Review*, **124**, 1225–1242.

Hubbard, M. C. and W. G. Cobourn, 1997: Development of a regression model to forecast ground-level Ozone concentrations in Louisville, Kentucky. *Atmospheric Environment*, **32**, 2637–2647.

Jacobson, M. Z., 1999: *Fundamentals of Atmospheric Modeling*. Cambridge University Press, New York.

Jonson, J. E., J. K. Sundet, and L. Tarrason, 2001: Model calculations of present and future levels of ozone and ozone precursors with a global and a regional model. *Atmospheric Environment*, **35**, 525–537.

Kahl, J. D. W., 1996: On the prediction of trajectory model error. *Atmospheric Environment*, **30**, 2945–2957.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, New York.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Willford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecast from multimodel superensemble. *Science*, **285**, 1548–1550.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, **102**, 409–418.

Liu, P.-W. G. and R. Johnson, 2002: Forecasting peak daily Ozone levels-1. a regression with time series errors model having a principal component trigger to fit 1991 Ozone levels. *Journal of the Air and Waste Management Association*, **52**, 1064–1074.

Lorenz, E. N., 1963: Deterministic non-periodic flow. *Journal of the Atmospheric Sciences*, **20**(130-141).

—, 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17**, 321–333.

—, 1993: *The Essence of Chaos*. University of Washington Press, Seattle.

McHenry, J. N., W. F. Ryan, N. L. Seaman, C. J. Coats Jr., J. Pudykiewicz, S. Arunachalam, and J. M. Vukovich, 2004: A real-time Eulerian photochemical model forecast system. *Bulletin of the American Meteorological Society*, **85**, 525–548.

Merrill, J. T., R. Bleck, and L. Avila, 1985: Modeling atmospheric transport to the Marshall Islands. *Journal of Geophysical Research*, **90**, 12927–12936.

Molina, M. J. and F. S. Roland, 1974: Stratospheric sink for chlorofluoromethanes: Chlorine atom catalysed destruction of ozone. *Nature*, **249**, 810–812.

Molteni, F. and T. N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quarterly Journal of Royal Meteorological Society*, **119**, 269–298.

Mullen, S. L. and D. P. Baumhefner, 1989: The sensitivity of numerical simulations of explosive oceanic cyclogenesis to changes in physics parameterizations. *Monthly Weather Review*, **116**, 2289–2339.

—, 1994: Monte Carlo simulation of explosive cyclogenesis. *Monthly Weather Review*, **122**, 1548–1567.

Nodop, K., R. Connolly, and G. Girardi, 1998: The field campaigns of the European Tracer Experiment (ETEX): overview and results. *Atmospheric Environment*, **32**, 4095–4108.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bulletin of the American Meteorological Society*, **74**, 49–64.

Pu, Z.-X., E. Kalnay, J. Sela, and I. Szunyogh, 1997: Sensitivity of forecast errors to the initial conditions with a quasi-inverse linear model. *Monthly Weather Review*, **125**, 2479–2503.

Rabier, F., E. Klinker, P. Courtier, and A. Hollingsworth, 1996: Sensitivity of forecast errors to initial conditions. *Quarterly Journal of Royal Meteorological Society*, **122**, 121–150.

Roeger, C., R. B. Stull, D. McClung, J. Hacker, X. Deng, and H. Modzelewski, 2003: Verification of mesoscale numerical weather forecasts in mountainous terrain for application to avalanche predicition. *Weather and Forecasting*, **18**, 1140–1160.

Ruiz-Suarez, J. C. and O. A. Mayora-Ibarra, 1995: Short-term Ozone forecasting by artificial neural networks. *Advances in Engineering Software*, **23**, 143–149.

Runeckles, V., 2002: Effects on vegetation and ecosystems. In Suzuki, D., editor, *A Citizen's guide to air pollution*. Vancouver, British Columbia, pages 177–216.

Ryan, W. F., 1994: Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmospheric Environment*, **29**, 2387–2398.

Scheele, M. P. and P. C. Siegmund, 2001: Estimating errors in trajectory forecasts using ensemble predictions. *Journal of Applied Meteorology*, **40**, 1223–1232.

Seinfeld, J. H. and S. N. Pandis, 1998: *Atmospheric Chemistry and Physics, from Air Pollution to Climate Change*. John Wiley and Sons Inc., New York.

Stohl, A., G. Wotawa, P. Seibert, and H. Kromp-Kolb, 1995: Interpolation errors in wind fields as a function of spatial and temporal resolution and their impact on different types of kinematic trajectories. *Journal of Applied Meteorology*, **34**, 2149–2165.

Straume, A. G., 2001: A more extensive investigation of the use of ensemble forecasts for dispersion model evaluation. *Journal of Applied Meteorology*, **40**, 425–445.

Straume, A. G., E. N. Koffi, and K. Nodop, 1998: Dispersion modeling using ensemble forecasts compared to ETEX measurements. *Journal of Applied Meteorology*, **37**, 1444–1456.

Stull, R. B., J. Hacker, and H. Modzelewski, 1997: Ensemble prediction of air-pollutant transport. In San Jose, R. and C. A. Brebbia, editors, *First International Conference on Measurements and Modelling in Environmental Pollution - MMEP 97*. Computational Mechanics Publications, Spain, pages 161–167.

Szunyogh, I., E. Kalnay, and Z. Toth, 1997: A comparison of Lyapunov vectors and optimal vectors in a low resolution GCM. *Tellus*, **49a**, 200–227.

Toth, Z., 2001: Ensemble forecasting in WRF. *Bulletin of the American Meteorological Society*, **82**, 695–697.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bulletin of the American Meteorological Society*, **74**, 2317–2330.

—, 1997: Ensemble forecasting at NCEP: the breeding method. *Monthly Weather Review*, **125**, 3297–3318.

Trevisan, A. and R. Legnani, 1995: Transient error growth and local predictability: a study in the Lorenz system. *Tellus*, **47a**, 103–117.

Tribbia, J. J. and D. P. Baumhefner, 1988: The reliability of improvements in deterministic short-range forecasts in the presence of initial state and modeling deficiencies. *Monthly Weather Review*, **116**, 2276–2288.

Vaughan, J., B. Lamb, C. Frei, R. Wilson, C. Bowman, C. Figueroa-Kaminsky, S. Otterson, M. Boyer, C. Mass, M. Albright, J. Koenig, A. Collingwood, M. Gilroy, and N. Maykut, 2004: A numerical daily air quality forecast system for the Pacific Northwest. *Bulletin of the American Meteorological Society*, **85**, 549–561.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multi-model ensemble system. *Monthly Weather Review*, **129**, 729–747.

Warner, T. T., R.-S. Sheu, J. F. Bowers, R. I. Sykes, G. C. Dodd, and D. S. Henn, 2002: Ensemble simulation with coupled atmospheric dynamic and dispersion models: illustrating uncertainties in dosage simulations. *Journal of Applied Meteorology*, **41**, 488–504.

WMO, 1986: Atmospheric ozone: 1985. global ozone research and monitoring project. Technical Report 16, World Meteorological Organization.

—, 1990: Report of the international Ozone trends panel: 1988. Global Ozone research and monitoring project. Technical Report 18, World Meteorological Organization.

—, 1998: Scientific assessment of ozone depletion: 1998. Technical report, World Meteorological Organization.

Chapter 2

An Ensemble Air-quality Forecast Over Western Europe During an Ozone Episode

2.1 Introduction

[1] Ensemble forecasting of the weather has been increasingly evaluated over the past decade, and found to provide better accuracy than any single Numerical Weather Prediction (NWP) model (Wobus and Kalnay, 1995; Molteni et al., 1996; Du et al., 1997; Hamill and Colucci, 1997; Toth and Kalnay, 1997; Stensrud et al., 1998; Krishnamurti et al., 1999; Evans et al., 2000; Kalnay, 2003). Transfer of this technique to air-quality (AQ) modeling can potentially

[1] A version of this chapter has been published. Delle Monache, L., and R. Stull, 2003: An ensemble air-quality forecast over western Europe during an ozone episode, *Atmospheric Environment*, **37**, 3469-3474. Published as "Fast Track", i.e., "...for papers that contain important and topical results whose significance merits fast publication.".

yield similar benefits. This note briefly reviews ensemble methods for NWP, and evaluates one method applied to AQ modeling.

NWP models are extremely complex computer codes that approximate with finite differences the nonlinear interaction among dynamic, thermodynamic, radiative, cloud microphysical, turbulent, and many other processes. Different models verify best on different days, usually in ways that cannot be anticipated. Sometimes one model is better because of its physical parameterizations, other times because of the underlying discretization methods, and other times because of different initial conditions.

But when output from different NWP models, or from different realizations of the same model, are considered together as an ensemble, it is found that their ensemble average is usually more accurate that any individual model realization. More specifically, the ensemble average is not the most accurate every day. But when verified for many cases, the ensemble average is the most accurate for more of the days than any other single ensemble member (Kalnay, 2003).

Modern photochemical AQ models are equally large and complex sets of computer code that describe hundreds to thousands of chemical reactions, plume rise from myriads of sources, dispersion induced by different turbulence mechanisms, and transport in boundary layers of varying stratification and in complex terrain. To make matters worse, the AQ models are often driven by NWP models, which introduce their own signature and imperfections. Different AQ models are better for different air-pollution episodes, also in ways that cannot always be anticipated. Sometimes one model might be better because of its choice of chemical reactions or rate constants, other times because of the turbulence and boundary-layer description, and other times because of more representative emission or meteorological inputs (Hass et al., 1997;

Russell and Dennis, 2000).

For NWP, ensembles have been created with different inputs (Toth and Kalnay, 1993; Molteni et al., 1996; Wandishin et al., 2001) (initial conditions, boundary conditions), different parameterizations within a single model (Stensrud et al., 1998) (physics packages, parameter values), different numerics within a single model (Thomas et al., 2002) (finite difference approximations and solvers, grid resolutions, compiler optimizations), and different models (Hou et al., 2001; Wandishin et al., 2001). This has been done in order to allow the ensemble to take into account different sources of uncertainties. For AQ, the ensemble-mean can be created similarly with different inputs (background concentrations, emissions inventories, meteorology), different parameterizations within a single model (chemistry mechanisms, rate constants, advection and dispersion packages), different numerics within a single model (finite difference approximations and solvers, grid resolutions, compiler optimizations), and different models.

For both NWP and AQ, the different models usually include differences in numerics and parameterizations. In this Chapter the multi-model (each model having different initial and boundary conditions) approach is tested, based on a reanalysis of the intercomparison study done by Hass et al. (1997). Using predicted ozone time series of concentrations from four models, an ensemble-mean is computed and tested against the observations, and its performance is compared with the performance of each single model.

2.2 Data

Hass et al. (1997) intercompared four photochemical dispersion models: the European Monitoring and Evaluation Programme (EMEP) model (Simpson, 1993), the European Air Pollution Dispersion (EURAD) model (Hass et al., 1993), the Long-Term Ozone Simulation (LOTOS) model (Builtjes, 1991) and the Regional Eulerian Model with three different chemistry schemes (REM3) (Stern, 1994). EMEP is a one-layer Lagrangian photochemical model. EURAD is a comprehensive, multi-layer Eulerian model. LOTOS and REM3 are three-layer Eulerian models, but REM3 also includes three different chemistry schemes.

These models have different computational domains (with different horizontal and vertical resolution), different initial and boundary conditions (for both emissions and meteorological fields), and different model formulations (different advection schemes and chemical mechanisms).

Each model also uses different emission data. The differences can be of the order of two (EURAD-LOTOS for the biogenic VOCs) or three (REM3-LOTOS for terpene). Moreover, the way the models split the VOC amount into anthropogenic and biogenic categories are significantly different. Also there are large differences in the importance terpene assumed in the four models.

The meteorological fields driving the four models are different. EMEP and LOTOS are driven by the Numerical Weather Prediction model of Gronås and Hellevik (1982), and both take the mixing heights from observation. EURAD is driven by the MM5 model (Grell et al., 1994), nudged by large-scale analysis from the European Center for Medium Weather Forecasts. The REM3 meteorological field is derived entirely from observations. Thus, the differences between the resulting meteorological fields are quite large. There are also differences in how the models consider the interactions between the meteorology and the chemistry. For example, to compute chemical-reaction rates, EMEP uses an average boundary-layer temperature, whereas the other models use the layer-average temperature for each layer within the boundary layer.

Hass et al. (1997) selected a case-study episode that covered the six-day time period of 31 July through 5 August, 1990. This was a hot summer period with high ozone concentrations (up to 140 pbbv) in northwestern and central Europe. A high-pressure ridge formed on 31 July over the North Sea, resulting in dry warm continental air over western Europe. This synoptic system moved toward Denmark on 2 August, and then to Poland on 4 August. The ozone episode ended after a frontal passage between 4-5 August. Further details about the models and the ozone episode, as well as about the emission data, can be found in Hass et al. (1997).

We verify the four model predictions, and the ensemble-mean against the observed ozone concentrations at five different sites. The sites are Sibton (United Kingdom), Kollumerwaard (The Netherlands), Waldhof (Germany), Lindenberg (Germany) and Rőervick (Sweden). The ensemble is computed as a simple, unweighted average over outputs from the four models.

2.3 Results

Figure 2.1 shows the ozone time series as predicted by the models and as observed at Sibton (U.K.), from 31 July to 5 August 1990. This is an example where the ensemble-mean concentration benefits from the spread of the predicted concentrations with respect to the observed values. The ensemble average is overall the best forecast, except the fourth day of the episode, when all the models considerably under-predict the observed ozone concentration. Table 2.1 shows, for each station, the following statistics:

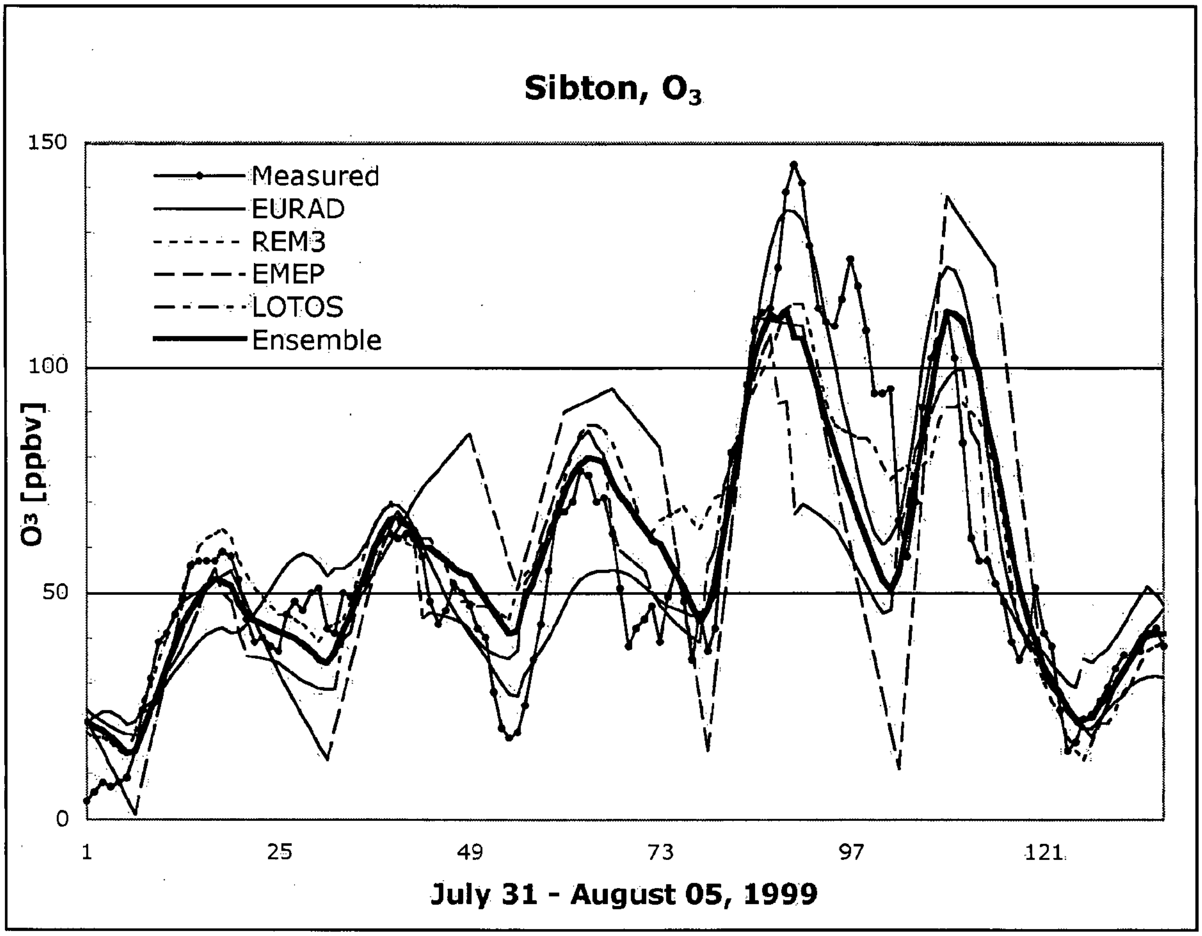

Figure 2.1: Observed, modeled, and ensemble-mean ozone concentration (ppbv) for the episode at the site Sibton (U.K.).

normalized gross error (GE) (herein "gross error", for hourly observed values of $O_3 > 60$ ppbv)

$$GE = \frac{1}{N}\sum_{i=1}^{N}\frac{|C_p(x,t_i) - C_o(x,t_i)|}{C_o(x,t_i)} \tag{2.1}$$

and

unpaired peak prediction accuracy ($UPPA$)

$$UPPA = \frac{C_p(x,t')_{max} - C_o(x,t')_{max}}{C_o(x,t')_{max}} \tag{2.2}$$

where N is the number of hourly concentrations over the episode, $C_o(x,t_i)$ is the observed value at the monitoring station located at x for hour t_i , $C_p(x,t_i)$ is the predicted value at the monitoring station located at x for hour t_i, $C_o(x,t')_{max}$ is the maximum 1-h observed concentration at a specific monitoring station over one day, and $C_p(x,t')_{max}$ is the maximum 1-h predicted concentration at a specific monitoring station over one day.

These two statistical parameters are included in the US Environment Protection Agency guidelines (EPA, 1997) to analyze historical ozone episodes using photochemical grid models. The EPA acceptable performance upper limit values are + 35 % for gross-error, and ± 20 % for unpaired peak prediction accuracy. In Table 2.1 the bold values are the ones that satisfy those criteria.

The gross-error values satisfy the EPA criteria in every case. The ensemble gives consistently the best or the second-best forecasts over the six monitoring stations. The second-best overall performance in terms of gross-error are given by both REM3 and LOTOS, while EURAD and EMEP have somewhat poorer performance. Performances from all the models are quite erratic compared to the smoother behavior of the ensemble, suggesting that the ensemble

Table 2.1: Model ozone-performance statistics [gross error (GE) and unpaired peak prediction accuracy (UPPA)] for 31 July to 5 August episode, at the sites Sibton (UK), Kollumerwaard (the Netherlands), Waldhof (Germany), Lindenberg(Germany) and Rőervik (Sweden). Values in bold are within the EPA acceptable performance criteria.

Station	**Model**	**GE (%)**	**UPPA (%)**					
			31 July	1 Aug.	2 Aug.	3 Aug.	4 Aug.	5 Aug.
Sibton	EURAD	**17**	**-13**	**10**	-28	**-6**	**-1**	**-19**
	REM3	**15**	**8**	**11**	**12**	-21	-25	**-7**
	EMEP	**29**	**-6**	28	23	-23	**11**	**-9**
	LOTOS	**22**	**-6**	**7**	**11**	-26	**-19**	**0**
	Ensemble	**16**	**-9**	**6**	**3**	-22	**-9**	**-2**
Kollumerw.	EURAD	**24**	**-12**	**-19**	-29	**3**	25	**-10**
	REM3	**20**	24	28	**12**	**10**	**5**	-41
	EMEP	**22**	**18**	20	-24	**-12**	**-8**	123
	LOTOS	**21**	**18**	22	23	-24	-24	22
	Ensemble	**13**	**11**	**9**	**-19**	**-6**	**-7**	42
Waldhof	EURAD	**17**	**-9**	1	**0**	**-10**	**-31**	**-9**
	REM3	**10**	**2**	41	35	22	**-3**	-39
	EMEP	**20**	**3**	**11**	**1**	**-10**	-25	**6**
	LOTOS	**22**	**-7**	**16**	**7**	**-7**	-35	**3**
	Ensemble	**13**	-5	**15**	**8**	**-3**	-29	**-11**
Lindenberg	EURAD	**17**	**-6**	36	40	25	**6**	**18**
	REM3	**28**	28	97	97	67	45	**-5**
	EMEP	**19**	**-10**	**13**	43	42	**15**	**15**
	LOTOS	11	**-19**	31	51	33	**4**	**5**
	Ensemble	**13**	**-6**	39	54	39	**12**	**7**
Rőervik	EURAD	**22**	**8**	**12**	**14**	**10**	33	**13**
	REM3	**24**	23	**10**	**2**	**-5**	-43	26
	EMEP	**15**	35	**-2**	**4**	**-1**	**-5**	**13**
	LOTOS	**19**	**19**	**13**	25	**-9**	-35	39
	Ensemble	**16**	19	**5**	**8**	**-5**	-35	25

might be able to take into account most of the uncertainties by filtering out the unpredictable components.

The unpaired peak accuracy for the six-day episode and over the five stations shows good performance of both EURAD and the ensemble, both having 73 % of the unpaired peak accuracy values within the EPA acceptance criteria. They are followed by EMEP with 66 %, LOTOS with 53 %, and REM3 with 43 %. A similar ranking is obtained when only observed peak ozone values above 60 ppbv are considered (not shown here).

2.4 Discussion

The case study investigated here suggests that a photochemical-model ensemble average can give a better result than a single model deterministic forecast. Because the limited size of the data set available, and most importantly because of the limited spatial separation among the stations relative to the coarse grid spacing of the models domains, these results are not spatially independent and cannot be generalized until further investigations are made.

Ideally the ensemble should be composed of state-of-the-art photochemical models that are run starting from the best possible emissions scenario, as well as with the best possible meteorological fields. The meteorological fields can be indeed different for different photochemical models, since each of them is obtained differently (from different mesoscale models, and then different starting analyses, map projections, domain grids, etc.). Moreover, the different model formulations, i.e., the different advection and turbulence transport schemes and the different chemical mechanisms implemented in each model, should assure a good ensemble spread, which is desirable to define likely bounds of possible pollutant-concentration fields. The uncertainty in each of those components is partially averaged out by the ensemble approach.

The ensemble tested in this study has many of those desirable features. For example, the differences between the emission data of each model (sometimes of the order of two or three), in both the initial and boundary conditions, can take in account the uncertainty in the emissions estimates (a factor of three or more), and is the dominant limitation in photochemical model performance (Russell and Dennis, 2000). As shown clearly by Hass et al. (1997) with backward trajectories, the difference in the modeled meteorological fields will strongly influence the final concentrations, and the ensemble might account for those uncertainties as well.

For NWP ensembles, errors typically grow linearly at first, and nonlinearly later (Kalnay, 2003). However, the linear period might me reduced in AQ ensembles because of the strongly nonlinear nature of many chemical reactions. For this reason, the differences among AQ ensemble members may account for the uncertainties associated with each component of the AQ process more rapidly than what is observed for NWP ensembles.

Because not all of the photochemical models used NWP meteorological fields as input for this study, it is not clear if the benefit of the ensemble accrued because of the ensemble of photochemical air-pollution forecasts, or because of the ensemble of input meteorological fields. The benefit of using a NWP ensemble for the meteorological input has been proven in other ensemble applications for air-quality forecasts, namely for transport and dispersion without the chemical processes (Stull et al., 1997; Straume et al., 1998; Dabberdt and Miller, 2000; Galmarini et al., 2001; Straume, 2001; Warner et al., 2002).

Another aspect that emphasizes the utility of the ensemble approach, is the fact that the model grids used in this study are completely different in both resolution and location. Again, these differences lead to different parcel trajectories, and this would allow the ensemble to take into account the uncertainties related to the different but plausible choices of the grid location and resolution adopted from each of the models that form the ensemble.

Once an ensemble forecasting system is implemented sufficiently long at a specific site, the ensemble-mean capabilities might be improved by taking into account the past performances of each single model in conditions similar to present conditions. Namely, one can by performing a weighted ensemble-mean, give more importance to the forecasts that historically perform better than the others. This approach has not been tested here due to the small size of the data set available.

Ensemble forecasting can also provide probabilistic forecasts based on the spread of the ensemble members. For instance, the probability that ozone concentration can be greater than a specific threshold on a specific site, can be easily computed as the ratio of the ensemble members that satisfy this condition, over the others that do not.

2.5 References for Chapter 2

Builtjes, P. J. H., 1991: The LOTOS model results. Comparison of three models for long term photochemical oxidant in Europe. Technical Report EMEP/MSC-W, DNMI.

Dabberdt, W. F. and E. Miller, 2000: Uncertainty, ensembles and air quality dispersion modeling: applications and challenges. *Atmospheric Environment*, **34**, 4667–4673.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Monthly Weather Review*, **125**, 2427–2459.

EPA, 1997: National ambient air quality standards for ozone; final rule. Technical Report Federal Register 62 (138), U.S. Environmental Protection Agency.

Evans, R. E., M. S. J. Harrison, and R. Graham, 2000: Joint mediumrange ensembles from the Met. Office and ECMWF systems. *Monthly Weather Review*, **128**, 3104–3127.

Galmarini, S., R. Bianconi, R. Bellasio, and G. Graziani, 2001: Forecasting the consequences of accidental releases of radionuclides in the atmosphere from ensemble dispersion modeling. *Journal of Environmental Radioactivity*, **57**, 203–219.

Grell, G. A., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation PENN State/NCAR Mesoscale Model (MM5). Technical Report NCAR/TN-398 +STR, National Center for Amospheric Research.

Gronås, S. and O. Hellevik, 1982: A limited area prediction model at the Norwegian Meteorological Institute. Technical Report 61, Norwegian Meteorological Institute.

Hamill, T. M. and S. J. Colucci, 1997: Verification of ETA-RSM shortrange ensemble forecasts. *Monthly Weather Review*, **126**, 1322–1327.

Hass, H., P. J. H. Builtjes, D. Simpson, and R. Stern, 1997: Comparison of model results obtained with several European regional air quality models. *Atmospheric Environment*, **31**, 3259–3279.

Hass, H., A. Ebel, H. Feldmann, H. Jakobs, and M. Memmesheimer, 1993: Evaluation studies with a regional chemical transport model (EURAD) using air quality data from EMEP monitoring network. *Atmospheric Environment*, **27**, 867–887.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX'98 ensemble forecasts. *Monthly Weather Review*, **129**, 73–91.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, New York.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Willford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecast from multimodel superensemble. *Science*, **285**, 1548–1550.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The new ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of Royal Meteorological Society*, **122**, 73–119.

Russell, A. and R. Dennis, 2000: NARSTO critical review of photochemical models and modeling. *Atmospheric Environment*, **34**, 2283–2324.

Simpson, D., 1993: Photochemical model calculations over Europe for two extended summer periods: 1985 and 1989. Model results and comparison with observations. *Atmospheric Environment*, **27**, 912–943.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 1998: Ensemble forecasting of mesoscale convective systems. In *12th Conference on Numerical Weather Prediction*. American Meteorological Society, Phoenix, Arizona, pages 265–268.

Stern, R., 1994: Entwicklung und anwendung eines drei-dimensionalen photochemischen ausbreitung-modelles mit verschiedenen chemischen mechanismen. Technical Report 8/1, Serie A, Institute for Meteorology, Free University Berlin.

Straume, A. G., 2001: A more extensive investigation of the use of ensemble forecasts for dispersion model evaluation. *Journal of Applied Meteorology*, **40**, 425–445.

Straume, A. G., E. N. Koffi, and K. Nodop, 1998: Dispersion modeling using ensemble forecasts compared to ETEX measurements. *Journal of Applied Meteorology*, **37**, 1444–1456.

Stull, R. B., J. Hacker, and H. Modzelewski, 1997: Ensemble prediction of air-pollutant transport. In San Jose, R. and C. A. Brebbia, editors, *First International Conference on Measurements and Modelling in Environmental Pollution - MMEP 97*. Computational Mechanics Publications, Spain, pages 161–167.

Thomas, S. J., J. P. Hacker, M. Desgagné, and R. B. Stull, 2002: An ensemble analysis of forecast errors related to floating point performance. *Weather and Forecasting*, **17**, 898–906.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bulletin of the American Meteorological Society*, **74**, 2317–2330.

—, 1997: Ensemble forecasting at NCEP: the breeding method. *Monthly Weather Review*, **125**, 3297–3318.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multi-model ensemble system. *Monthly Weather Review*, **129**, 729–747.

Warner, T. T., R.-S. Sheu, J. F. Bowers, R. I. Sykes, G. C. Dodd, and D. S. Henn, 2002: Ensemble simulation with coupled atmospheric dynamic and dispersion models: illustrating uncertainties in dosage simulations. *Journal of Applied Meteorology*, **41**, 488–504.

Wobus, R. and E. Kalnay, 1995: Three years of operational prediction of forecast skill. *Monthly Weather Review*, **123**, 2132–2148.

Chapter 3

Ozone Ensemble Forecasts. A New Ensemble Design

3.1 Introduction

[1] The harmful effects of tropospheric ozone on humans (Horvath and McKee, 1994; Brauer and Brook, 1995), vegetation (Runeckles, 2002) and materials (Brown et al., 2001) motivate the issuance of air-quality (AQ) forecasts, and the need to limit and control anthropogenic emissions of ozone precursors. To alert the population about impending AQ degradation, Dabberdt and Miller (2000) discussed the need for an operational AQ forecast system. The first experiences with this kind of system are described in Delle Monache et al. (2004), McHenry et al. (2004) and Vaughan et al. (2004). A probabilistic approach to AQ forecasting is recommended by the U.S. Weather Research Program and its Prospectus Development Team on Air-Quality

[1] A version of this chapter has been accepted for publication. Delle Monache, L., X. Deng, Y. Zhou, and R. B. Stull, 2005: Ozone ensemble forecasts. Part I: a new ensemble design, *manuscript accepted in November 2005 to be published in the Journal of Geophysical Research.*

Forecasting (Dabberdt et al., 2003) due to the chaotic nature of the atmosphere.

Some dynamical systems are called chaotic if they show divergent behavior, meaning that two different solutions starting from similar but not identical initial states would eventually diverge nonlinearly in solution space (Lorenz, 1963). In such cases we don't know a priori which of the two solutions is closest to the true evolution of the system.

The atmosphere exhibits this behavior, and is thus a chaotic system. We are not able to accurately measure the initial state of the atmosphere, due to instrumentation errors and large gaps between observation sites. Moreover, we are able to solve only a simplified version of the equations describing the atmosphere, and those solutions are usually numerical approximations; i.e., they are sources of error as well. As a consequence, there is an upper limit in time on the predictive skill of weather forecasts. The ensemble approach is one method to represent the time evolution of the probability density function (PDF) describing the atmosphere's initial state and its uncertainty. Practically, the PDF can be represented by a limited set of points (e.g., Leith, 1974). The evolution of each of those points would be a member of the ensemble. Each of those members should ideally represent an equally likely evolution of the dynamical system.

It has been found for numerical weather prediction (NWP) that the ensemble-mean is more accurate that an individual model realization, when verified for many cases. NWP ensembles have been created using different model input values (Toth and Kalnay, 1993; Molteni et al., 1996; Wandishin et al., 2001), different parameterizations within a single model (Stensrud et al., 1998), different numerical schemes (Thomas et al., 2002), and different models (Hou et al., 2001; Wandishin et al., 2001). This allows the ensemble to take into account different sources of uncertainty.

The ensemble technique can yield similar benefits to AQ prediction, because there are similar model complexities and constraints. Different AQ models can be better for different air-pollution episodes, in ways that cannot always be anticipated. Similar to NWP ensembles, AQ ensemble members can be created with different meteorological and/or emission inputs, different parameterizations within a single model, different numerics within a single model, and different models.

For NWP ensembles, errors typically grow linearly at first, and nonlinearly later. However, the linear period might be reduced in AQ ensembles because of the strongly nonlinear nature of many chemical reactions. For this reason, the differences among AQ ensemble members may account for the uncertainties associated with each component of the AQ process more rapidly than what is observed for NWP ensembles.

In Chapter 2 it has been discussed the benefit of the ensemble approach in studies involving not only pollutant transport, but also the associated photochemical reactions. Their ensemble was composed of four Chemistry Transport Models (CTMs), and was tested for a 6-day summer period over five monitoring stations in northwestern and central Europe. The ensemble approach presented in that study showed promising results, performing better than the models individually, including good performance for ozone peak-value prediction.

Another successful implementation of the ensemble approach can be found in Galmarini et al. (2004b), where the authors describe an application to long-range transport and dispersion studies. They used the data collected during the ETEX experiment (Nodop et al., 1998) to quantitatively estimate the concepts and parameters introduced in Part I of their coupled papers (Galmarini et al., 2004a). They tested a multi-model ensemble dispersion system by considering several operational long-range transport and dispersion models used to support decision making in case of accidental releases. The median member of the forecast ensemble exhibited the best forecast skill.

McKeen et al. (2005) present results for a multi-model (i.e., seven CTMs) Ozone Ensemble Forecast System (OEFS), statistically evaluated for 53 days (summer 2004), against 340 monitoring stations over eastern U.S. and southern Canada. The high correlation coefficients and low root-mean-square-error (RMSE) points to the ensemble mean as the preferred forecast when compared to any individual model.

Recently O'Neill and Lamb (2005) presented an interesting intercomparison of the Community Multiscale Air Quality Model (CMAQ) (Byun and Ching, 1991) with the California Photochemical Grid Model (CALGRID) (Carmichael et al., 1992). They tested an ensemble averaged prediction based on the two CTM models run with different meteorology and chemical mechanisms. They found the ensemble skillful for the 8-hour averaged forecasts, while with the 1-hour predictions the ensemble mean did not necessarily showed more skill than the single deterministic runs. However, the standard deviation about the 1-hour mean forecast provides a useful measure of overall model uncertainty.

A new OEFS is presented here using predicted ozone concentrations from 12 different ensemble members. An ensemble-mean is computed (as a linear average of the ensemble-member predicted hourly concentrations) and tested against observations from five different stations over the Lower Fraser Valley (LFV), British Columbia (BC), Canada (see Figure 3.1). This is a region where ozone modeling is particularly challenging, because of the complex coastal mountain setting (McKendry and Ludgren, 2000). OEFS performance is compared with the performance of each single forecast for a 5-day period (11-15 August 2004).

Galmarini et al. (2004b) showed that the ensemble-median (the median of the ensemble-

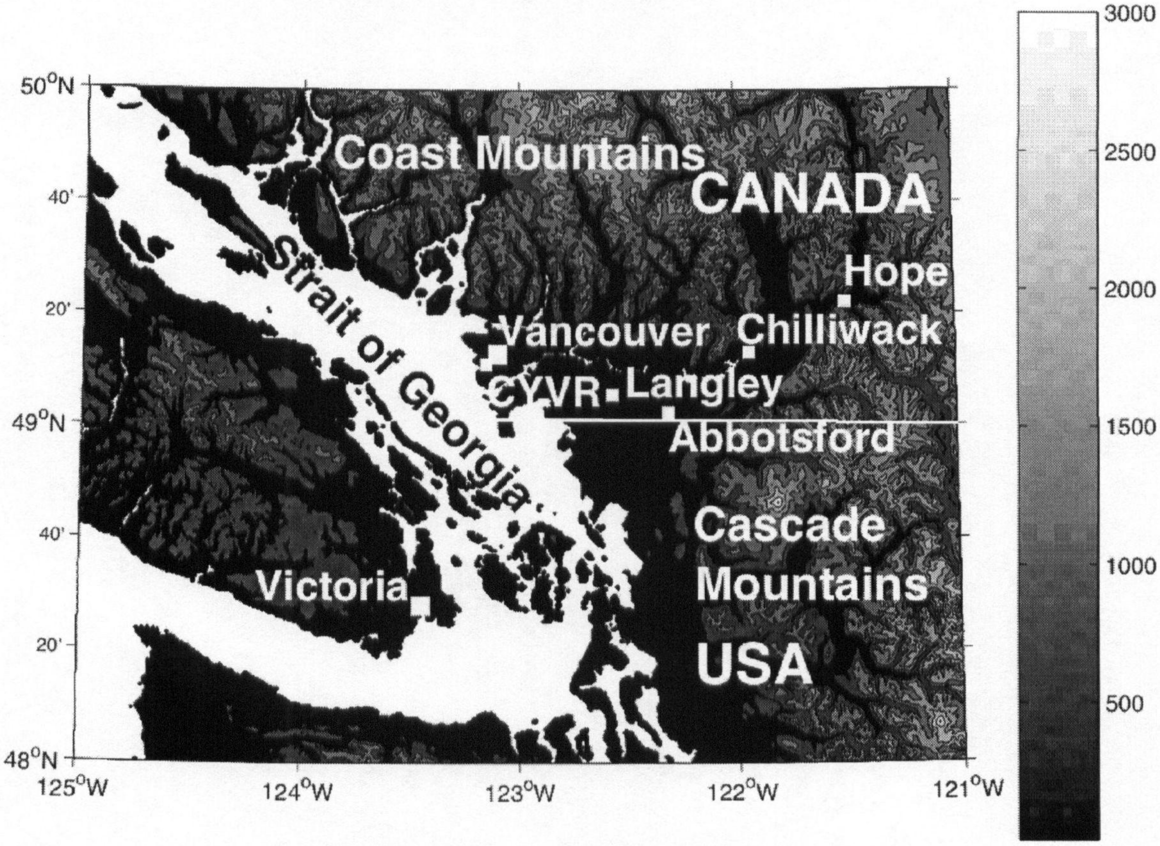

Figure 3.1: The Lower Fraser Valley is a floodplain spanning the ozone stations of Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope. The triangular valley is widest near CYVR along the coast of the Georgia Strait, and tapers to a narrow gorge between steep mountain walls near Hope. Shading (vertical bar at right) indicates terrain elevation above sea level.

member predicted hourly concentrations) has better forecast skill than the ensemble-mean. For ensembles with many members that all capture likely forecast outcomes, one would expect statistically that the ensemble mean and median member should be nearly identical. However, if some ensemble members are distant outliers because of any number of model or initial-condition errors, then they would not contribute to a realistic estimate of the probability distribution of realistic forecast outcomes. This is a particular problem if there is a cluster of outliers. For such cases the ensemble average is unduly biased by the outliers, allowing the one median ensemble member to give the best forecast. In this study the ensemble mean resulted in a more skillful forecast than the ensemble median, implying that we did not have a problem with unphysical or unrepresentative outliers.

For situations where ensemble outliers might be problem, there are some solutions. One is to build a record of error variances for each member based on past forecasts, and then weight each member inversely with its error to compute a weighted ensemble mean. Another is to reduce their systematic errors (Chapter 5), and then combine these corrected forecasts into an uniformly-weighted average.

Section 3.2 describes the case study and the data, while a detailed description of the OEFS is given in Section 3.3. Section 3.4 presents the results and their analysis, and a discussion followed by the conclusions can be found in the last section.

3.2 Case Study and Data

The LFV lies across the western edge of the Canada/US border (Figure 3.1). The main metropolitan area is located at the northwest end of the valley, where the Greater Vancouver region has a population slightly greater than two million people. The valley is triangular-shaped, oriented approximately west-to-east, with the Strait of Georgia on the west side, the Coast Mountains to the north, and the Cascade Mountains range limiting the valley's southeastern side.

The synoptic conditions observed during the period 11-15 August 2004 were typical of conditions that lead to high ground-level ozone concentrations in the LFV, as described by McKendry (1994). Those conditions are associated with a northward progressing low-level thermal trough, extending from California northward through Oregon and Washington State reaching the southern part of BC. An associated stationary upper-level ridge was situated across southern BC. The upper-level ridge started to weaken on 14 August, allowing clouds to spread over the LFV on 15 August, leading to lower observed ozone concentrations at four stations out of five. Over the LFV, sea-breeze circulations combine with valley and slope flows to make ozone modeling (that includes photochemistry) quite challenging (McKendry and Ludgren, 2000).

This study uses hourly observed ozone concentrations from five stations across the LFV: Vancouver International Airport (CYVR) (urban), Langley (suburban), Abbotsford (urban), Chilliwack (suburban), and Hope (rural) (Figure 3.1). These stations span the LFV from west to east, and being apart one from each other more than 12 km, they fall in different grid cells for all the forecasts. The observed ozone hourly concentrations for the period 11-15 August 2004 vary considerably from west to east. This reflects the easterly advection of ozone and its precursors by the sea-breeze circulation, leading to higher concentrations further inland. Thus, at CYVR the values are low (peak value always below 50 ppbv) and close to typical background summer values, due to its proximity to the coast. At Langley (further inland), the observed maxima for the 5-day period are between 60 and 70 ppbv, with the lower peak value observed on 15 August. Ozone maximum values between 60 and 80 ppbv are observed at Abbotsford, while at Chilliwack the observed peak is above 70 ppbv except on 15 August. The ozone concentrations at Hope (furthest inland) exceed 82 ppbv (the Canadian National Ambient Air Quality Objective for maximum 1-h average concentration) during the first four days (with values between 85 and 90 ppbv). At all five stations, the nighttime values are very low (< 15 ppbv). Secondary nocturnal maxima ozone concentrations are observed at all stations as discussed by Salmond and McKendry (2002).

Studies of ozone photochemistry in the LFV (Ainslie, 2004, with a scaling-level model as described in Section 1.3.2) show that the present and projected AQ is in a regime affected roughly equally by NO_x and VOC emissions (Figure 3.2). Namely, in a maximum-ozone-concentration isopleth plot as a function of NO_x and VOC emissions, the state of the LFV is above the ridgeline of ozone relative maxima. Those results (specific to the LFV), are considered in building the ensemble design presented in the next session.

3.3 Ensemble Design

At the University of British Columbia (UBC), the Mesoscale Compressible Community (MC2) NWP model (Benoit et al., 1997) and the Penn State/NCAR mesoscale (MM5) model (Grell et al., 1994) have been running daily for several years (http://weather.eos.ubc.ca/wxfcst/). MC2 is a fully compressible, non-hydrostatic model using semi-implicit semi-Lagrangian techniques. The model is initialized using the National Centers for Environmental Prediction (NCEP) North American Mesoscale (NAM) model at 108-km grid spacing. One-way nesting is applied to produce model output at horizontal grid spacing of 108, 36, 12, 4, and 2 km. MM5 is a fully compressible, non-hydrostatic, primitive-equation meteorological model that uses a terrain-following sigma (non-dimensionalized pressure) vertical coordinate. The MM5 model is initialized from the same analysis and for the same five nested grids as MC2, but with 2-way nesting.

Both MC2 and MM5 produce meteorological fields that are used in this study to drive the U.S. Environmental Protection Agency (EPA) Models-3/CMAQ Chemistry Transport Model

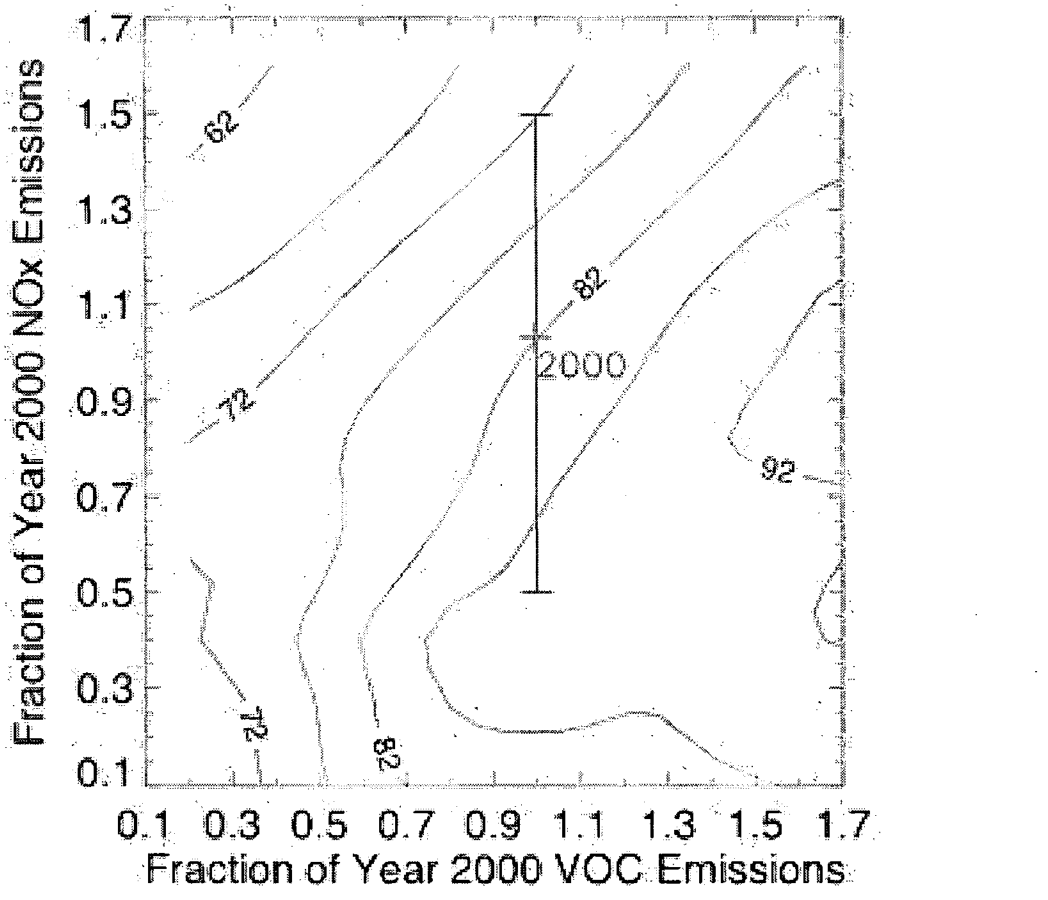

Figure 3.2: Isopleths of maximum ozone concentration (ppbv) are given as a function of year 2000 VOC and NO_x emissions over the Lower Fraser Valley (adapted from Ainslie (2004)). The total annual VOC and NO_x emissions are 111,196 and 99,897 tonnes, respectively (GVRD, 2002). The vertical bar shows the ± 50% NO_x used for the ensemble perturbations.

(CTM) (Byun and Ching, 1991). CMAQ has been run at UBC daily real-time for two and a half years (Delle Monache et al., 2004). The CBM-IV chemical mechanism (Gery et al., 1989), and the Modified Euler Backward Iterative (MEBI) chemistry solver (Huang and Chang, 2001) are used. The emissions used as input to CMAQ are prepared using the Sparse Matrix Operator Kernel Emission (SMOKE) system (Coats, 1996). The boundary conditions are a time-invariant vertical concentration profile for the coarser domain (based on typical summertime background ozone concentrations in the LFV), while the finer grids are initialized each day with the previous day's prediction.

Ideally, for the ensemble to be a skillful forecast, the ensemble members should span all the uncertainties associated with different phases of the modeling process: initial conditions and boundary conditions, meteorological and emission fields, numerics, chemical mechanisms, etc. Unfortunately, to consider all those modeling aspects would require an ensemble with an unfeasibly large number of ensemble members. For this reason, we present an OEFS that considers only the uncertainties associated with the meteorological and emission fields. These fields are considered to cause the main uncertainties in photochemical modeling (Russell and Dennis, 2000). For example, NO_x emission estimates can be in error by a factor of two or more (Hanna et al., 2001).

A related question is what ensemble size and perturbed attributes are necessary for capturing most of the forecast uncertainty, based on ensemble-mean metrics. We demonstrate here that a limited-size ensemble with only meteorology and emission perturbations can indeed yield an ensemble average that is better than individual members, on average.

A flowchart of the OEFS tested in this Chapter is shown in Figure 3.3. CMAQ is run with a 12-km horizontal resolution domain covering southern BC, Washington State, and the

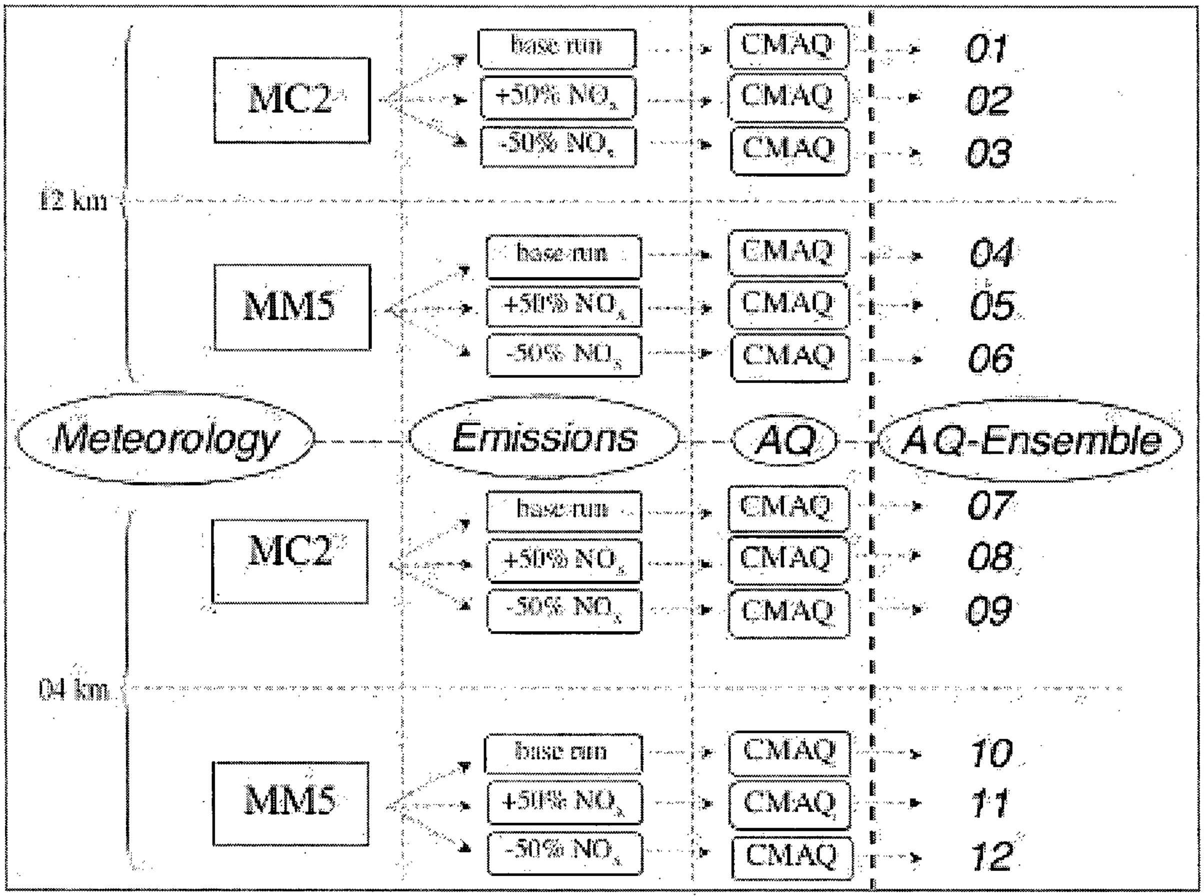

Figure 3.3: The 12-member (01, 02, $\cdots$, 12) Ozone Ensemble Forecast System is sketched. It is formed with four different meteorological fields (MC2 at 4 and 12 km, and MM5 at 4 and 12 km), and three different emission scenarios: a control run (CTRL), a run with plus 50 % NO_x (NOXP), and a run with minus 50 % NO_x (NOXN).

northern portion of Oregon, with a nested 4-km resolution domain covering southwestern BC and northwestern Washington State. Both domains are centered over the LFV. MC2 and MM5 provide the meteorological inputs for CMAQ, for the 12 and 4-km domains. Moreover, for each of the four possible meteorological input combinations, CMAQ is run with three emission scenarios: a control run (CTRL), a run with 50 % more NO_x (NOXP), and a run with 50 % less NO_x (NOXN) (also see Figure 3.2). These scenarios were chosen because NO_x emissions are mainly anthropogenic (Jacobson, 1999) and strongly influence ground-level ozone concentrations (Steyn et al., 1997). This leads to a system with 12 ensemble members (01, 02, $\cdots$, 12), as shown in Figure 3.3. An example (Abbotsford, 11-15 August) of the ensemble members (black lines) and their ensemble-mean (thick black line) temporal evolution, compared with the observed ozone concentrations (circles), can be found in Figure 3.4.

Since the six 12-km resolution ensemble members are run for 48 hours, the second half of the $(N-1)^{th}$ forecast day can be added to the N^{th} forecast day ensemble forecast. Figure 3.5 depicts the resulting 18-member OEFS tested in this study, built as a lagged-averaged ozone ensemble (see Section 3.4.4).

3.4 Results and Analysis

3.4.1 Verification Statistics

The forecast skill of each ensemble member and the ensemble-mean has been evaluated using the following statistical parameters:

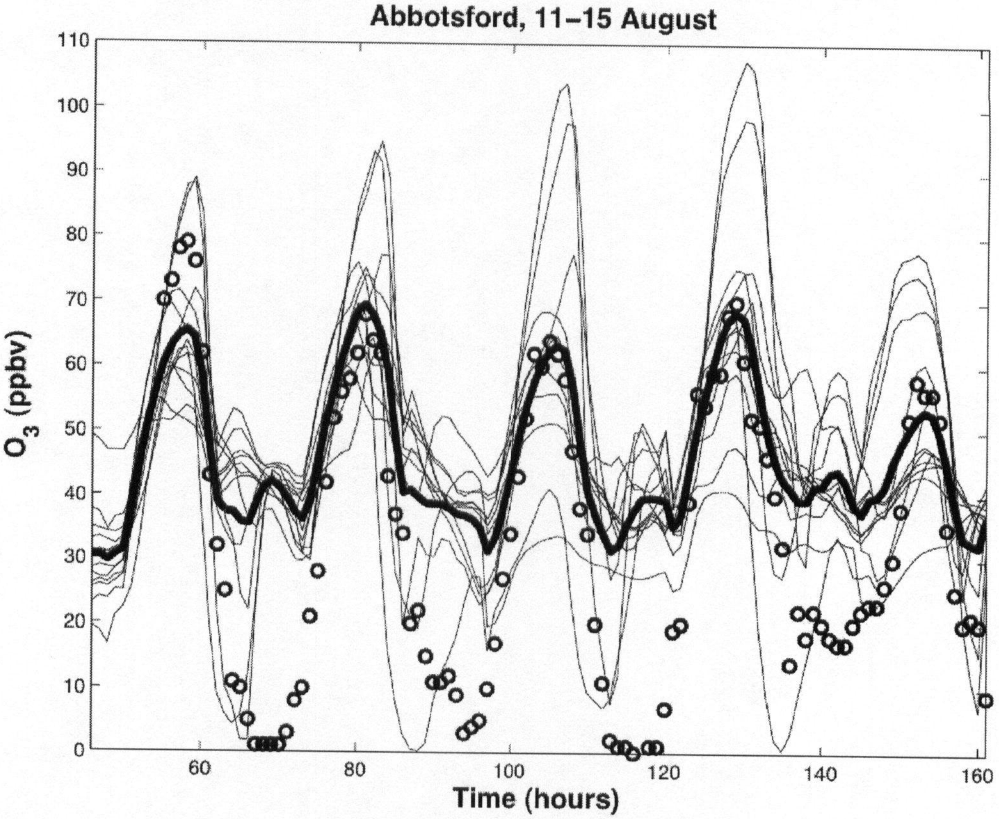

Figure 3.4: The 12 ensemble members (black lines) and the ensemble-mean (thick black line) predictions, along with the observations (circles), at Abbotsford, 11-15 August 2004.

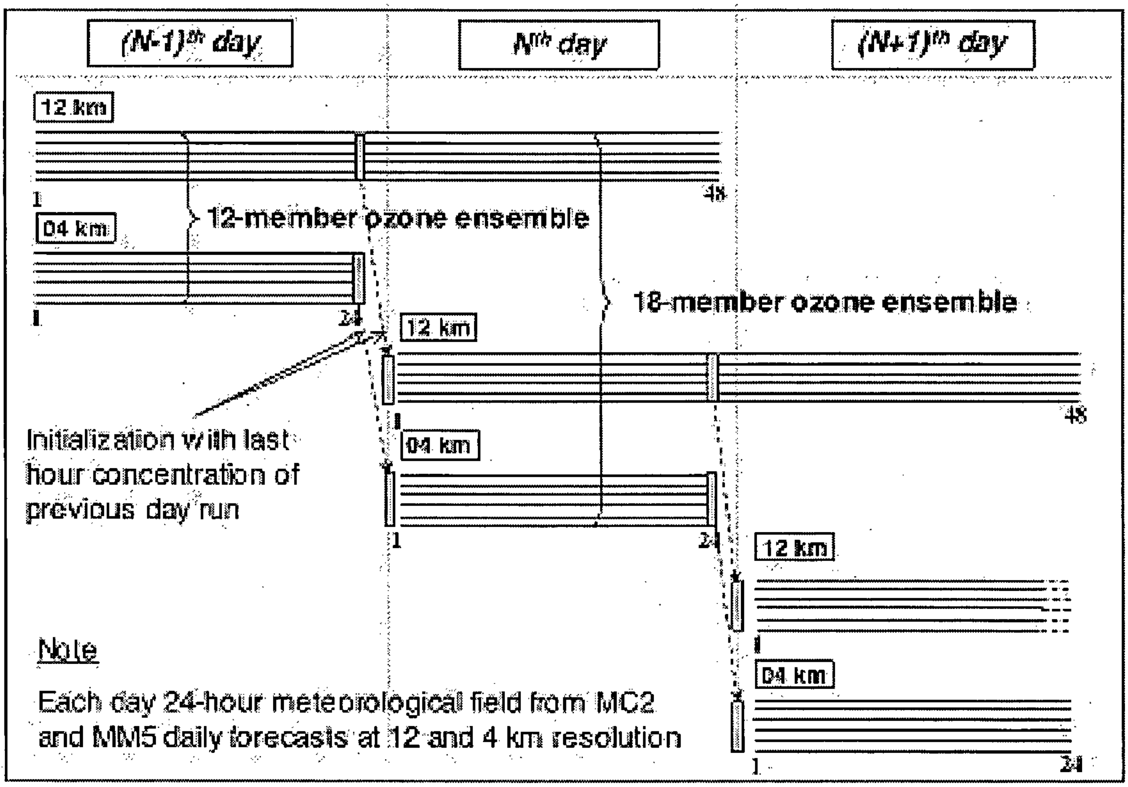
(N-1)th day
Nth day
(N+1)th day
12 km
1
48
04 km
12-member ozone ensemble
1
24
18-member ozone ensemble
12 km
1
48
Initialization with last hour concentration of previous day run
04 km
1
24
12 km
1
Note
Each day 24-hour meteorological field from MC2 and MM5 daily forecasts at 12 and 4 km resolution
04 km
1
24

Figure 3.5: The 18-member Ozone Ensemble Forecast System (OEFS) is shown. The six 12-km resolution ensemble members are run for 48 hours. The second half of the $(N-1)^{th}$ forecast day can be added the N^{th} day 12-member OEFS to form a lagged-averaged ozone 18-member ensemble.

- Pearson product-moment coefficient of linear correlation (herein "correlation"):

$$corr(station) = \frac{\sum_{t=1}^{N_{hour}}[C_o(t, station) - \overline{C_o(station)}][C_p(t, station) - \overline{C_p(station)}]}{\sqrt{\sum_{t=1}^{N_{hour}}[C_o(t, station) - \overline{C_o(station)}]^2 \sum_{t=1}^{N_{hour}}[C_p(t, station) - \overline{C_p(station)}]^2}} \tag{3.1}$$

- normalized gross error (herein "gross error", for hourly observed values of $O_3 > 30$ ppbv):

$$gross\ error(station) = \frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} \frac{|C_p(t, station) - C_o(t, station)|}{C_o(t, station)} \tag{3.2}$$

- root mean square error (RMSE):

$$RMSE(station) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C_p(t, station) - C_o(t, station)]^2} \tag{3.3}$$

- unpaired peak prediction accuracy (UPPA):

$$UPPA = \frac{1}{N_{day}} \sum_{day=1}^{N_{day}} \frac{|C_p(day, station)_{max} - C_o(day, station)_{max}|}{C_o(day, station)_{max}} \tag{3.4}$$

where N_{hour} is the number of 1-h average concentrations over the 5-day period, N_{day} is the number of days, $C_o(t, station)$ is the 1-h average observed concentration at a monitoring station for hour t, $C_p(t, station)$ is the 1-h average predicted concentration at a monitoring station for hour t, $\overline{C_o(station)}$ is the average of 1-h average observed concentrations at a monitoring station over the 5-day period, $\overline{C_p(station)}$ is the average of 1-h average predicted concentrations at a monitoring station over the 5-day period, $C_o(day, station)_{max}$ is the maximum 1-h average observed concentration at a monitoring station over one day, and $C_p(day, station)_{max}$

is the maximum 1-h average predicted concentration at a monitoring station over one day.

The gross error and UPPA are included in the U.S. EPA guidelines (EPA, 1991) to analyze historical ozone episodes using photochemical grid models. The EPA acceptable performance upper-limit values are + 35 % for gross error, and ± 20 % for unpaired peak prediction accuracy. UPPA is computed here as an average (over the five days available) of the absolute value of the normalized difference between the predicted and observed maximum at each station (Equation 3.4). Thus, UPPA is non-negative; hence, only the + 20 % acceptance performance upper limit is used in the next sections.

We selected this set of statistics for the following reasons. We choose correlation to get an indirect indication of the differences between the predicted and measured ozone time series at a specific location. The closer the correlation is to one, the better is the correspondence of timing of ozone maxima and minima between the two signals. RMSE (measured in ppbv) gives important information about the skill in predicting the magnitude of ozone concentration, even though alone it does not draw a complete picture of a forecast value. It is very useful also for understanding ensemble averaging effects, because it can be decomposed into systematic and unsystematic components as discussed in detail in Section 3.4.2.

The gross-error statistic has been considered in this analysis because it is included in the U.S. EPA guidelines (EPA, 1991). Also, being computed for hourly observed values of $O_3 >$ 30 ppbv, it gives useful information about the forecast skill for higher concentration values, which are important for health-related issues. It gives information about the error magnitude (as RMSE), but as a portion of the observed ozone concentration (i.e., is measured in %).

UPPA (%) is also used because it measures the ability of the forecasts to predict the ozone peak maximum on a given day. Peak concentrations have been in the past the main concern for public health. However, in recent years over midlatitudes of the Northern Hemisphere, a rising trend of background ozone concentrations has been observed, while peak values are steadily decreasing (Vingarzan, 2004).

3.4.2 12-member OEFS Results

The performance of the OEFS presented in Section 3.3 has been tested by computing the statistical parameters introduced in Section 3.4.1, using the data described in Section 3.2.

Correlation

Figure 3.6 shows the results for the correlation between the observed hourly ozone concentration and the predicted concentrations from the 12 ensemble members and the ensemble-mean. Those values are computed for the 5-day period from 11 to 15 August 2004, and at five different stations: CYVR, Langley, Abbotsford, Chilliwack and Hope.

Generally, correlation values tend to be lower moving towards the east side of the LFV, with all the forecasts having their poorest performance at Hope. Indeed Hope is located in a very steep narrow valley (less than 4 km wide), which none of the models are able to resolve. Because the 12 km runs do not see this valley, in the afternoon the ozone plume is advected past Hope (instead of being trapped there), resulting in decreasing values (after the plume passage) while in reality the concentration is increasing. Also, during the nighttime return flow (a land breeze, going back westward) is established, causing the 12 km run to bring back the plume, and resulting in increasing predicted concentrations when the observed ozone is decreasing. This causes negative correlation values for the 12 km runs, as shown in Figure 3.6. Thus, the ensembles using finer resolution runs have better correlation values at Hope and

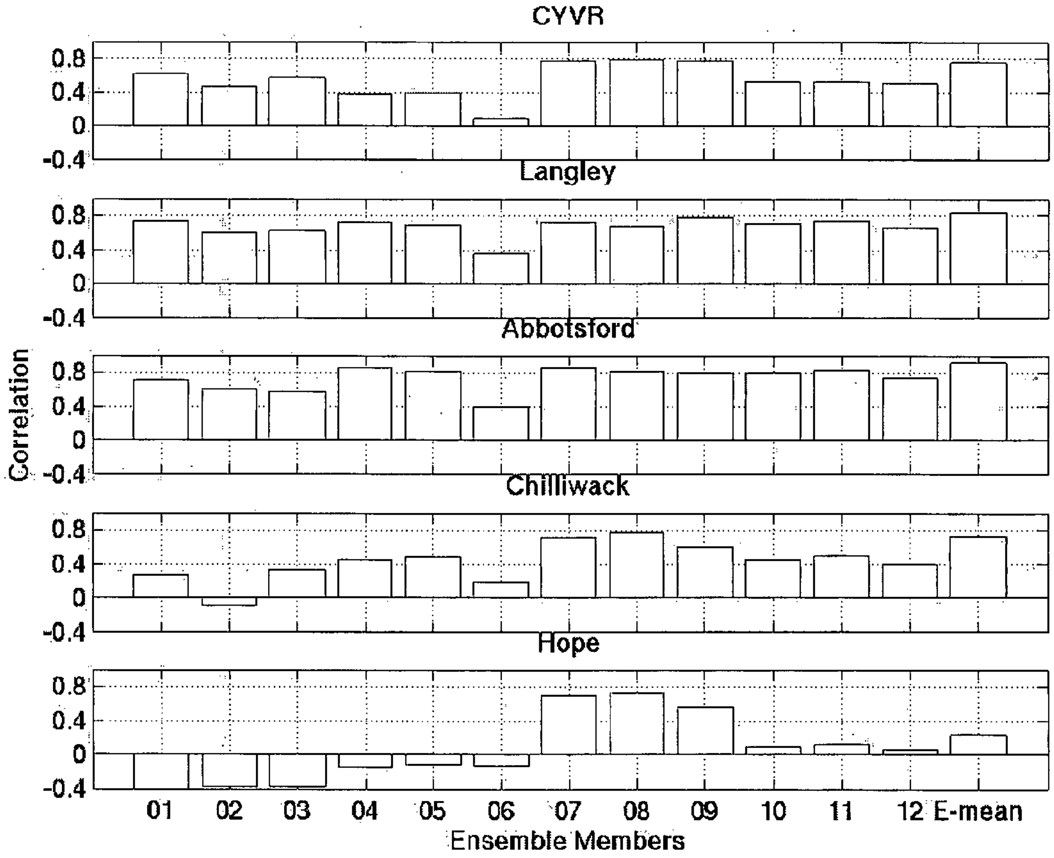

Figure 3.6: Correlation values between observed and predicted ozone 1-h average concentrations are plotted at five stations [Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope], for the 12-member Ozone Ensemble Forecast System (01, 02, $\cdots$, 12) and the ensemble-mean (E-mean), for the 5-day period 11-15 August 2004. Values are within the interval $[-1,\ 1]$, with correlation $= 1$ being the best possible value.

Table 3.1: Ranking for correlation of the 12 ensemble members (01, 02, $\cdots$, 12) and the ensemble-mean (E-mean) at the Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack and Hope stations. The lowest sum of rankings indicates the best overall performance.

	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**Ensemble-mean**
CYVR	5	10	6	12	11	13	3	1	2	8	7	9	4
Langley	4	12	11	6	8	13	5	9	2	7	3	10	1
Abbotsford	10	11	12	2	6	13	3	5	7	8	4	9	1
Chilliwack	11	13	10	8	6	12	3	1	4	7	5	9	2
Hope	13	12	11	10	8	9	2	1	3	6	5	7	4
Ranking Sum	33	58	50	38	39	60	16	17	18	36	24	44	12

Chilliwack (particularly with MC2; i.e., forecasts 07, 08 and 09), where the topography is most complex. Spatial resolutions even finer than 4 km would be needed to better capture these topographic effects.

CYVR is located adjacent to the water in the Georgia Strait, and the meteorological models have difficulty capturing accurately the thermally driven sea-breeze flows generated by the water/land discontinuity. At this location the finer resolution runs tends to have better correlation with the observation (again, particularly with MC2), probably because they better represent the complex coastline and the associated land-use data. The ensemble-mean has the best performance at Langley and Abbotsford, and is second best at Chilliwack.

Table 3.1 shows for each station the ranking (from 1 to 13) of each ensemble member and the ensemble-mean, where the best (highest) correlation value has a ranking of 1, and the worst (lowest) has 13. Overall the ensemble-mean has the best ranking as measured by the lowest sum of rankings. The only ensemble members with similar (but worse) skill are 07, 08, and 09, with members 08 having a number of first rankings.

The ensemble-mean has mediocre skill at CYVR and Hope because both stations are located in areas where all the individual ensemble members have difficulties, as explained

Table 3.2: Rankings similar to Table 3.1, but for gross error.

	01	02	03	04	05	06	07	08	09	10	11	12	Ensemble-mean
CYVR	1	5	4	6	9	2	13	11	12	7	10	3	8
Langley	2	7	12	5	4	8	13	11	10	6	9	3	1
Abbotsford	2	5	11	3	4	10	13	12	9	6	8	7	1
Chilliwack	9	8	1	5	7	11	12	13	10	4	6	3	2
Hope	11	12	10	6	7	13	1	2	9	5	3	8	4
Ranking Sum	25	37	38	25	31	44	52	49	50	28	36	24	16

above. The correlation values are significantly improved (closer to one) with Kalman-filter (KF) post-processing, as shown in Chapter 5.

Gross error

The gross-error results are shown in Figure 3.7, and the rankings are summarized in Table 3.2. Overall the ensemble-mean is the best for these cases when compared to each ensemble member, as indicated by the ranking sum. Forecast 08 for the correlation has similar performances to the ensemble-mean, but has large gross error (very poor skill), except at Hope where it ranks second. Note that the 4-km MC2-driven ensemble members (07, 08 and 09) at CYVR, Langley and Abbotsford have relatively poor skill using the gross-error metric, but have much better performance using the correlation metric.

The ensemble-mean is well within the 35 % EPA acceptance value at Langley, Abbotsford and Chilliwack. At CYVR and Hope the ensemble-mean has the highest gross-error values, confirming the difficulties for all the ensemble members at those two locations. In Chapter 5 it is shown that application of the KF post-processing improves (brings closer to zero) the gross-error performance of most forecasts, with an improvement up to 20 %.

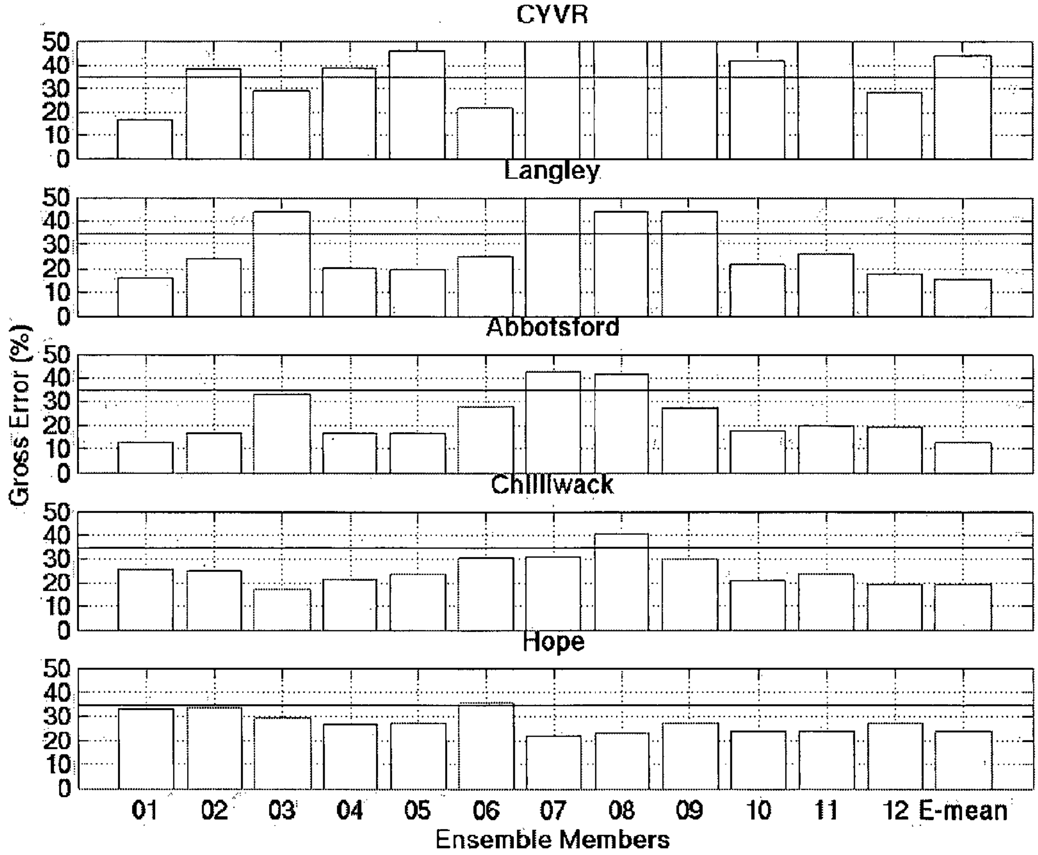

Figure 3.7: Similar to Figure 3.6, but for gross-error values (%). The continuous line is the EPA acceptance value (+ 35 %). Values are within the interval [0, + ∞], with a perfect forecast having gross error = 0.

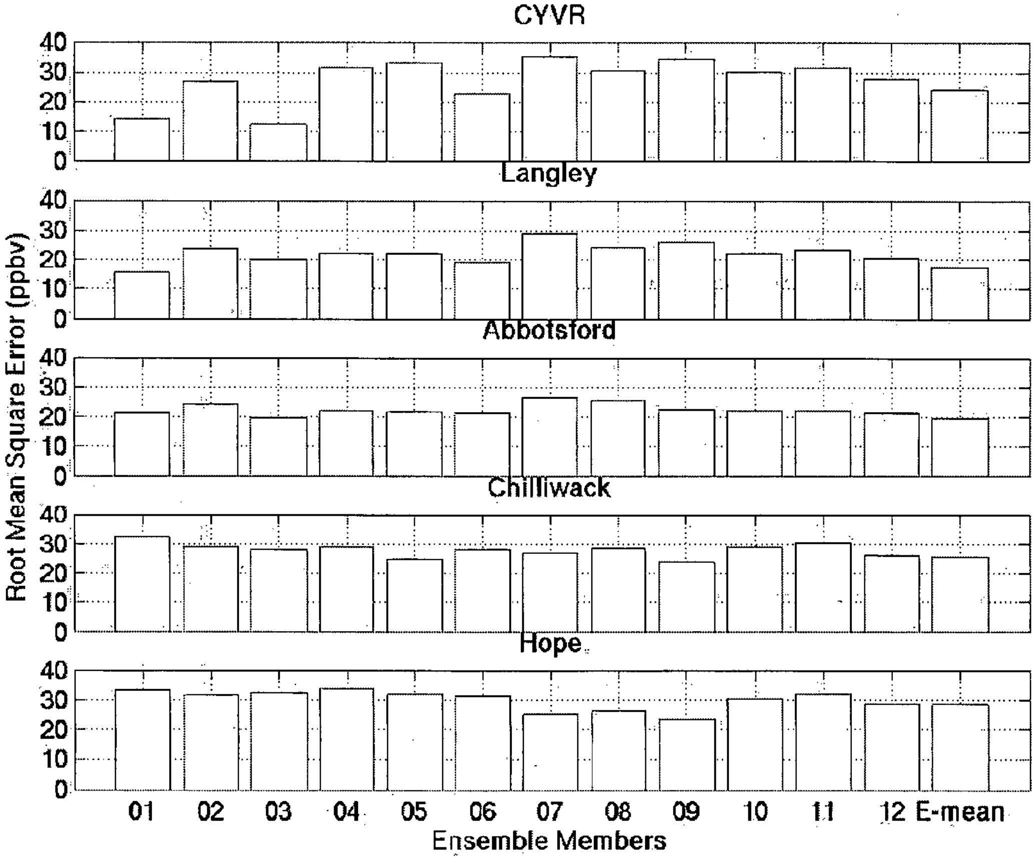

Figure 3.8: Similar to Figure 3.6, but for root mean square error (RMSE) values (ppbv). Values are within the interval [0, + ∞], with a perfect forecast when RMSE = 0.

RMSE

The RMSE results are shown in Figure 3.8 and summarized in Table 3.3. In general, the values of this statistical parameter are between 20 and 30 ppbv. However, the KF correction presented in Chapter 5 shows substantial improvements up to 20-25 %, with values often between 10 and 20 ppbv. Nevertheless, the ensemble mean is the best. Forecast 03 ranks first at CYVR and Abbotsford, but still is worse than the ensemble-mean at three stations (Langley, Chilliwack and Hope). Forecast 03 is one of the worst for the correlation metric, and

Table 3.3: Rankings similar to Table 3.1, but for root mean square error.

	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**Ensemble-mean**
CYVR	2	5	1	9	11	3	13	8	12	7	10	6	4
Langley	1	10	4	6	7	3	13	11	12	8	9	5	2
Abbotsford	4	11	1	7	6	3	13	12	10	8	9	5	2
Chilliwack	13	10	6	9	2	7	5	8	1	11	12	4	3
Hope	12	8	11	13	9	7	2	3	1	6	10	4	5
Ranking Sum	32	44	23	44	35	23	46	42	36	40	50	24	16

worse than average for gross error. Again, the ranking sum shows that the ensemble mean is the best.

RMSE can be separated in different components. One decomposition was proposed by Willmott (1981). First, an estimate of concentration $C^*(t, station)$ is defined as follows:

$$C^*(t, station) = a + bC_o(t, station) \tag{3.5}$$

where a and b are the least-square regression coefficients of $C_p(t, station)$ and $C_o(t, station)$ (the predicted and observed ozone concentrations, respectively, as defined in Section 3.4.1). Then the following two quantities can be defined:

$$RMSE_s(station) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C^*(t, station) - C_o(t, station)]^2} \tag{3.6}$$

$$RMSE_u(station) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C^*(t, station) - C_p(t, station)]^2} \tag{3.7}$$

where $RMSE_s(station)$ is the RMSE systematic component, while $RMSE_u(station)$ is the unsystematic one. $RMSE_s$ indicates the portion of error that depends on errors in the model, while $RMSE_u$ depends on random errors, on errors resulting by a model skill deficiency in predicting a specific situation, and on initial-condition errors. The following relates RMSE to its components:

$$RMSE^2 = RMSE_s{}^2 + RMSE_u{}^2 \tag{3.8}$$

Ensemble averaging is expected to reduce some of the unsystematic component of the error (i.e., $RMSE_u$), while the systematic component ($RMSE_s$) should be little affected by the averaging process. In fact, since $RMSE_s$ reflects errors in the model affecting each individual forecast similarly, it should not be reduced (when compared with the ensemble members) for the ensemble mean.

Figure 3.9 shows the RMSE systematic (bottom bar) and unsystematic components (top bar). CYVR (and to a lesser extent Langley) shows among the highest $RMSE_u$ values, indicating an intrinsic lack of predictive skill at this location, as already discussed in Section 3.4.2. Martilli and Steyn (2004) discuss the effects of the superimposed valley, slope, and thermal flows over the LFV. Often the pollution plume is transported during night over the Georgia Strait waters as a result of the combination of several transport processes. This makes it very challenging for the models to accurately predict the spatial and temporal evolution of ozone concentration in near-water locations, such as CYVR, where the over-strait pool of pollutants can be re-advected over land by the daytime sea breeze. The 12-km runs (forecasts 01-06) have their highest systematic error at Hope. All these forecasts poorly reproduce the real topography at this location, and this leads to systematic misrepresentations of ozone temporal and spatial distributions. Conversely, the 4-km runs have their highest systematic error at CYVR (in particular for MC2 driven runs; ensemble members 07-09), where their ability to capture complex terrain more accurately than the 12-km runs is not an advantage, since at

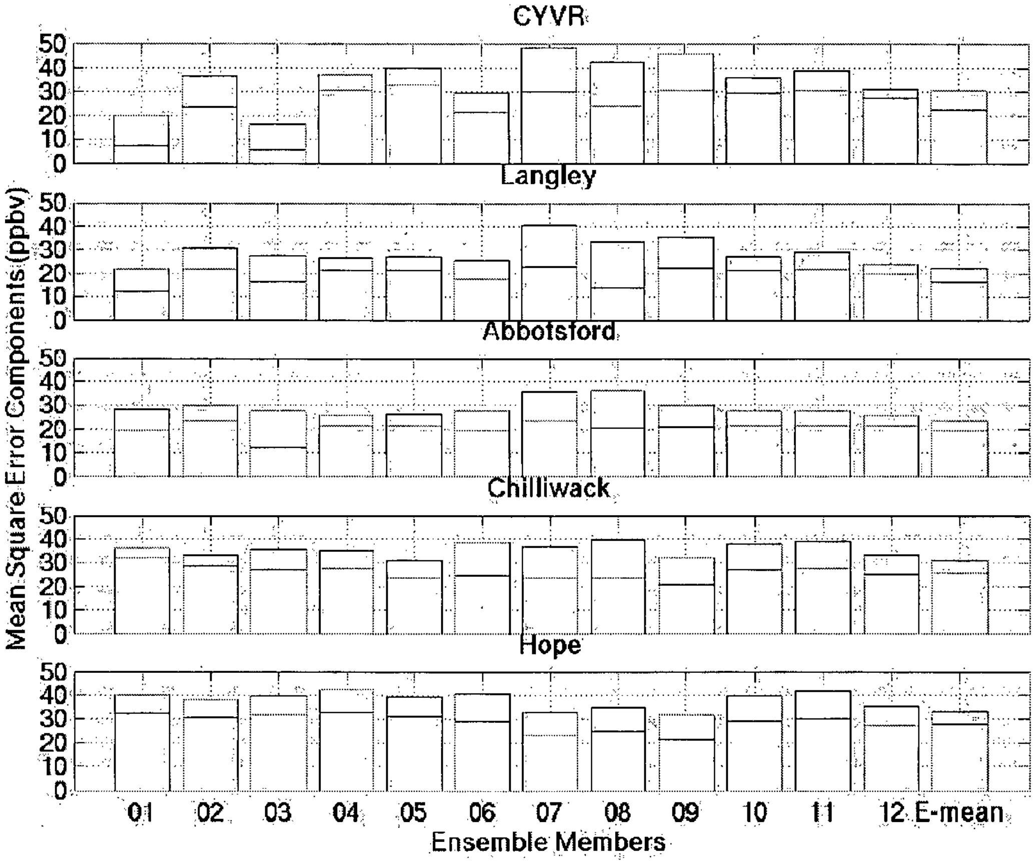

Figure 3.9: Similar to Figure 3.8, but segregating the root-mean-square-error into its systematic (bottom bar) and unsystematic components (top bar). The sum of these components squared equals the square the root-mean-square-error (Equation 3.8).

CYVR the terrain is flat.

Overall, the ensemble mean has among the lowest $RMSE_u$ when compared with the other forecasts, being the second best after forecast 12 (MM5, at 4 km, with NOXN) and before forecast 04 (MM5, at 12 km, NOXP). The ensemble mean has the lowest $RMSE_u$ at Hope, the second best at Abbotsford, the third at Chilliwack, the fourth at Langley and the sixth at CYVR. Conversely, the ensemble mean $RMSE_s$ is never the lowest and is always close to the average $RMSE_s$ of the individual forecasts. This confirms the usefulness of ensemble averag-

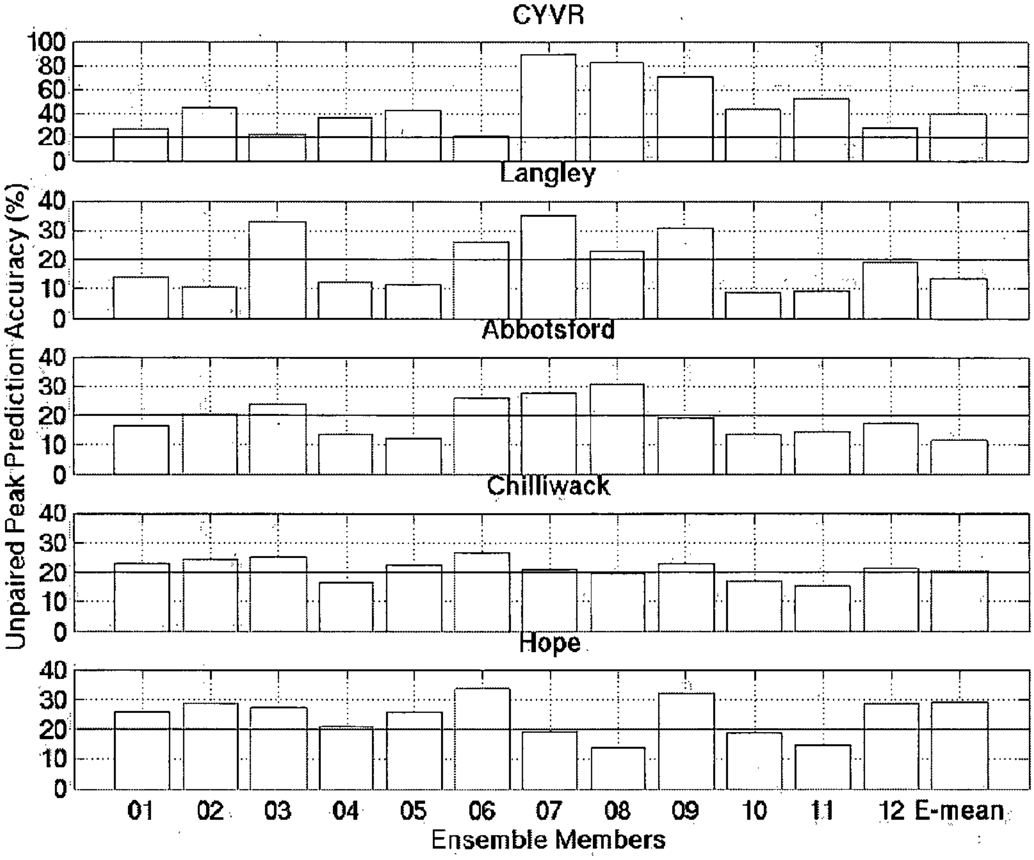

Figure 3.10: Similar to Figure 3.6, but for unpaired peak prediction accuracy (UPPA) values (%). The continuous lines are the EPA acceptance values (+ 20 %). Values are within the interval [0, + ∞], with a perfect peak forecast when UPPA = 0.

ing: it is able to remove part of the unpredictable components of the physical and chemical processes involved in the ozone fate, resulting in a more skillful forecast when compared to any deterministic ensemble member.

UPPA

Figure 3.10 shows the UPPA results. At CYVR, forecasts 07, 08 and 09 largely overestimate the observed ozone peak concentration, even though they have at this station a high correlation

Table 3.4: Rankings similar to Table 3.1, but for unpaired peak prediction accuracy.

	01	02	03	04	05	06	07	08	09	10	11	12	Ensemble-mean
CYVR	3	9	2	5	7	1	13	12	11	8	10	4	6
Langley	7	3	12	5	4	10	13	9	11	1	2	8	6
Abbotsford	6	9	10	3	2	11	12	13	8	4	5	7	1
Chilliwack	9	11	12	2	8	13	6	4	10	3	1	7	5
Hope	6	10	8	5	7	13	4	1	12	3	2	9	11
Ranking Sum	31	42	44	20	28	48	48	39	52	19	20	35	29

value (close to 0.8). The UPPA rankings in Table 3.4 are computed using absolute values, so that under and over-prediction of the observed peak concentrations have the same weight when the ranking is computed. For this parameter the ensemble-mean is the best only at Abbotsford when compared with the 12 individual ensemble members. It has a slightly better than average performance at CYVR, Langley at Chilliwack, and it has poor performance at Hope. A possible reason for the poor average performance (i.e., high ranking sum) of the ensemble mean with UPPA (observed in this study), is that ensemble averaging might lead to excessive smoothing of the peak values.

Except at CYVR, forecasts 10 and 11 (MM5, at 4 km, with CTRL and NOXP) have good forecast skill for UPPA, while for all other statistical parameters they are average or worse than average. In Chapter 5 is shown that application of the KF post-processing modestly improves (brings closer to zero) the UPPA performance.

3.4.3 11-member OEFS Results

Since the previous analysis shows that different ensemble members contribute differently to the ensemble-mean performance, we eliminate each individual member in turn from the 12-member ensemble, and re-compute the four statistical parameters for the 5-day period and five stations, for the resulting 11-member ensemble. This way, one can gauge the effect of each single ensemble member on the ensemble-mean.

Figure 3.11 shows the median (over the five stations) of the correlation of the 11-member ensemble-mean, where each bar represents the correlation value for the ensemble-mean without the one corresponding ensemble member indicated in the label below the bar. Superimposed as a dashed line is the correlation value for the full 12-member ensemble. If the value shown is below the dashed line, it implies that the ensemble-mean without that specific member has worse performance, and vice versa.

First, all the correlation values are between 0.7 and 0.8, regardless of which forecast is removed from the ensemble. The forecasts with MC2 at 4 km (07, 08 and 09) removed give generally worse correlation values, and the contrary is true for the runs with MM5 at 4 km (10, 11, and 12). In other words, the ensemble average is better if MC2 at 4 km is included. Also, all the runs without MM5 at 12 km give better correlation, while the runs with MC2 at 12 km improve the correlation two times out of three.

Figure 3.12 shows a similar analysis, but for the gross error. All the values are close to 19 ppbv without any evident trend, except that for all the runs at 12 km, NOXN is better than NOXP, which are both better than the CTRL run.

Similar results for RMSE are shown in Figure 3.13. If the value is below the dashed line, it implies that the ensemble-mean without that specific member has better performance. Here the differences are more pronounced, with maximum difference (of about 10 %) between the value of the ensemble-mean without forecast 03 and the one without forecast 05. The only ensemble members that positively contribute to the RMSE ensemble-mean value (i.e., increasing RMSE when removed, which is equivalent to reducing errors when included in the

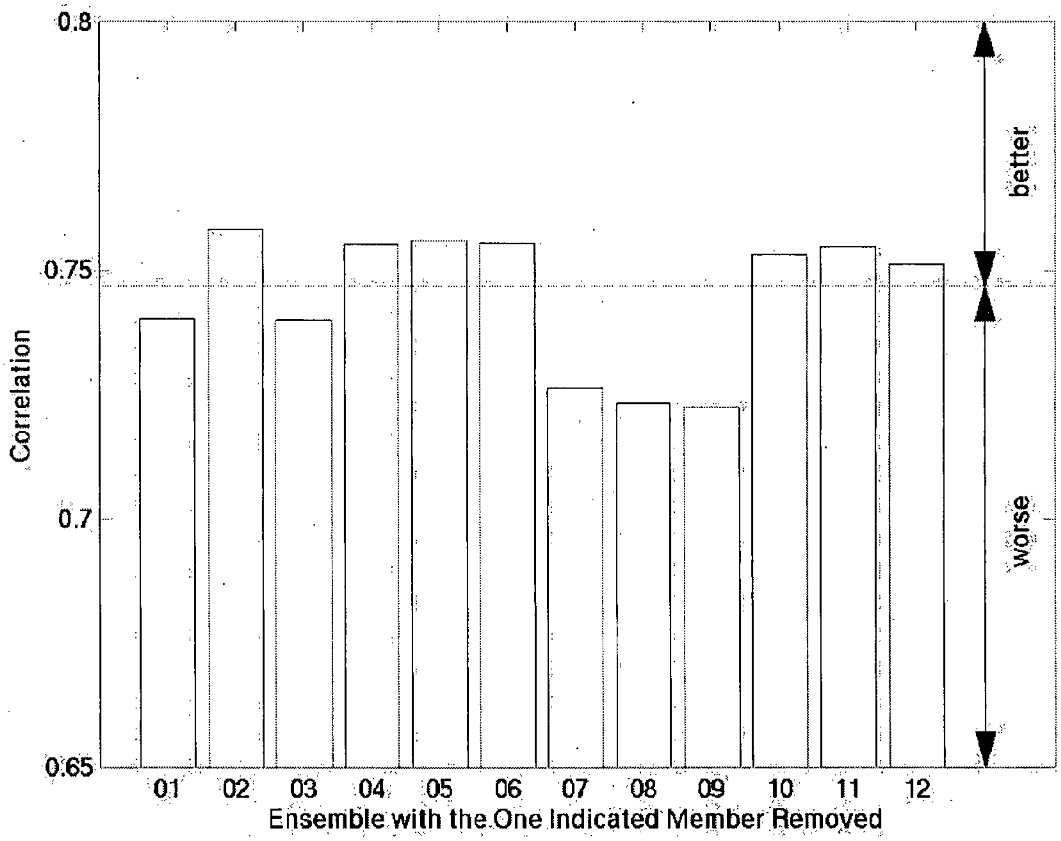

Figure 3.11: Median (over the five stations) of the correlation of the 11-member ensemble-mean, given for the 5-day period 11-15 August 2004. Each bar represents the correlation value for the ensemble-mean without the corresponding ensemble member (the label below the bar). The dashed line is the correlation value for the full 12-member ensemble and the better-worse designation at right is relative to this full ensemble. Values are within the interval [−1, 1], with correlation = 1 being the best possible value.

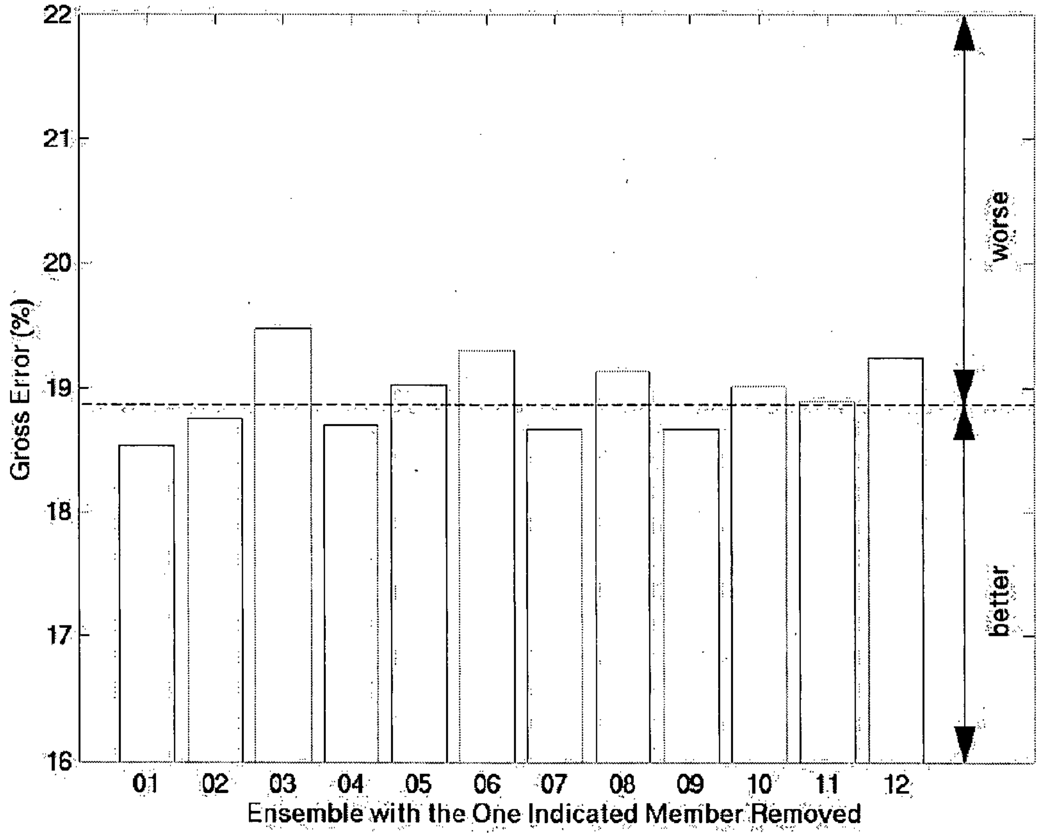

Figure 3.12: Similar to Figure 3.11, but for the gross error. Values are within the interval [0, $+\infty$], with perfect forecast when gross error = 0.

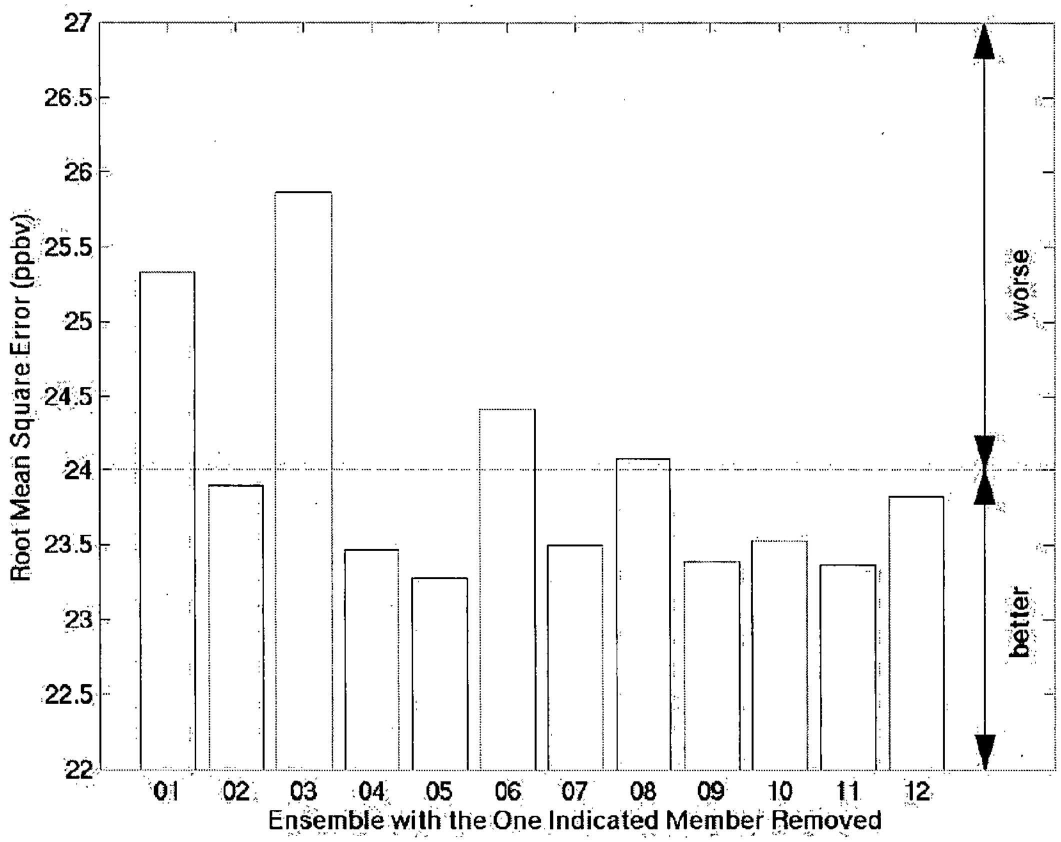

Figure 3.13: Similar to Figure 3.11, but for the root mean square error (RMSE). Values are within the interval [0, + ∞], with a perfect forecast when RMSE = 0.

ensemble) are forecasts 01, 03, 06, and barely 08, while removing the others from the ensemble results in a better RMSE ensemble-mean.

UPPA results are shown in Figure 3.14. The values are between 19.5 and 22.5 %, meaning that none of the models change dramatically this statistical parameter when excluded from the ensemble. Notably, when the 4-km runs (for both MM5 and MC2) with the CTRL and NOXP emission run (forecasts 07, 08, 10, and 11) are removed separately from the ensemble, the UPPA gets worse. The only other forecast that makes UPPA better (i.e., UPPA is worse if removed) is forecast 04 (MM5, 12-km, CTRL run). All the other forecasts make this statistical

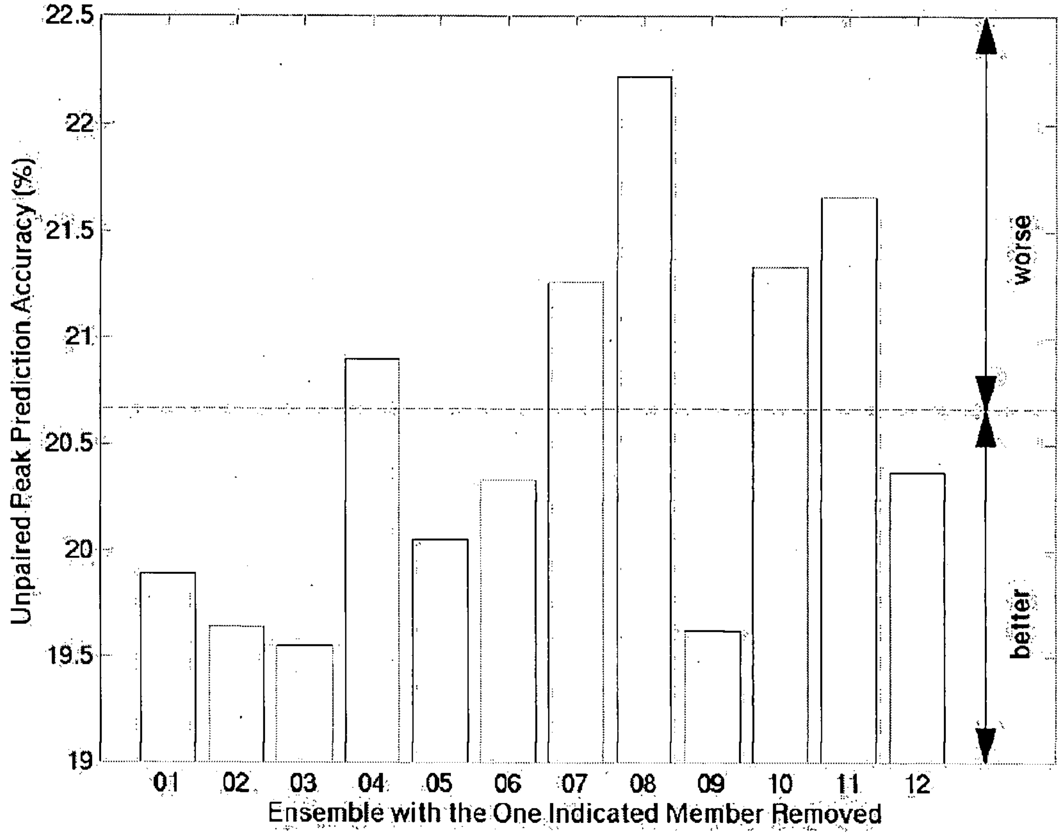

Figure 3.14: Similar to Figure 3.11, but for the unpaired peak prediction accuracy (UPPA). Values are within the interval [0, + ∞], with a perfect peak forecast when UPPA = 0.

parameter worse when they are retained, when they contribute to the ensemble.

3.4.4 18-member OEFS Results

Hoffman and Kalnay (1983) introduced the lagged-average weather forecast. The forecasts initialized at the current initial time, t = 0, as well as forecast from the previous times, $t = -\tau, -2\tau, \cdots, (N-1)\tau$ are combined at a common valid time to form an ensemble. They tested this approach using a primitive-equation NWP model to represent the true atmospheric evolution, and a quasi-geostrophic NWP model as the forecast. They found the lagged-average

Table 3.5: Correlation, gross error (%), root mean square error (RMSE) (ppbv), and unpaired peak prediction accuracy (UPPA) (%) for a 12-member (12-ens) and an 18-member (18-ens) Ozone Ensemble Forecast System, are listed at five stations [Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope], for the 5-day period 11-15 August 2004.

	Correlation		**Gross Error (%)**		**RMSE (ppbv)**		**UPPA (%)**	
	12-ens	**18-ens**	**12-ens**	**18-ens**	**12-ens**	**18-ens**	**12-ens**	**18-ens**
CYVR	0.74	0.72	44	37	24	23	39	35
Langley	0.84	0.85	15	15	17	17	13	13
Abbotsford	0.91	0.90	12	11	19	19	11	13
chilliwack	0.71	0.72	18	19	25	26	20	21
Hope	0.23	0.06	24	25	28	29	29	31

forecast to be slightly better than a Monte Carlo forecast (introduced assuming a perfect model by Leith (1974)), and they found higher correlation between error growth and ensemble spread in their approach. These improvements were because the lagged-average forecast perturbations are not randomly chosen, but better capture the error of the day. In the literature some other applications of this ensemble approach can be found, as for example in Dalcher et al. (1988).

In our study, we tested a lagged-averaged ozone ensemble. Each of the six 12-km resolution ensemble members is run for more than 48 hours. This allows the expansion of the 12-member OEFS to an 18-member OEFS, by adding the second half of the six 12-km "yesterday" forecasts to the "today" ensemble forecast, as shown in Figure 3.5.

Table 3.5 shows the results of the 12-member and 18-member OEFS, for the same statistical parameters as in the previous subsections, and for the same 5-day period and the same stations. Only in few occasions is the 18-member OEFS slightly better than the 12-member one, as for example for the gross error and UPPA at CYVR. In general the two ensemble systems have very similar forecast skill, meaning that the computation effort of adding the six lagged members to the original system does not provide valuable results.

Ideally, each ensemble member should give an equally likely time evolution and space distribution of ozone concentration, and they should all give equally good estimates of truth. The ensemble members should thus be "independent", in the sense that none of them should rely on other members for their realizations. This is not the case when nested grids are used, as for 12-member OEFS presented in this study. Namely, CMAQ domains are linked using a 1-way nesting approach (similarly for MC2, but MM5 runs are implemented with 2-way nesting), all the 4 km runs cannot be considered independent of the runs where the driving meteorology is their 12 km coarser domain. Moreover, the fact that the addition of six lagged members leave the OEFS performances substantially unvaried, suggests that no independent information on errors is added with those members.

3.5 Discussion

3.5.1 Taylor Diagrams

A concise way to display and study the results is to use a Taylor diagram (Taylor, 2001). It can be used to create a multi-statistic plot of correlation, centered RMSE (CRMSE: RMSE computed after the average is removed from the time series), and standard deviation. This is done for each forecast, for the ensemble-mean, and for the observations. CRMSE is the distance on the diagram between the point representing the forecast and the one representing the observations.

At the Vancouver International Airport (Figure 3.15), the ensemble has the best performance, as indicated by being closest to the observations. Forecasts 07, 08, and 09 (MC2, 4-km) are the worst, being the farthest. At Langley (Figure 3.16) the ensemble-mean is the closest,

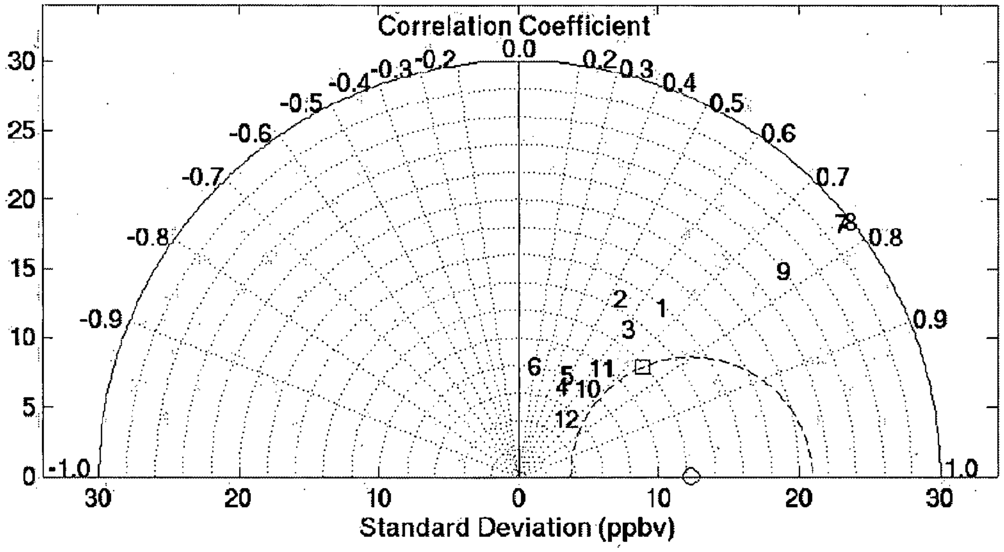

Figure 3.15: Taylors diagram is plotted for Vancouver International Airport (CYVR). The azimuthal position gives the correlation, while the radial distance from the origin is proportional to the standard deviation (ppbv). The circle represents the observations, and the square is the ensemble-mean. The numbers correspond to the ensemble-member indices. The distance between the observation and a given point is proportional to the centered root mean square error (CRMSE) between the observations and the forecast having the correlation and standard deviation of the given point. The dashed line indicates the ensemble-mean CRMSE centered over the point representing the observation.

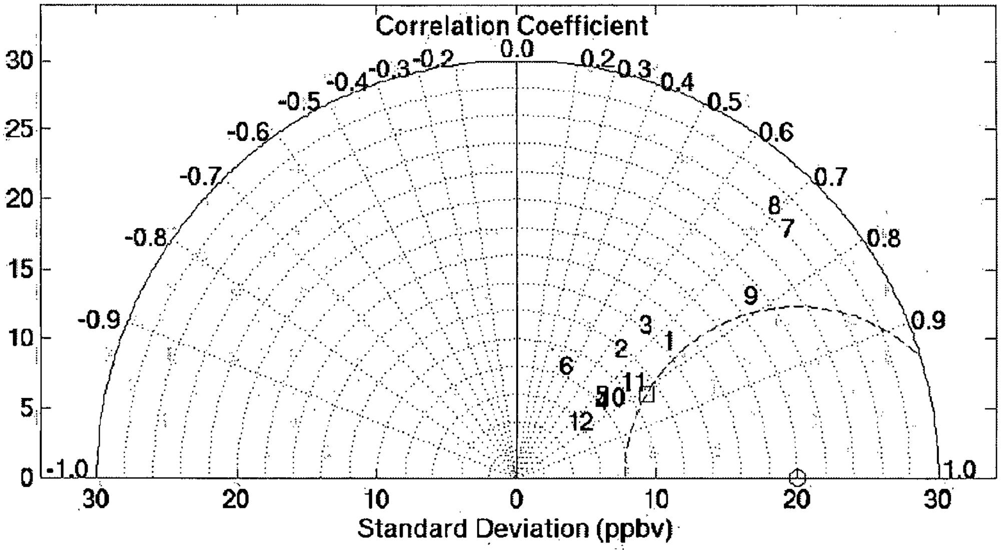

Figure 3.16: Taylor diagram for Langley (similar to Figure 3.15).

while forecasts 07 and 08 are the worst, and 09 has an average performance. At Abbotsford (Figure 3.17) 07 is the best, with 09 and the ensemble-mean having similar distance from the observations and being the second closest. At Chilliwack (Figure 3.18) the ensemble-mean and 09 have again the same distance from the observations, and 08 and 07 are closest and the second closest, respectively. Finally at Hope (Figure 3.19) forecasts 07, 08, and 09 are all closer to the observations than the ensemble-mean.

The ensemble-mean forecast is not the best at every location and for any given observed ozone concentration. However, overall it is indeed the most skillful forecast when tested against

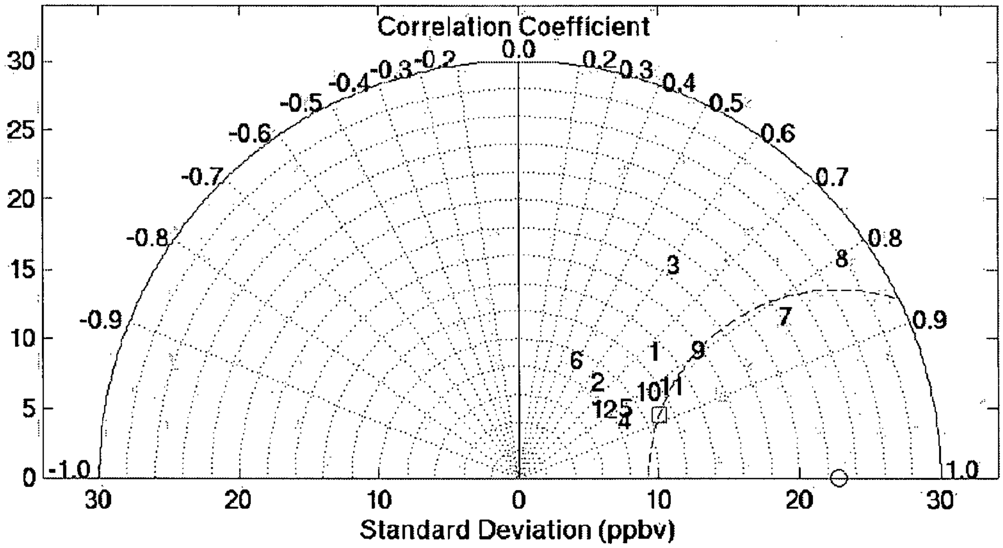

Figure 3.17: Taylor diagram for Abbotsford (similar to Figure 3.15).

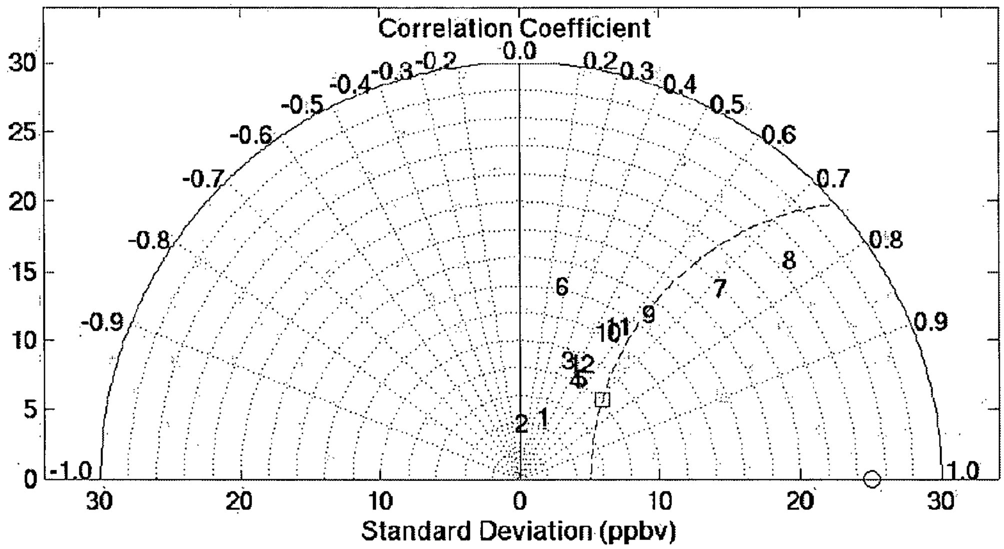

Figure 3.18: Taylor diagram for Chilliwack (similar to Figure 3.15).

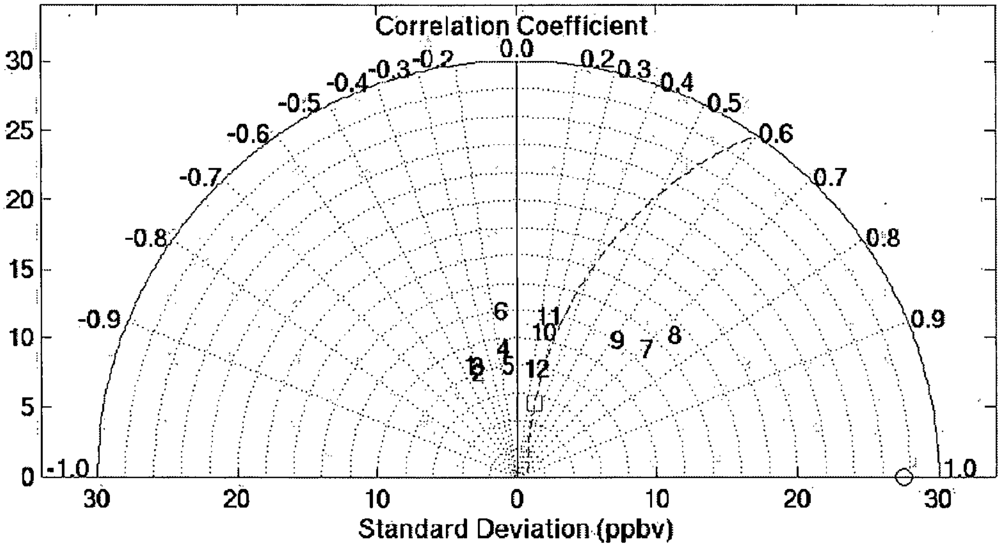

Figure 3.19: Taylor diagram for Hope (similar to Figure 3.15).

observations, and compared to any other individual ensemble member. The key point in favor of the ensemble-mean is that it is not possible to establish a priori which specific ensemble member will outperform the ensemble-mean in any specific situation.

3.5.2 Meteorology versus Emission Perturbations

Ensemble members 01, 04, 07 and 10 (MC2 and MM5 control runs at 12 km, and MC2 and MM5 control runs at 4 km) are the control runs, where the non-perturbed emission data are used. Namely, only the meteorology is perturbed. Any one of those control runs can be compared with runs driven by the same meteorological field but with an emission perturbation (plus or minus 50 % NO_x). This means comparing ensemble member 01 with 02 and 03, 04 with 05 and 06, 07 with 08 and 09, and 10 with 11 and 12. This methodology allows one to infer information about the contribution to the ensemble performance of meteorology versus emission perturbations.

The control runs have good correlation statistics relative to the runs driven by the same meteorology but with emission perturbations. This could reflect the importance of meteorology perturbations in capturing the ozone temporal and spatial distributions. However, by looking at RMSE, the emission-perturbation runs seem to produce better (i.e., lower) RMSE values overall when compared with the control runs. Thus, emission perturbations are needed to better predict ozone-concentration magnitude.

The analysis above suggests that both perturbations are needed to have a skillful forecast. This is another reason why the ensemble average is the best. However, further investigations using other case studies could help to test this hypothesis.

3.5.3 Spread versus Skill

The standard deviation of the ensemble members about the ensemble mean is called spread. The relationship between ensemble spread and forecast error is not yet well defined (Kalnay, 2003). Nevertheless, it often provides very useful information about ensemble skill. Ensemble weather forecasts often provide information on the reliability of the forecast: if the ensemble members have large spread, this implies less confidence in the forecast.

In this study no correlation or relationship between ozone ensemble spread and forecast error has been found. This is caused by a lack of accuracy of one or more aspects of the modeling process, which creates similar errors in the forecasts for specific circumstances. For instance overnight most of the forecasts are close to each other resulting in a small spread, as shown in Figure 3.20 at Langley, for the 5-day period 11-15 August 2004 (shaded areas represent nighttime periods). At the same time those forecasts are far from the observations, and this results in an ensemble where there is small spread with large errors. In this case, the correlation that the ensemble skill and spread may have in other parts of the day is partially mitigated by what occurs in those specific circumstances.

3.6 Summary and Conclusions

A new Ozone Ensemble Forecast System (OEFS) has been tested. Twelve ensemble members are obtained by driving U.S. Environmental Protection Agency (EPA) Models-3/Community Multiscale Air Quality Model (CMAQ) with two mesoscale models, the Mesoscale Compressible Community (MC2) model and the Penn State/NCAR mesoscale (MM5) model, each run at two resolutions, 12 and 4-km. CMAQ is run for three emission scenarios for each of the

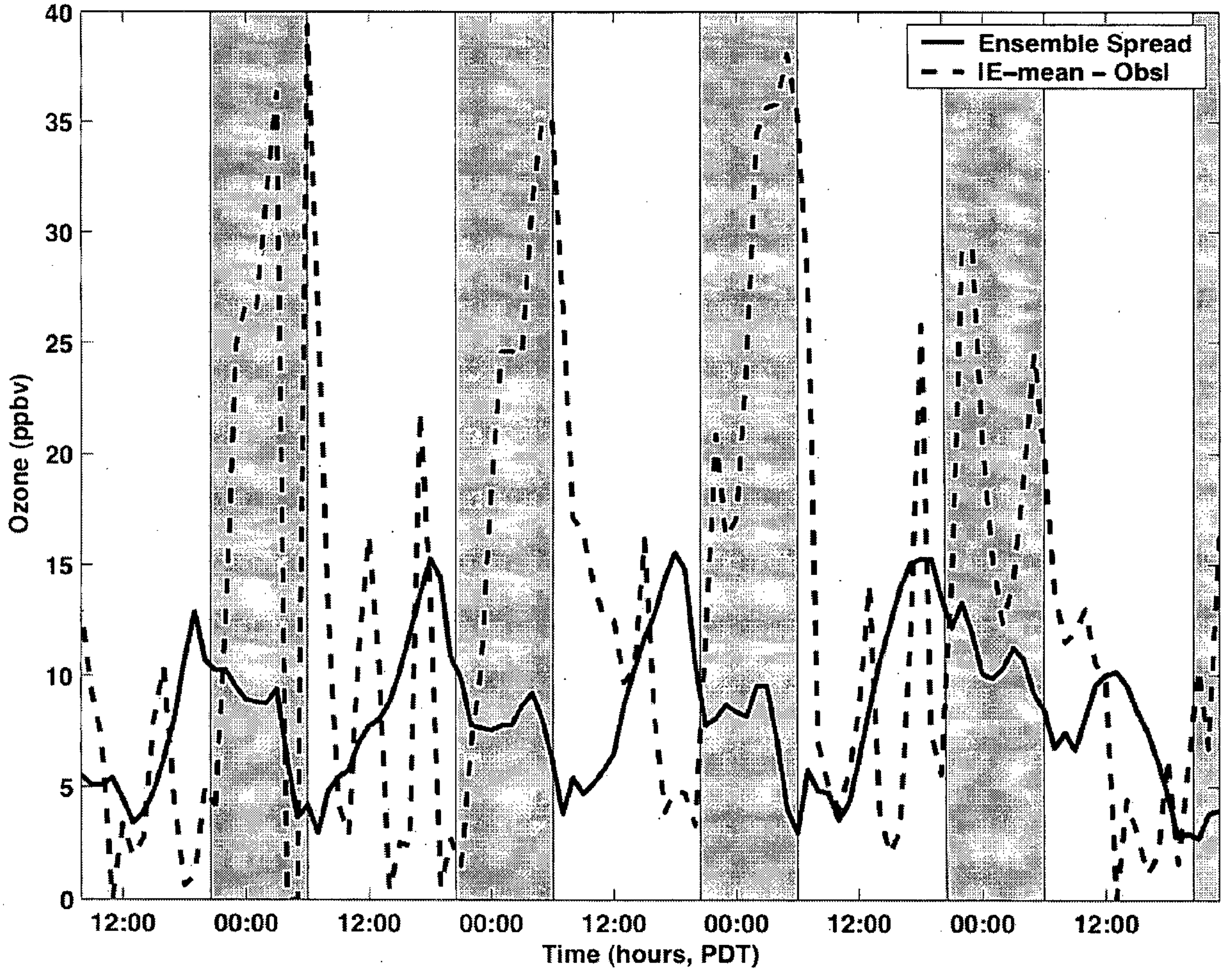

Figure 3.20: Spread (standard deviation of the ensemble members about the ensemble mean (E-mean)) and E-mean absolute error (absolute values of the difference between E-mean and observations (Obs)) at Langley, for the 5-day period 11-15 August 2004. Shaded areas represent nighttime periods. Local Pacific Daylight Time (PDT) is UTC - 7 h.

four available meteorological fields: a control run, 50 % more NO_x, and 50 % less NO_x.

The performance of the ensemble-mean and 12 different forecasts is compared with the individual forecasts and tested against observations for a 5-day period (11-15 August, 2004), over five monitoring stations in the Lower Fraser Valley (LFV), British Columbia (BC). In summary, for the locations and days used to test this new OEFS, one finds strong evidence for the following:

- The ensemble-mean is usually the best AQ forecast if ranked using correlation, gross error, or RMSE.

- The ensemble-mean has an average performance with UPPA. One possible reason could be that ensemble averaging could cause excessive smoothing of the peak values.

- The ensemble-mean forecast is not the best at every location and for any given observed ozone concentration. However, it is indeed the most skillful forecast when tested against observations, and compared to any other ensemble member, since it is able to remove part of the unpredictable components of the individual deterministic forecasts.

- The ranking sum is useful for comparing overall performance.

- Sporadically (in space and time) there are few ensemble members that have better performance than the ensemble-mean when the forecasts are ranked based on a particular statistical parameter. The key point in favor of the ensemble-mean is that it is not possible to establish a priori which specific ensemble member will outperform the ensemble-mean in any specific situation.

- Meteorology perturbations could be important to better capture the ozone temporal and spatial distributions, while emission perturbations could be necessary to better predict the ozone-concentration magnitude. If this is the case, then both perturbations are useful for maximizing the skill of ozone forecasts, but further investigations are needed to validate this hypothesis.

- The 11-member ensembles, given by removing each of the 12-members in turn from the original ensemble, show results close to the 12-member system for correlation, gross error, RMSE and UPPA. In general, no particular 11-member ensemble consistently outperforms the other possible 11-member combinations. This reflects the fact that there is not one of the 12 forecasts that clearly outperform the others, based on the four statistical parameters considered here.

- The 18-member ensemble did not improve the ensemble-mean forecast skill. This is probably because the added six lagged forecasts did not span more uncertainty than the original 12-member ensemble, and that no independent information on errors is added with those members.

These results indicate that ensemble averaging improves the forecast timing of maximum and minimum concentrations with respect to the observations, because the correlation is closer to one. From the improved (decreased) RMSE and gross-error values, we infer that ensemble averaging does improve the forecast accuracy by reproducing the magnitude of ozone concentrations. The ensemble-mean average performance with UPPA could be caused by excessive smoothing of the peak values.

The results presented in this study suggest that an air-quality (AQ) ensemble design built on meteorological and emission-field perturbations is a promising approach. For NWP ensembles, the multi-model approach is the more promising approach, especially for short-range forecasts (Hou et al., 2001; Wandishin et al., 2001). So, even if only two different NWP models are used (each with two different resolutions), the results found here indicate that the multi-model approach is an efficient way to perturb the meteorological input in an AQ ensemble design as well.

Furthermore, the emission errors are expected to behave in a more systematic fashion than the errors in the initial conditions. They should depend much less on temporal variations of the atmosphere. So the issue of capturing the "error of the day", which each NWP ensemble system strives for (Kalnay, 2003, and references therein), should be less pronounced for emission perturbations within an AQ ensemble design. This could be a reason why the simple emission perturbation tested here (combined with the multi-NWP model perturbation) gives good results. Further investigation is needed to clarify this point.

A refinement of the system could focus on the emission perturbations. Ideally, a multi-model approach, using the Sparse Matrix Operator Kernel Emission (SMOKE) model and other state-of-the-art emission pre-processors, would take into account many of the uncertainties generated by the several approximations embedded in the emission-data gathering and computation processes. An alternative way could be to run the same emission pre-processor (e.g., SMOKE) with different configurations, and starting from different emission inventories to generate different (but equally likely) emission fields.

Future work could focus also on a VOC-based perturbation OEFS, and the comparison with this study should help to understand the effects of different emission perturbations (NO_x or VOC) when combined with meteorology perturbations. Moreover, interesting experiments could result from generating ensemble members by also perturbing other phases of the AQ modeling process, such as the chemistry. For instance, Hanna et al. (2001) found the NO_2 photolysis rate to be "the variable whose uncertainties are most strongly correlated to the uncertainties in predictions of maximum hourly averaged ozone concentrations". This would make it a strong candidate as a parameter to be perturbed. Perturbing the chemistry likely would be more important in predicting particulate matter rather than ozone, because of the higher uncertainties on how the models represent hetereogeneous chemistry when compared to gas-phase chemistry.

Also, the perturbations of the meteorological field presented here are not spatially independent, because two NWP models are used to produce forecasts over four domains. A likely improvement could be obtained by using different NWP models for each domain.

Finally, ensemble averaging is able to remove part of the unpredictable components of the physical and chemical processes involved in the ozone fate, resulting in a more skillful forecast when compared to any deterministic ensemble member. In Chapter 5, it is shown how a Kalman filter can be used to reduce systematic errors. Thus, using both ensemble averaging and Kalman filtering, significantly improved real-time AQ forecasts are possible even in complex coastal mountain setting as in the LFV. There are no intrinsic limitations to these methods that would prevent their application in real time to other pollutants in other geographic settings.

3.7 References for Chapter 3

Ainslie, B., 2004: *A photochemical model based on a scaling analysis of ozone photochemistry.* Ph.D. thesis, 311 pp., University of British Columbia, Vancouver, Canada.

Benoit, R., M. Desgagne, P. Pellerin, S. Pellerin, Y. Chartier, and S. Desjardins, 1997: The Canadian MC2: A semi-Lagrangian, semi-implicit wide band atmospheric model suited for fine scale process studies and simulation. *Monthly Weather Review*, **125**, 2382–2415.

Brauer, M. and J. R. Brook, 1995: Personal and fixed-site ozone measurements with a passive sampler. *Journal of the Air and Waste Management Association*, **45**, 529–537.

Brown, R. P., T. Butler, and S. W. Hawley, 2001: *Ageing of Rubber - Accelerated Weathering and Ozone Test Results.* Rapra, Shawbury, United Kingdom.

Byun, D. W. and J. K. S. Ching, 1991: Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) modeling system. Technical Report EPA/600/R-99/030, U.S. Environmental Protection Agency.

Carmichael, G. R., Y. S. Chang, J. S. Scire, and R. J. Yamartino, 1992: The CALGRID mesoscale photochemical grid model - I. Model formulation. *Atmospheric Environment*, **26**, 1493–1512.

Coats, C. J. J., 1996: High-performance algorithms in the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system. In *9th AMS Joint Conference on Applications of*

Air Pollution Meteorology with Air and Waste Management Association. American Meteorological Society, Atlanta, Georgia.

Dabberdt, W. F., M. A. Carroll, D. Baumgardner, G. Carmichael, R. Cohen, T. Dye, J. Ellis, G. Grell, S. Grimmond, S. Hanna, J. Irwin, B. Lamb, S. Madronich, J. McQueen, J. Meagher, T. Odman, J. Pleim, H. P. Schmid, and D. Westphal, 2003: Meteorological research needs for improved air quality forecasting: report of the 11^{th} Prospectus Development Team of the U.S. Weather Research Program. Technical report, National Center for Atmospheric Research.

Dabberdt, W. F. and E. Miller, 2000: Uncertainty, ensembles and air quality dispersion modeling: applications and challenges. *Atmospheric Environment*, **34**, 4667–4673.

Dalcher, A., E. Kalnay, and R. N. Hoffmann, 1988: Medium range lagged average forecasts. *Monthly Weather Review*, **116**, 402–416.

Delle Monache, L., X. Deng, Y. Zhou, H. Modzelewski, G. Hicks, T. Cannon, R. B. Stull, and C. di Cenzo, 2004: Air quality ensemble forecast over the Lower Fraser Valley, British Columbia, Canada. In *27th NATO/CCMS Conference on Air Pollution Modeling*. Banff, Alberta.

EPA, 1991: Guideline for regulatory application of the urban airshed model. Technical Report EPA-450/4-91-013, U.S. Environmental Protection Agency.

Galmarini, S., R. Bianconi, W. Klug, T. Mikkelsen, R. Addis, S. Andronopoulos, P. Astrup, A. Bklanov, J. Bartniki, J. C. Bartzis, R. Bellasio, F. Bompay, R. Buckley, M. Bouzom, H. Champion, R. D'Amoursn, E. Davakis, H. Eleveld, G. T. Geertsema, H. Glaab, M. Kollax,

M. Ilvonen, A. Manning, U. Pechinger, C. Persson, E. Polreich, S. Potemski, M. Prodanova, J. Saltbones, H. Slaper, M. A. Sofiev, D. Syrakov, J. H. Sørensen, L. Van der Auwera, I. Valkama, and R. Zelazny, 2004b: Ensemble dispersion forecasting-Part II: application and evaluations. *Atmospheric Environment*, **38**, 4619–4632.

—, 2004a: Ensemble dispersion forecasting-Part I: concept, approach and indicators. *Atmospheric Environment*, **38**, 4607–4617.

Gery, M. W., G. Z. Whitten, J. P. Killus, and M. C. Dodge, 1989: A photochemical kinetics mechanism for urban and regional scale computer modeling. *Journal of Geophysical Research*, **94**, 12,925–12,956.

Grell, G. A., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation PENN State/NCAR Mesoscale Model (MM5). Technical Report NCAR/TN-398 +STR, National Center for Amospheric Research.

GVRD, 2002: 2000 emissions inventory for the Lower Fraser Valley airshed. Technical report, Greater Vancouver Regional District.

Hanna, S. R., L. Zhigang, H. C. Frey, N. Wheeler, J. Vukovich, S. Arunachalam, M. Fernau, and D. Hansen, 2001: Uncertainties in predicted ozone concentrations due to input uncertainties for the UAM-V photochemical grid model applied to the July 1995 OTAG domain. *Atmospheric Environment*, **35**, 100–118.

Hoffman, R. N. and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus*, **35**, 100–118.

Horvath, S. M. and D. J. McKee, 1994: Acute and chronic health effects of Ozone. In *Tropospheric ozone, human health and agricultural aspects*. Lewis Publisher, Boca Raton, Florida, pages 39–84.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX'98 ensemble forecasts. *Monthly Weather Review*, **129**, 73–91.

Huang, H.-C. and J. S. Chang, 2001: On the performance of numerical solvers for a chemistry submodel in three-dimensional air quality models. *Journal of Geophysical Research*, **106**, 20,175–20,188.

Jacobson, M. Z., 1999: *Fundamentals of Atmospheric Modeling*. Cambridge University Press, New York.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, New York.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, **102**, 409–418.

Lorenz, E. N., 1963: Deterministic non-periodic flow. *Journal of the Atmospheric Sciences*, **20**(130-141).

Martilli, A. and D. G. Steyn, 2004: A numerical study of recirculation processes in the Lower Fraser Valley (British Columbia, Canada). In *27th NATO/CCMS Conference on Air Pollution Modeling*. Banff, Alberta.

McHenry, J. N., W. F. Ryan, N. L. Seaman, C. J. Coats Jr., J. Pudykiewicz, S. Arunachalam, and J. M. Vukovich, 2004: A real-time Eulerian photochemical model forecast system. *Bulletin of the American Meteorological Society*, **85**, 525–548.

McKeen, S. A., J. M. Wilczak, G. A. Grell, I. Djalalova, S. Peckham, E.-Y. Hsie, W. Gong, V. Bouchet, S. Menard, R. Moffet, J. McHenry, J. McQueen, Y. Tang, G. R. Carmichael, M. Pagowski, A. Chan, T. Dye, G. Frost, P. Lee, and R. Mathur, 2005: Assessment of an ensemble of seven real-time ozone forecasts over Eastern North America during the summer of 2004. *To appear on Journal of Geophisical Research.*

McKendry, I. G., 1994: Synoptic circulation and summertime ground-level ozone concentrations at Vancouver, British Columbia. *Journal of Applied Meteorology*, **33**, 627–641.

McKendry, I. G. and J. Ludgren, 2000: Tropospheric layering of ozone in regions of urbanized complex and/or coastal terrain: a review. *Progress in Physical Geography*, **24**, 329–354.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The new ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of Royal Meteorological Society*, **122**, 73–119.

Nodop, K., R. Connolly, and G. Girardi, 1998: The field campaigns of the European Tracer Experiment (ETEX): overview and results. *Atmospheric Environment*, **32**, 4095–4108.

O'Neill, S. M. and B. K. Lamb, 2005: Intercomparison of the Community Multiscale Air Quality Model and CALGRID using process analysis. *Evironmental Science and Technology*, **39**, 5742–5753.

Runeckles, V., 2002: Effects on vegetation and ecosystems. In Suzuki, D., editor, *A Citizen's guide to air pollution*. Vancouver, British Columbia, pages 177–216.

Russell, A. and R. Dennis, 2000: NARSTO critical review of photochemical models and modeling. *Atmospheric Environment*, **34**, 2283–2324.

Salmond, J. A. and I. G. McKendry, 2002: Secondary ozone maxima in a very stable nocturnal boundary layer: observations from the Lower Fraser Valley, BC. *Atmospheric Environment*, **36**, 5771–5782.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 1998: Ensemble forecasting of mesoscale convective systems. In *12th Conference on Numerical Weather Prediction*. American Meteorological Society, Phoenix, Arizona, pages 265–268.

Steyn, D. G., J. W. Bottenheim, and R. B. Thomson, 1997: Overview of tropospheric ozone in the Lower Fraser Valley, and the Pacific '93 field study. *Atmospheric Environment*, **31**, 2025–2035.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, **106**, 7183–7192.

Thomas, S. J., J. P. Hacker, M. Desgagné, and R. B. Stull, 2002: An ensemble analysis of forecast errors related to floating point performance. *Weather and Forecasting*, **17**, 898–906.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bulletin of the American Meteorological Society*, **74**, 2317–2330.

Vaughan, J., B. Lamb, C. Frei, R. Wilson, C. Bowman, C. Figueroa-Kaminsky, S. Otterson, M. Boyer, C. Mass, M. Albright, J. Koenig, A. Collingwood, M. Gilroy, and N. Maykut, 2004: A numerical daily air quality forecast system for the Pacific Northwest. *Bulletin of the American Meteorological Society*, **85**, 549–561.

Vingarzan, R., 2004: A review of surface ozone background levels and trends. *Atmospheric Environment*, **38**, 3431–3442.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multi-model ensemble system. *Monthly Weather Review*, **129**, 729–747.

Willmott, C. J., 1981: On the validation of models. *Physical Geography*, **2**, 184–194.

Chapter 4

Probabilistic and Ensemble-averaged Regional Ozone Forecasts

4.1 Introduction

[1] Exposure to ozone in the troposphere may have adverse effects on humans (Horvath and McKee, 1994; Brauer and Brook, 1995), vegetation (Runeckles, 2002) and materials (Brown et al., 2001). To alert the population about impending air-quality (AQ) degradation, Dabberdt and Miller (2000) discussed the need for an operational AQ forecast system. Experiences with such numerical forecast systems are described in Delle Monache et al. (2004), McHenry et al. (2004) and Vaughan et al. (2004). The U.S. Weather Research Program and its Prospec-

[1] A version of this chapter will be submitted for publication. Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2005: Probabilistic and ensemble-averaged regional ozone forecasts, *manuscript to be submitted in October 2005 to the Journal of Geophysical Research.*

tus Development Team on Air-Quality Forecasting (Dabberdt et al., 2003) recommended a probabilistic approach to AQ forecasting due to the chaotic nature of the atmosphere.

It has been found for numerical weather prediction (NWP) than the ensemble-mean is more accurate that an individual model realization (e.g., Toth and Kalnay, 1993; Molteni et al., 1996). Chapters 2 and 3 and recent studies (e.g., McKeen et al., 2005) have shown that the ensemble average yields similar benefits for AQ prediction, because there are similar model complexities and constraints. Moreover, NWP ensembles have been very useful by providing information about the likelihood of possible future evolution of the atmosphere. Similarly, AQ ensembles may be able to provide reliable probabilistic information about possible AQ scenarios. Given the nonlinear nature of photochemical reactions, the differences among ensemble members of an Ozone Ensemble Forecast System (OEFS) may be able to account for some of the uncertainties associated with each component of the modeling process.

Chapter 2 discussed the benefit of the AQ ensemble approach in studies involving not only pollutant transport, but also the associated photochemical reactions. An ensemble composed of four Chemistry Transport Models (CTMs) was tested for a 6-day summer period over five monitoring stations in northwestern and central Europe. The ensemble mean presented in that study showed promising results, performing better than the models individually, and giving good performance for ozone peak-value prediction.

Another successful implementation of the ensemble approach for ozone forecasts can be found in McKeen et al. (2005), where the authors present results for a multi-model (i.e., seven CTMs) OEFS, statistically evaluated for 53 days (summer 2004), against 340 monitoring stations over eastern U.S. and southern Canada. The high correlation coefficients and low root-mean-square-error (RMSE) points to the ensemble mean as the preferred forecast when compared to any individual model.

Chapter 3 introduced a new OEFS design (12 ensemble members), generated by including both meteorology and emission (NO_x) perturbations. They tested the ensemble mean for a 5-day episode (August 2004) over the Lower Fraser Valley (LFV), British Columbia, Canada, and found that the ensemble average is the best forecast, having the best timing of maxima and minima values, and predicting the ozone magnitude more accurately than any other individual forecast.

These successful experiments prompted the work presented here. Studies of ozone photochemistry in the LFV (Ainslie, 2004) show that the present and projected AQ is in a regime affected roughly equally by NO_x and VOC emissions (Figure 4.1). Namely, in a maximum ozone concentration plot as a function of NO_x and VOC emissions, the state of the LFV is above the ridgeline of ozone relative maxima. In Chapter 3 the emission perturbations are generated with 50 % more NO_x emissions (point A in Figure 4.1), and 50 % less (point B in Figure 4.1). In this Chapter, VOC perturbations are also considered, and the 12-member ensemble has been expanded to 28 members. Hanna et al. (2001) reported that both NO_x and VOC estimates can be in error by a factor of two or more.

The different forecasts are grouped in 13 different OEFS protocols. One includes all the forecasts, one includes only the meteorology perturbations, four have only emission perturbations, three have both meteorology and emissions perturbations, one contains only fine-resolution runs, one has only coarse-resolution forecasts, and two drive the AQ forecast with two different Numerical Weather Prediction (NWP) models.

The performance of these OEFS groups are investigated here by comparing their forecast skill as both probabilistic and ensemble-averaged forecasts. The effects of different perturba-

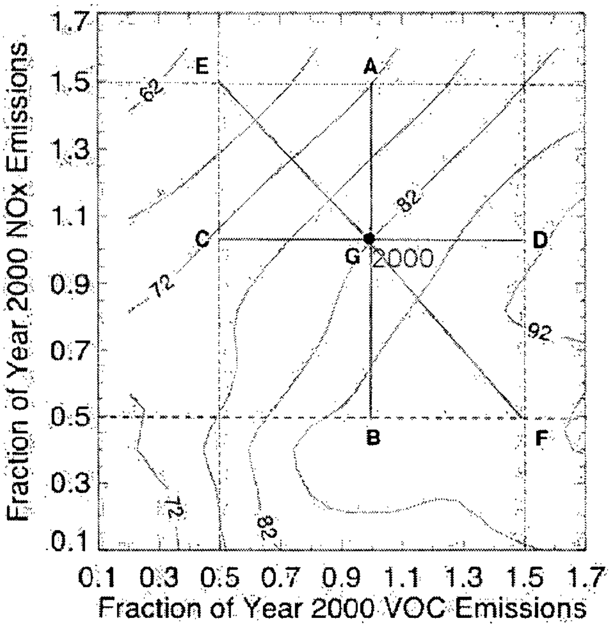

Figure 4.1: Isopleths of maximum ozone concentration (ppbv) are given as a function of year 2000 VOC and NO_x emissions over the Lower Fraser Valley (adapted from Ainslie (2004)). The total annual VOC and NO_x emissions are 111, 196 and 99, 897 metric tonnes, respectively (GVRD, 2002). The vertical bar is along the plus (point A) and minus (point B) 50% NO_x perturbations. The horizontal bar shows the plus (point D) and minus (point C) 50% VOC perturbations. The diagonal bar (approximately perpendicular to the isopleths) follows the plus 50% NO_x and minus 50% VOC perturbation (point E) and the minus 50% NO_x and plus 50% VOC perturbation (point F). Point G is the control run with no perturbations.

tions, resolutions, and driving models on the ensemble skill are analyzed.

Section 4.2 describes in detail the OEFS groups generated in this study. Section 4.3 and 4.4 present the probabilistic forecast skill metrics and results, respectively. This is followed by an analysis of the results of the ensemble-averaged forecasts (Section 4.5). In Section 4.6 conclusions are drawn.

4.2 Ozone-Ensemble Methodology

Following the work in Chapter 3, both the meteorology and emissions are perturbed in this new study. Two NWP mesoscale models are each run with two horizontal grid spacings: 12 and 4 km, yielding four meteorological fields. The mesoscale models are the Mesoscale Compressible Community (MC2) NWP model (Benoit et al., 1997) and the Penn State/NCAR mesoscale (MM5) model (Grell et al., 1994), which have been running daily for a decade at the University of British Columbia (UBC), [http://weather.eos.ubc.ca/wxfcst/]. The AQ forecasts were produced with the U.S. Environmental Protection Agency (EPA) Models-3/Community Multiscale Air Quality Model (CMAQ) Chemistry Transport Model (CTM) (Byun and Ching, 1991).

In this new study, both VOC and NO_x perturbations are considered. For each one of the four available meteorological input fields, runs are made with plus and minus 50 % VOCs (point D and C in Figure 4.1, respectively). Also, the NO_x and VOC perturbations have been combined, to make perturbations that are perpendicular to the ozone maximum isopleths. This better captures more of the ozone uncertainty than when perturbing only NO_x or VOC. Hence, perturbations combining plus 50 % NO_x and minus 50 % VOC (point E, Figure 4.1), and minus 50 % NO_x and plus 50 % VOC (point F, Figure 4.1) were generated as well. Ensemble members with the original points A and B from Chapter 3 are also included to allow comparison with NO_x-only perturbations. Including the control run with no emissions perturbations, there are a total of seven emission fields, corresponding to the seven points in Figure 4.1.

The 28 AQ forecasts resulting from the above perturbation combinations (four meteorology times seven emission) are tested here using the same episode analyzed in Chapter 3, with hourly observed ozone concentrations from five stations across the Lower Fraser valley (LFV), British Columbia (BC), Canada: Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope (Figure 4.2). The study period is 11-15 August 2004, and further details about the data and episode can be found in Section 3.2.

The 28 ensemble members are grouped into the following subsets, to form 13 different ensemble groups, as also summarized in Table 4.1. These are identified with abbreviation as follows:

- All the forecasts available (ALL, 28 members).
- Meteorology and NO_x perturbations combined together, as presented in Chapter 3 (MET+NO_x, 12 members).
- Meteorology and VOC perturbations (MET+VOC, 12 members).
- Meteorology and NO_x combined with VOC perturbations (MET+NO_xVOC, 12 members).
- All the ensemble members driven by MC2 at 12 km (MC2-12, seven members).

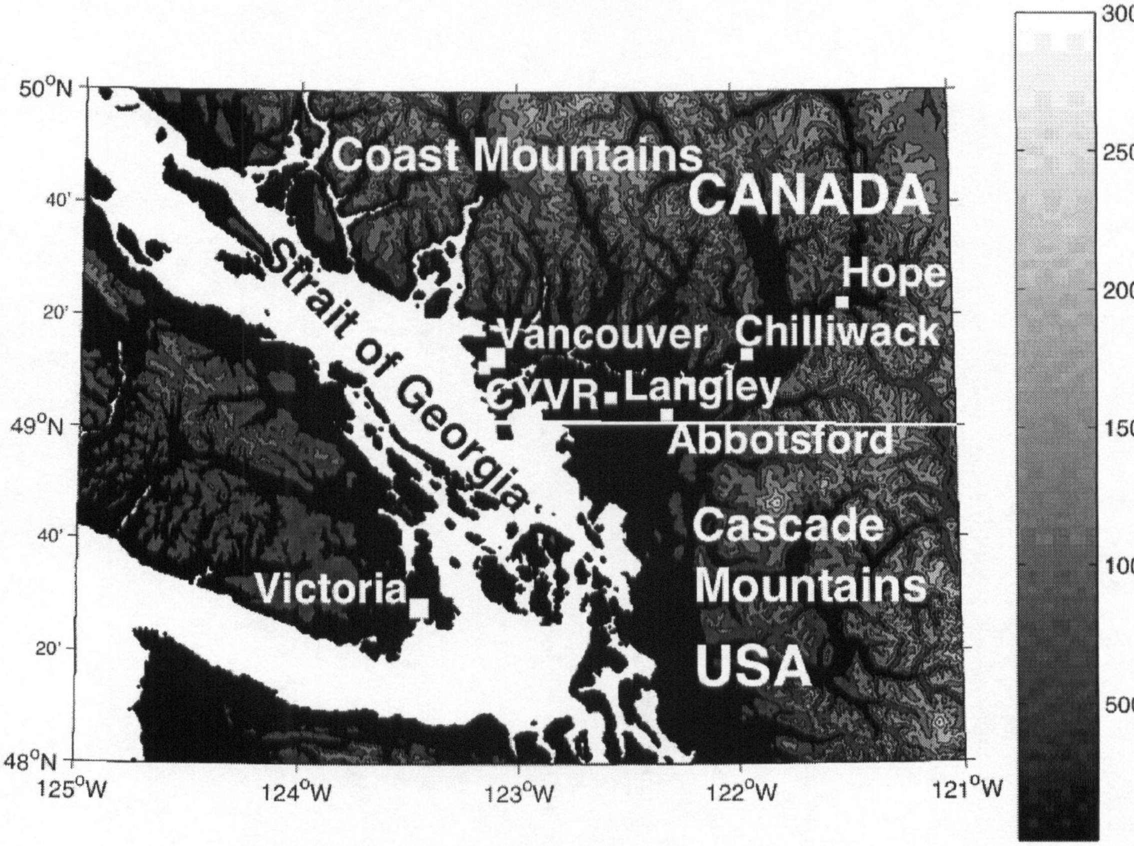

Figure 4.2: The Lower Fraser Valley is a floodplain spanning the ozone stations of Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope. The triangular valley is widest near CYVR along the coast of the Georgia Strait, and tapers to a narrow gorge between steep mountain walls near Hope. Shading (vertical bar at right) indicates terrain elevation above sea level.

Table 4.1: Ensemble members included in each of the 13 ensemble groups. "Base" is the forecast obtained by running CMAQ with the base emissions at one of the two possible resolutions (12 or 4 km) driven by NWP models (MC2 or MM5). NO_x indicates runs with perturbations of ± 50% NO_x, VOC includes the ± 50% VOC runs, and NO_xVOC represents the run with plus 50% NO_x combined with minus 50% VOC, and the run with minus 50% NO_x combined with plus 50% VOC. Last column indicates the size (number of forecasts included in the ensemble) of each of the 13 ensemble groups.

Ensemble	MC2-CMAQ								MM5-CMAQ								Size
	12 km				4 km				12 km				4 km				
	Base	NO_x	VOC	NO_xVOC	Base	NO_x	VOC	NO_xVOC	Base	NO_x	VOC	NO_xVOC	Base	NO_x	VOC	NO_xVOC	
ALL	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	28
MET+NO_x	•	•			•	•			•	•			•	•			12
MET+VOC	•		•		•		•		•		•		•		•		12
MET+NO_xVOC	•			•	•			•	•			•	•			•	12
MC2-12	•	•	•	•													7
MC2-04					•	•	•	•									7
MM5-12									•	•	•	•					7
MM5-04													•	•	•	•	7
MET	•				•				•				•				4
12-km	•	•	•	•					•	•	•	•					14
04-km					•	•	•	•					•	•	•	•	14
MC2-ALL	•	•	•	•	•	•	•	•									14
MM5-ALL									•	•	•	•	•	•	•	•	14

- All the ensemble members driven by MC2 at 4 km (MC2-04, seven members).

- All the ensemble members driven by MM5 at 12 km (MM5-12, seven members).

- All the ensemble members driven by MM5 at 4 km (MM5-04, seven members).

- All the control runs (MET, four members).

- All the ensemble members with 12 km resolutions (12-km, 14 members).

- All the ensemble members with 4 km resolution (04-km, 14 members).

- All the ensemble members driven by MC2 (MC2-ALL, 14 members).

- All the ensemble members driven by MM5 (MM5-ALL, 14 members).

MET+NO$_x$, MET+VOC, and MET+NO$_x$VOC are ensembles generated with both meteorology and emission perturbations, while MC2-12, MC2-04, MM5-12, and MM5-04 are ensembles where only emissions perturbations are considered (i.e., the members in each of them are driven by the same meteorological input field). MET, being formed by the four control runs, takes into account meteorology perturbations from NWP model differences alone.

Ensembles 12-km and 04-km will help to understand the effects of different horizontal grid spacing. Finally, MC2-ALL and MM5-ALL give insights about the different contributions from different NWP models (MC2 and MM5) while including different spatial resolutions.

4.3 Probabilistic-Forecast Verification Statistics

A probabilistic forecast system (PFS) can be built from a given set of ensemble members by computing the probability of an event occurrence. This probability can be computed as the

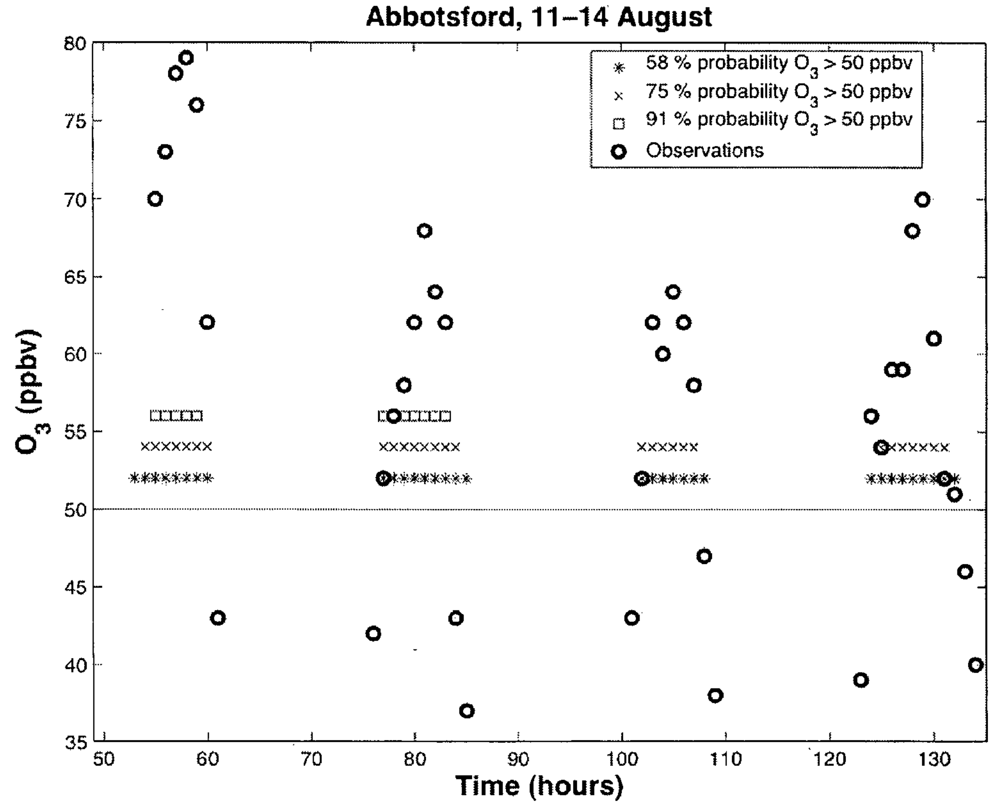

Figure 4.3: Probabilities of ozone concentrations above 50 ppbv, as predicted by MET-NO$_x$ at Abbotsford, 11-14 August 2004. Asterisks indicates hours when the forecasted probability is 58 % (seven out of 12 ensemble members are predicting the event), crosses when it is 75 % (nine out of 12), and squares when it is 91 % (11 out of 12). The continuous line represents to ozone 50 ppbv concentration threshold. Circles indicate observations.

ratio of the number of the ensemble members that predict the event over the total number of members. For an ozone PFS, the event can be the probability of ozone concentration above a certain threshold. Figure 4.3 is an example of the probabilities forecasted by the MET+NO$_x$ ensemble, at Abbotsford, 11-14 August 2004.

Probabilistic forecast skill can be evaluated by determining the predictive accuracy of a forecast distribution, and also the ability to distinguish the relative frequency of different events. With this in mind two important forecast attributes can be defined: resolution and reliability. Both are concerned with the conditional probability $p(o|f)$ of observation (o) given

Table 4.2: Out of the 549 valid observation points available, this table shows the percent of observations with ozone concentration greater than the given threshold.

Threshold (ppbv)	10	20	30	40	50	60	70	80
Occurrence (%)	79	63	46	34	25	15	7	3

forecast (f). An in-depth discussion of those and other attributes of probabilistic forecasts can be found in Joliffe and Stephenson (2003).

4.3.1 Resolution

Resolution measures the ability of the forecast to sort, a priori, the observed events into separate groups, when the events considered have a frequency different from the climatological frequency. For an ozone PFS, two different events could be the probabilities of ozone concentrations above two different thresholds. A PFS with good resolution should be able to separate the observed concentrations when the two different probabilities are forecasted. Table 4.2 shows the concentration threshold values used in this study. As the threshold concentration increases, the percentage of the available event occurrences greater than this threshold decreases. For threshold values above the 60 ppbv limit (an event occurring 15 % of the time) the low number of observation points available yields a large sampling uncertainty. Nevertheless, these threshold values are included in this analysis, since it is interesting to see how the ensembles behave for high (important for health-related issues) but rarely observed ozone-concentration values.

Resolution can be measured with Relative Operating Characteristics (ROC), developed in the field of signal-detection theory for discrimination of two alternative outcomes (Mason, 1982). A contingency table of observed versus forecasted event occurrences is built separately for individual forecast probability values (a probability value can be defined as the percentage of ensemble members forecasting a given event). The hit rate is computed as the ratio of the number of correct forecasts of the event to the total number of occurrences of the event, while the false-alarm rate is computed as ratio of the number of non-correct forecasts of the event to the total number of non-occurrences of the event. Then, hit rates are plotted on the ordinate against the corresponding false-alarm rates on the abscissa to generate the ROC curve. For a PFS with good resolution, the ROC curve is close to the upper left hand corner of the graph. The area under the ROC quantifies the ability of an ensemble to discriminate between events, which can be equated to forecast usefulness, and is known also as the ROC score (Mason and Graham, 1999). The closer the area is to one, the more useful the forecast is. A value of 0.5 indicates that the forecast system has no skill, as when the predicted events have a climatoligical frequency. The ROC curve does not depend on the forecast bias, hence is independent of reliability (defined below). It represents the PFS intrinsic value.

Figure 4.4 shows an example of a ROC curve for the "ALL" ensemble (28 members), for observed ozone concentration above 50 ppbv. The shaded portion of the plot represents the ROC area, and the dashed line is the ROC curve for a chance forecast. A contingency table is constructed for each probability threshold (the labels adjacent to the asterisks), where the probability threshold in this example assumes the values from 0/28 to 28/28, with increments of 1/28. Hit and false-alarm rates are computed for each contingency table (i.e., for each probability threshold). In this example, a correct forecast of the event occurs if the forecast probability (ratio of the number of the ensemble members that predict the event over the total number of members) is above the given probability threshold when the observed ozone concentration is above 50 ppbv. Similar curves can be produced for the other concentration thresholds.

4.3.2 Reliability

Reliability measures the capability of the PFS to predict unbiased estimates of the observed frequency associated with different forecast probabilities. In a perfectly reliable forecast, the forecasted probability of the event should be equal to the observed frequency of the event for all the cases when that specific probability value is forecasted. It can be improved with a forecast calibration such as bias correction; e.g., by re-assigning the forecast probability values based on a long series of past forecasts, or by Kalman filtering each individual forecast based on recent past bias values, as discussed in Chapter 5. Reliability alone is not sufficient to establish if a PFS produces valuable forecasts or not. For instance, a system that always forecasts the climatological probability of an event is reliable, but not useful.

Reliability can be measured with a Talagrand diagram (Talagrand and Vautard, 1997), also known as a rank histogram (Hamill and Colucci, 1997). First, the ensemble members are ranked for each prediction. Then, the frequency of an event occurrence in each bin of the rank histogram is computed and plotted against the bins. The number of bins equals the number of members plus one. A perfectly reliable PFS shows a flat Talagrand diagram, where all the bins have the same height ("ideal bin height"). In fact, if each ensemble member represents an equally likely time evolution and space distribution of the ozone concentration, then the ensemble exhibits a perfect spread, and the observations are equally likely to fall between any two members.

In this study a new summary index, called a "reliability index" (RI), is introduced as the

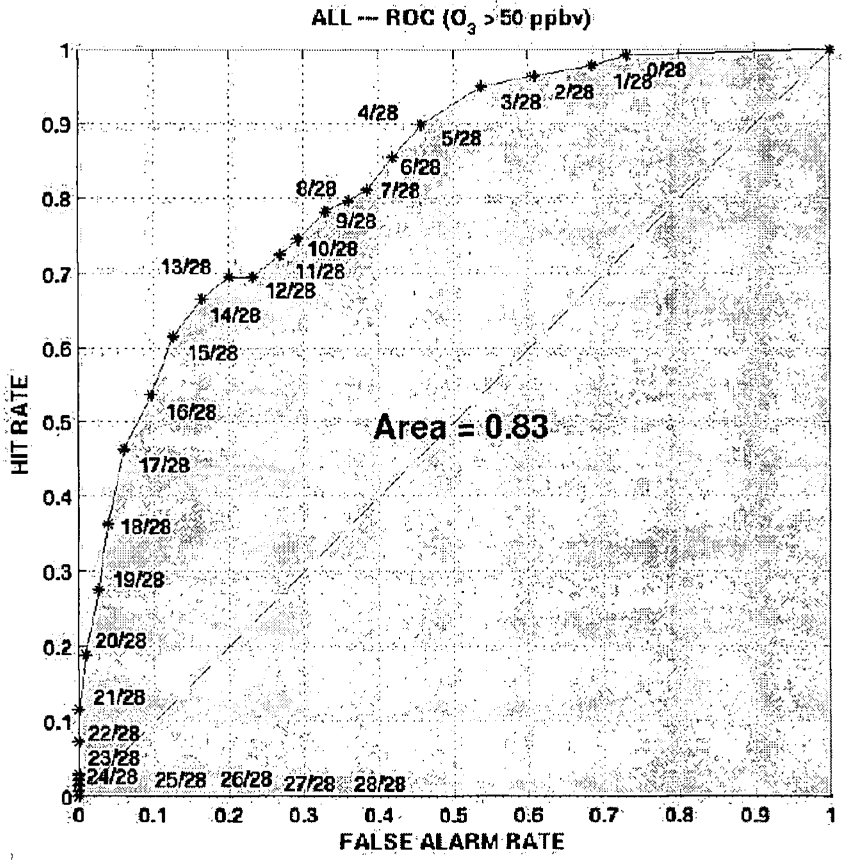

Figure 4.4: ROC curve for the "ALL" ensemble (28 members), for observed ozone concentration above 50 ppbv. The better the probabilistic forecast, the closer the ROC curve is to the upper left corner. The shaded portion of the plot represents the ROC area (large areas are better), and the dashed line is the ROC curve for a chance forecast. Hit rates are plotted on the ordinate against the corresponding false-alarm rates on the abscissa, to generate the ROC curve for each probability threshold (the labels adjacent to the asterisks), where the probability threshold assumes values from 0/28 to 28/28, with increments of 1/28.

reliability attribute. It is computed as follows:

$$\frac{mean\,bin\,distance\,from\,ideal\,bin\,height}{ideal\,bin\,height} \times 100 \tag{4.1}$$

$$= \frac{\frac{1}{N_{bin}} \sum_{i=1}^{N_{bin}} |\frac{count_i}{N_{point}} - \frac{1}{N_{bin}}|}{\frac{1}{N_{bin}}} \times 100 \tag{4.2}$$

$$= \sum_{i=1}^{N_{bin}} |\frac{count_i}{N_{point}} - \frac{1}{N_{bin}}| \times 100 \tag{4.3}$$

where N_{bin} is the Talagrand diagram number of bins (corresponding to the number of ensemble members plus one), $count_i$ is the number of times the observed event falls into the ith bin, and N_{point} is the sum of $count_i$, for $i = 1, \cdots, N_{bin}$ (i.e., the sample size).

Lower RI (i.e., closer to zero) means that the bins are closer to the ideal bin height. The Talagrand diagram of the 13 PFSs all have similar shapes, as shown in the next Section, so this index can be useful to better discriminate between their reliabilities. The RI does not provide any information about the Talagrand diagram shape.

The RI index, as define in Equation 4.3 tends to increase with increasing ensemble size, if the ensembles are samples drawn from the same distribution. This would prevent its application in cases as here, where ensembles with different sizes are compared with each other. For this reason, Equation 4.3 is normalized as follows:

$$RI = \frac{\sum_{i=1}^{N_{bin}} |\frac{count_i}{N_{point}} - \frac{1}{N_{bin}}| \times 100}{\sqrt{\frac{esize}{esizemin}}} \tag{4.4}$$

where *esize* is the size of the ensemble for which RI is computed, and *esizemin* is the size of the smallest ensemble considered. Hereafter, this normalized expression is used because it makes RI independent of ensemble size. Again, lower RI is better.

Tests of this normalization are performed by computing RI (using Equation 4.4) with ensembles predicting the same distribution but having different size. Results gave the same RI value, as desired, plus noise. The variance of the noise can be interpreted as an estimate of the sampling uncertainty, where a sample is an individual ensemble.

The RI (%) measures the degree of closeness of a Talagrand diagram to its ideal flat shape. Recently, a similar index (δ) measuring the "deviation of the histogram from flatness" has been introduced by Candille and Talagrand (2005). This index takes into account the deviation from the ideal bin height by considering a sum over the squares of the differences of $count_i$ minus N_{point}/N_{bin} for $i = 1, \cdots, N_{bin}$, and by normalizing this quantity. When used to compare the reliability of different ensemble systems, it gives the same relative rankings as RI, but its values interpretation differs from RI. In fact, $\delta = 1$ means a perfectly reliable system, $\delta \gg 1$ suggests unreliability, and $\delta \ll 1$ indicates that "successive realization of the prediction process are not independent".

4.4 Probabilistic Forecast Results

In this section the resolution and reliability of the 13 PFSs are evaluated and discussed. The PFSs are divided into three groups: ensembles considering perturbations of both meteorology and emissions, ensembles based on only emission perturbations or only meteorology perturbations, and ensembles formed using the same model resolution or the same model. A summary of these analyses concludes this section.

4.4.1 Ensembles with both Meteorology and Emission Perturbations

The following are the ensembles generated by including both meteorology and emission perturbations: MET+NO_x, MET+VOC, MET+NO_xVOC (all three with 12 members), and ALL (28 members). These ensembles will be referred generally as PERT.

Figure 4.5 shows the area under the ROC curve and its variation using eight different concentration thresholds for each ensemble. The event being forecast is ozone concentration above the threshold. The higher the threshold, the less often the event occurs. Table 4.2 shows the percentage of occurrence of each event associated with the eight thresholds.

The probabilistic forecasts are best (ROC area larger than 0.8) for those threshold values between 40 and 70 ppbv (except MET+NO_xVOC with 70 ppbv). For low concentration values (10 and 30 ppbv) almost all the ROC-area values are below 0.7. For the highest threshold (80 ppbv) only ALL is above 0.7, and ensembles MET+VOC and MET+NO_xVOC have poor skill, with the latter below the 0.5 line. ALL and MET+NO_x most often outperform the other ensembles.

Figure 4.6 shows the Talagrand diagram for the PERT ensembles. The solid lines indicate the ideal shape (for a perfectly reliable diagram). All the panels show, to different degrees, a combination of a "U-shape" and a "L-shape". The U-shape indicates that spread of the ensemble is too small, because the observed event often falls outside the range of values sampled by the ensemble. In fact, the left-most bin contains an absolute frequency maximum (compared with the frequency of the other bins), while the right-most bin contains a relative frequency maximum. Furthermore, the asymmetric L-shape (maximum on the first bin) indicates that the ensembles are biased towards higher values compared to the observed ozone concentrations.

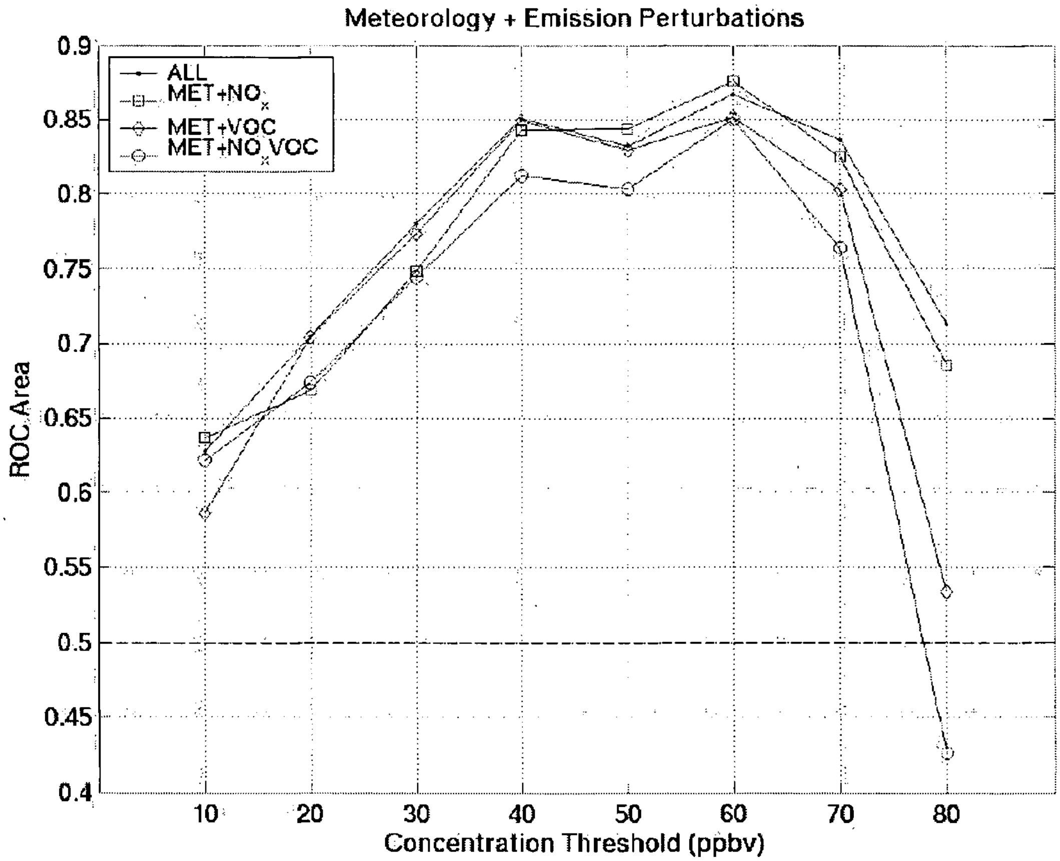

Figure 4.5: ROC-area values for 10 different concentration thresholds (from 10 to 80 ppbv, with increments of 10) and for the ensembles generated by including both meteorology and emission perturbations: ALL (28 members), MET+NO$_x$, MET+VOC, and MET+NO$_x$VOC (all three with 12 members). Values are within the interval [0, 1], with the perfect ROC-area = 1, and a no-skill ROC-area of 0.5 (dashed line).

Table 4.3 shows the RI values and the relative ranking based on these values. Among the PERT ensembles, ALL visually shows the least deficiencies (associated with the different shapes), followed by similar reliability for MET+NO$_x$VOC and MET+VOC (with the former slightly better, having a lower left-most maximum). MET+NO$_x$ is the ensemble showing the greatest positive bias among the four analyzed in this section, having the highest maximum in the first bin. This is confirmed by looking at RI, where ALL is most reliable within PERT (overall ranking 2) followed by MET+VOC (4), MET+NO$_x$VOC (7), and MET+NOx (9).

The MET+NO$_x$ tendency of overestimating more than the other ensembles in this group suggests that the ± 50 % NO$_x$ perturbation is not centered over an optimal estimate, and shifting the perturbations toward lower values could improve its forecast skill by reducing the positive bias. MET+VOC and MET+NO$_x$VOC also overestimate the measured ozone concentrations, suggesting that the same kind of perturbation shifting towards lower values could improve their forecast skill as well. This is confirmed by noticing in Figure 4.6 that all ensembles have a bump (around the fifth bin for ALL and around the third or fourth bin for the others) meaning that the observations fall more often in those bins than the neighboring bins. Ideally this bump should appear at the middle bin, so a perturbation shift towards lower values may move the bump more centrally.

Based on the above considerations, ALL is the best forecast by looking at both the probabilistic forecast resolution and reliability. ALL is formed by the largest number of members (28), and the observations fall more often within the maximum and minimum concentration predicted by its members at any given hour, compared with the other ensembles having only 12 members each (a subset of the ALL 28 members). This is certainly a desirable feature of

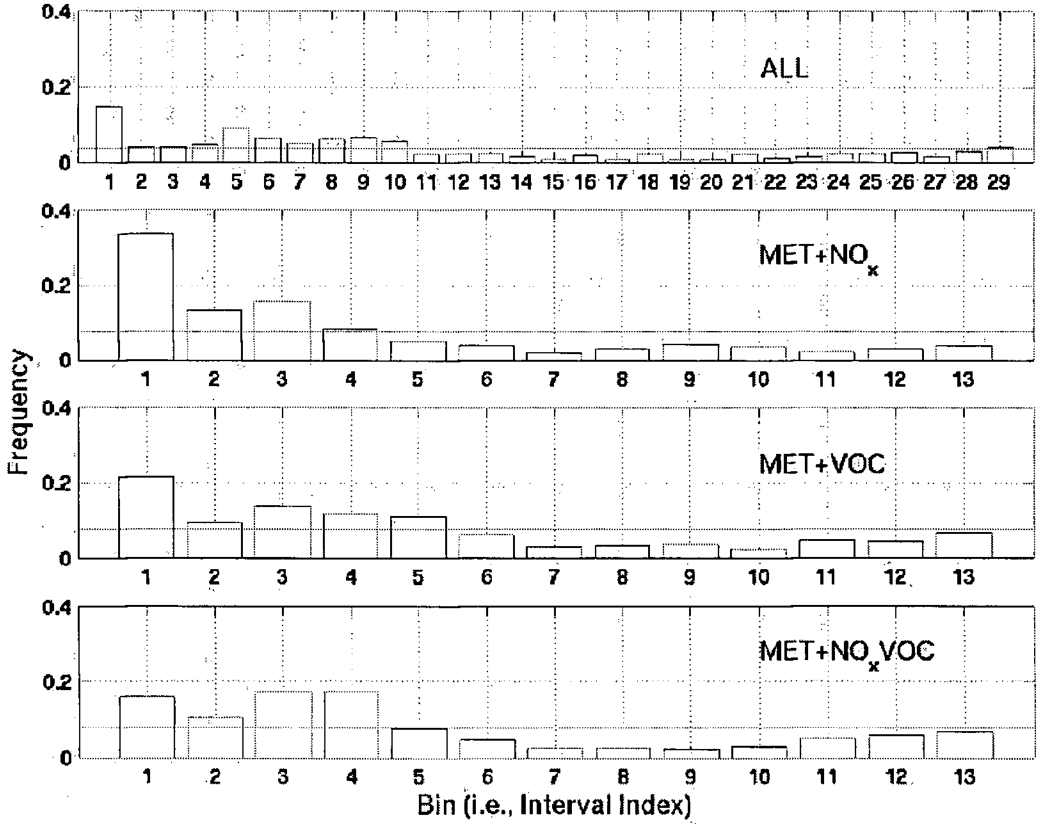

Figure 4.6: Talagrand diagram (rank histogram) for the ensembles generated by including both meteorology and emission perturbations (from top to the bottom panel): ALL (28 members), MET+NO$_x$, MET+VOC, and MET+NO$_x$VOC (all three with 12 members). The number of bins equals the number of ensemble members plus one. The solid horizontal line represents the perfect Talagrand diagram shape (flat). The closer the diagram to this horizontal line, the better.

Table 4.3: Normalized Reliability Index (RI) computed as in Equation 4.4, and the relative ranking based on RI values for each of the 13 ensemble groups. Smaller RI is better.

	ALL	MET+NO$_x$	MET+VOC	MET+NO$_x$VOC	MC2-12	MC2-04	MM5-12	MM5-04	MET	12-km	04-km	MC2-ALL	MM5-ALL
RI (%)	24	46	32	35	29	43	68	88	82	34	33	20	62
Ranking	2	9	4	7	3	8	11	13	12	6	5	1	10

an ensemble system in general. Moreover, with the NO_x, VOC and NO_x combined with VOC perturbations ALL is able to span more emission uncertainty than the other three forecasts.

Even though MET+NO_x is the most biased ensemble in this group, it shows very good probabilistic predictive skills, having ROC values similar to ALL, and better than any other PERT ensemble with a threshold value of 10, 50, and 60 ppbv. Over the five stations this means that the NO_x perturbation is more efficient than the VOC (or VOC combined with NO_x) perturbations in spanning the emission-uncertainty subspace.

The NO_x perturbation has much better predictive skill than the VOC perturbation for ozone above 80 ppbv. These high concentrations were observed in the afternoon mainly at Hope, except on 11 August at Chilliwack when a peak of 89 ppbv exceeded for three hours the 82 ppbv Canadian maximum 1-hour average acceptable ozone level. The fact that the NO_x perturbations outperform the VOC perturbations for ozone values above 80 ppbv suggests that when (afternoon) and where (eastern side of the LFV) these values are observed, the predominant chemical regime is NO_x-sensitive. NO_x-sensitive means that a percent change in NO_x results in a significantly greater change in ozone concentration relative to the same percent change in VOC (Sillman, 1999). It is beyond the goal of this study to analyze in-depth the predominant chemical regimes in the region, which would require several runs of a photochemical model with different VOC/ NO_x ratios (here only seven values of this ratio are utilized). Other studies using different approaches, i.e., without running complex 3-D CTM models, (e.g., Pryor, 1998; Ainslie, 2004) have instead suggested that the LFV is climatologically VOC-sensitive for the daily maximum.

Nevertheless, the results of this study suggest a NO_x-sensitive regime at Hope for this particular 11-15 August 2004 event, which can be explained as follows. The aged air mass from the urban core (the main NO_x source, located in the west and central parts of the LFV) is transported eastward by sea breezes. In the aged air mass, NO_x concentrations are reduced by the chemistry that produces ozone. In a NO_x-sensitive regime, a NO_x perturbation is more likely than a VOC one to capture ozone-concentration variability, and that is why MET+NO_x has much higher ROC-area values with the threshold of 80 ppbv than MET+VOC or MET+NO_xVOC. Also, the good probabilistic skill of MET+NO_x suggests that the ± 50 % limit for NO_x is appropriate, even though the perturbations themselves could be shifted towards lower values as discussed above.

4.4.2 Ensembles with only Meteorology or Emissions Perturbations

In this subsection the following ensembles are considered: MC2-12, MC2-04, MM5-12, and MM5-04 (all formed by seven members), and MET (four members). Since each of the first four PFSs is driven with the same meteorological input, they can be viewed as ensembles where only the emissions are perturbed. These ensembles are compared here with MET, that is an ensemble where only meteorology is perturbed. MET has only four members (while the others in this group have seven members), so the comparison with larger ensembles is a more stringent test for the meteorology perturbation than for the emissions perturbations.

Nevertheless, MET has the best ROC area for concentration thresholds of 40, 60 and 70 ppbv, and is very close to the best (MC2-04) for 50 ppbv (Figure 4.7). However, it has the worst performance for 80 ppbv (where the best is again MC2-04) because only one of its four ensemble members is predicting concentrations above this value. As will be shown in Section 4.5, the ensemble-averaged MET forecast is skillful in predicting the ozone peak. Even though three out of four of its members are always below 80 ppbv, they balance the highest

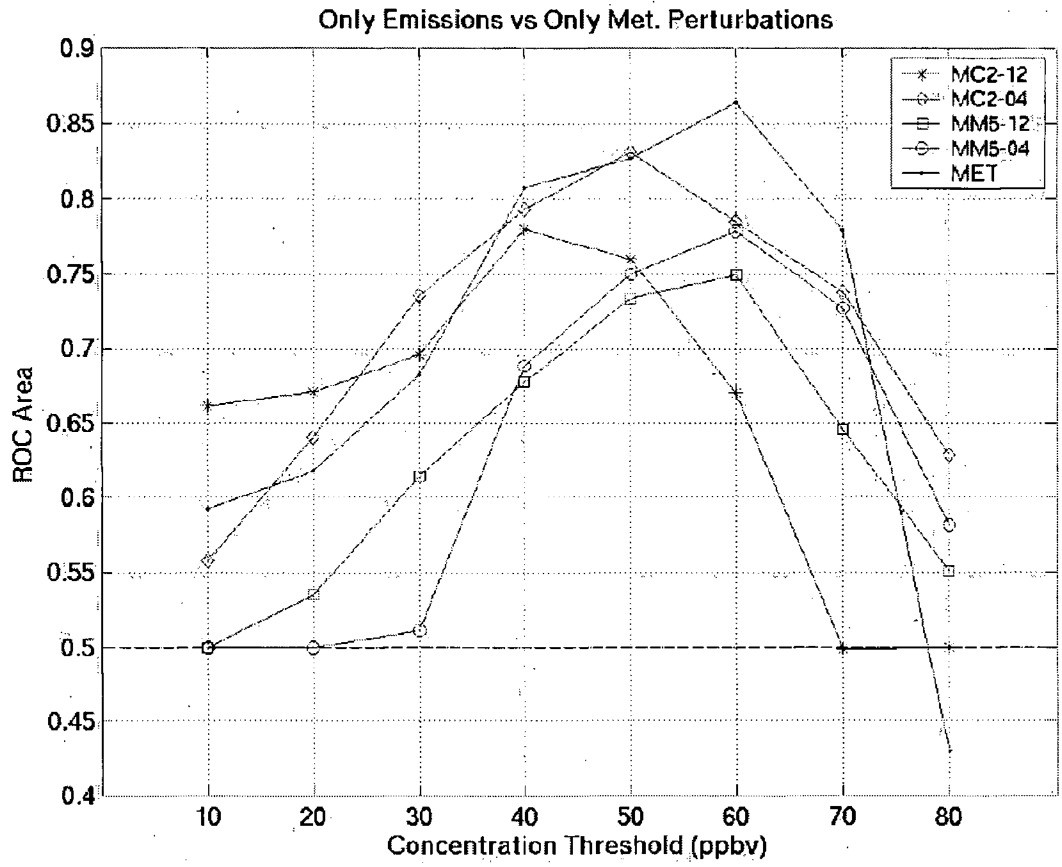

Figure 4.7: Similar to Figure 4.5, but with ROC-area values for the ensembles generated with only emission perturbations, i.e., the ensembles formed by forecasts driven with the same meteorological input (MC2-12, MC2-04, MM5-12, and MM5-04, all with seven members), or with only the meteorology perturbations (MET, four members).

peak prediction, resulting in a skillful ensemble mean for the ozone peak.

Among the ensembles with only the emission perturbations, the one showing the highest ROC-area values is MC2-04, and it is the best of this group for ozone thresholds from 30 to 80 ppbv. The MM5 ensembles including only emission perturbations (MM5-12 and MM5-04) have low ROC area values until 40 ppbv, and improve their performance relative to the other ensembles for threshold values above 40 ppbv. MC2-12 is the best for 10 and 20 ppbv, and the worst with 60 and 70 ppbv. At 80 ppbv it has a ROC area value of exactly 0.5, because it never predicts concentrations above this threshold. The 12 km runs are worse than the 4 km runs for high ozone values (with the thresholds of 70 and 80 ppbv), because the high values are mostly observed at Chilliwack and Hope, where the valley is much narrower than at the other locations, resulting in an advantage for the finer horizontal resolution runs.

Figure 4.8 shows the Talagrand diagram for these PFSs, where the solid lines have the same meaning as in Figure 4.6. Similar to Figure 4.6, U- and L-shaped diagrams are observed here. At the same time a maximum frequency is observed for MC2-04 at the central fifth bin, and to a lesser extent (relative maximum at fourth bin) for MC2-12. The central peak indicates less bias in the ensemble forecasts. These ensembles also have a larger spread than the ones with only the U- and L-shapes, as for the PERT ensemble set, but the spread is still too small.

Overall MC2-12 has the third best RI value (29 %), followed by MC2-04 (43 %). The two MM5 and the MET PFSs all have very high RI values (between 68 and 88), resulting in a worse overall ranking (11-13) as shown in Table 4.3. The reason is that they are highly positively biased, and this also results in the first bin being considerably higher than the others in the Talagrand diagram.

By comparing Figures 4.5 and 4.7, the utility of the meteorology and emission perturbations and their combination can be inferred. The predictive skill of the PERT ensembles (generated with both meteorology and emission perturbations) is superior to the ensembles with only the meteorology or only the emission perturbations for threshold values from 10 to 70 ppbv. For 80 ppbv, the best among those ensembles is MET+NO_x, while MC2-04, MM5-04, and MM5-12 are better than MET+VOC and MET+NO_xVOC.

Therefore the following can be deduced: both meteorology and emission perturbations are

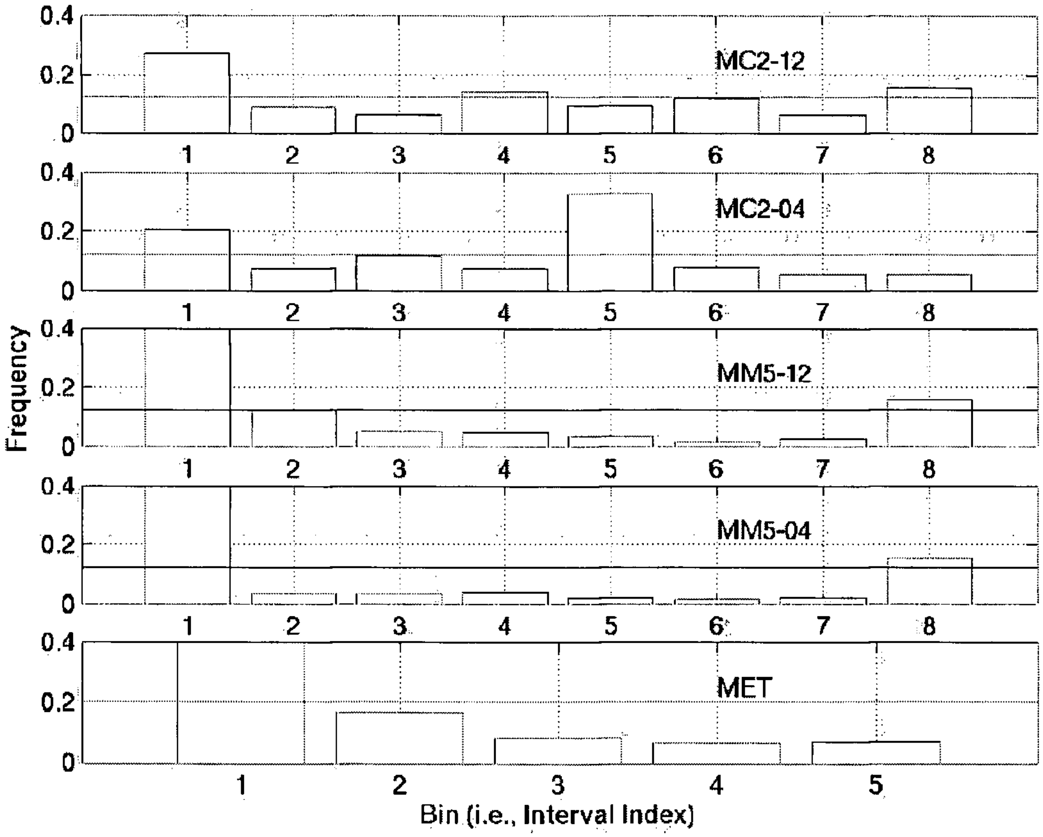

Figure 4.8: Similar to Figure 4.6, but for the ensembles generated with only emission perturbation, i.e., the ensembles formed by forecasts driven with the same meteorological input (MC2-12, MC2-04, MM5-12, and MM5-04, all with seven members), or with only the meteorology perturbation (MET, four members).

needed to have a skillful PFS, and neither one is sufficient to form a reliable PFS with a good resolution for all the threshold values. Moreover, the emission perturbations (particularly with NO_x) appear most important to capturing ozone concentrations above 80 ppbv.

4.4.3 Ensembles Generated with the Same Model or the Same Resolution

Here the PFS resolution and reliability for 12-km, 04-km, MC2-ALL and MM5-ALL are analyzed (they are all formed by 14 members). The intent is to observe the effect on the PFS skill of different horizontal grid resolutions, and different driving meteorological models.

Figure 4.9 shows the ROC area for these ensembles. MM5-ALL has the lowest values from 10 to 60 ppbv, and is slightly better than MC2-ALL with the concentration thresholds of 70 and 80 ppbv. 12-km is better than 04-km with thresholds of 10 or 20 ppbv and worse with the others, and 04-km is the best at 60, 70, and 80 ppbv. This may reflect the fact that higher concentrations were observed often in the eastern end of the LFV, where the topography progressively becomes more and more complex, giving a clear advantage to the finer resolution runs (as also discussed in Section 4.4.2). 04-km and MC2-ALL have very good ROC-area values (above 0.8) between 40 and 70 ppbv, while 12-km is above 0.8 only with 40 ppbv. MM5-ALL always has a ROC-area below approximately 0.78.

Figure 4.10 shows the Talagrand diagram for these PFSs. MC2-ALL has the smallest bias and MM5-ALL the largest. This corresponds to the overall best (20 %) and among the worst (62 %) RI values, respectively, as shown in Table 4.3. MM5-ALL has the smallest spread and MC2-ALL the largest (but still too small), by comparing the first and last bin heights. This suggests that the MC2 model has more of the needed variability than MM5 in the 5-day period analyzed in this study. Moreover, 04-km has a bigger spread and slightly less bias than 12-km

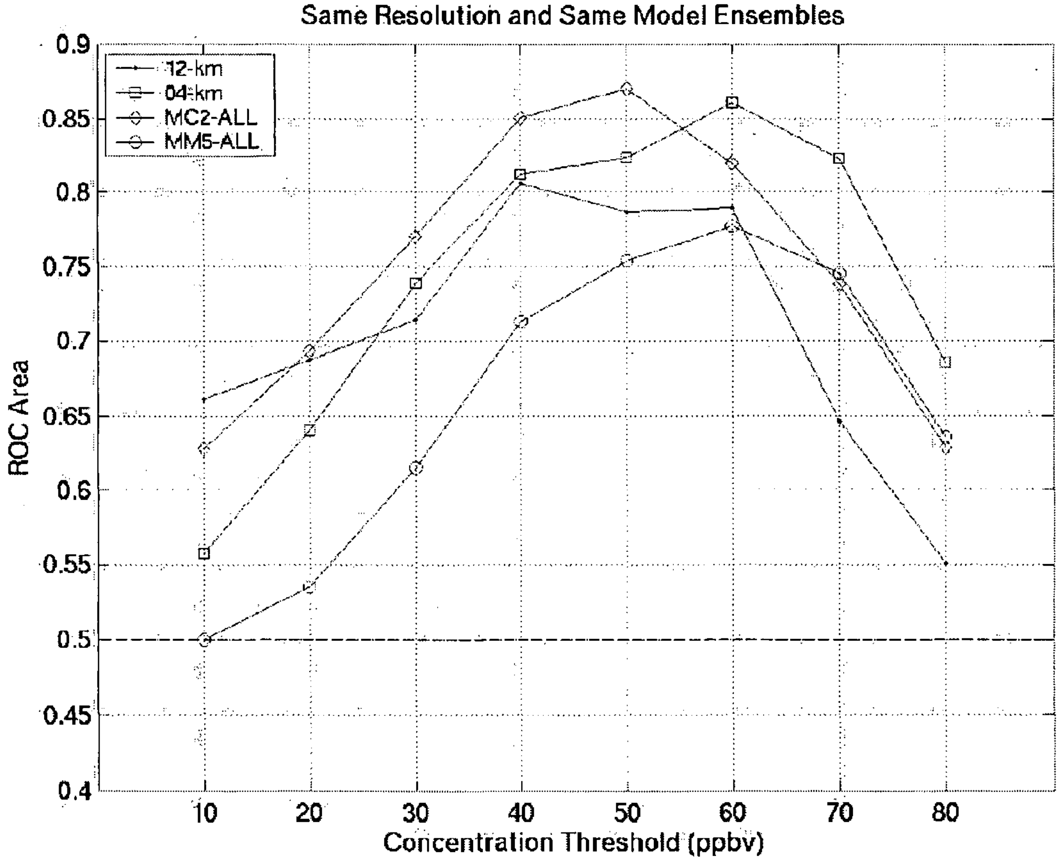

Figure 4.9: Similar to Figure 4.5, but for the ensembles formed with the same resolution runs (12-km and 04-km) or driven by the same Numerical Weather Prediction model (MC2-ALL or MM5-ALL).

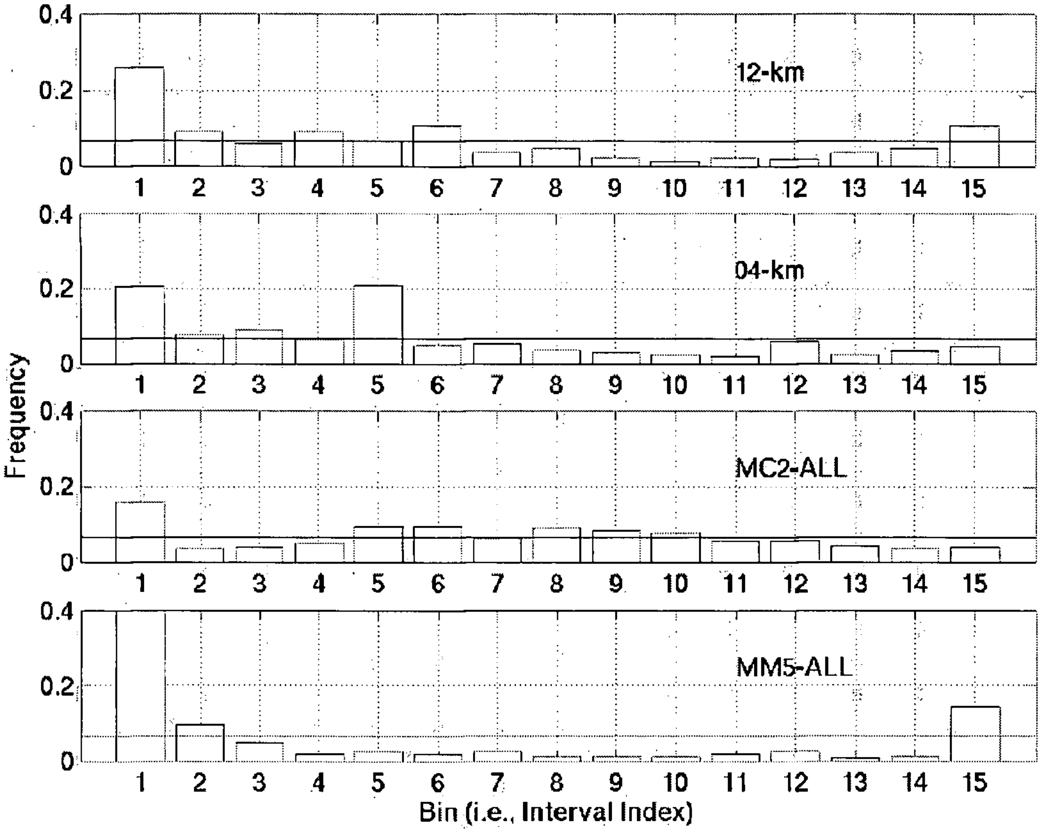

Figure 4.10: Similar to Figure 4.6, but for the ensembles formed with the same resolution runs (12-km and 04-km) or driven by the same Numerical Weather Prediction model (MC2-ALL or MM5-ALL).

(resulting in the fifth and sixth overall RIs, respectively).

Overall, by looking at the resolution and reliability of these ensembles built with different resolutions and models, MC2-ALL is the best for observed ozone concentrations below 60 ppbv, and 04-km has similar or better skills when higher ozone concentrations are measured, because it has better ROC-area values but is less reliable.

4.4.4 Summary

Figure 4.11 shows the ROC area for all the 13 PFSs, allowing an overall comparison of the PFS resolutions. ALL demonstrates the highest resolution, being the best at 30, 70 and 80 ppbv, and close to the best with the other thresholds. Figure 4.11 shows also that MET (with only four ensemble members) has improved resolution relative to the other PFSs at 40, 50 and 60 ppbv, while at 80 ppbv is among the worst along with MET+NO$_x$VOC. The subset of ensembles that includes only emission perturbations usually have low ROC area values, with the exception of MC2-12 which has the highest value (but still well below 0.7) for 10 ppbv. Perturbing only the meteorology, or only the emissions, results in a PFS with lower verification resolution than when both perturbations are considered. However, the emission perturbations are more important than the meteorology perturbations in capturing the highest ozone concentrations (above 80 ppbv).

If ALL is excluded from the PFS set, then MET+NO$_x$ and 04-km have the highest ROC area at 60, 70 and 80 ppbv. MET+NO$_x$ stays among the best even for lower concentration thresholds, while 04-km tends to lower verification resolution skill with lower ozone concentrations. Instead, by looking at the Talagrand diagram, 04-km (Figure 4.10) is certainly more reliable than MET+NO$_x$ (Figure 4.6), which is one of the most positively biased PFSs. However, the MET+NO$_x$ bias could be efficiently removed by Kalman filtering its forecasts (as discussed in Chapter 5), resulting in a reliable prediction.

The most reliable PFS is MC2-ALL, followed closely by ALL and then MC2-12. ALL certainly benefits from the highest number of ensemble members, making the extra computational effort worthwhile.

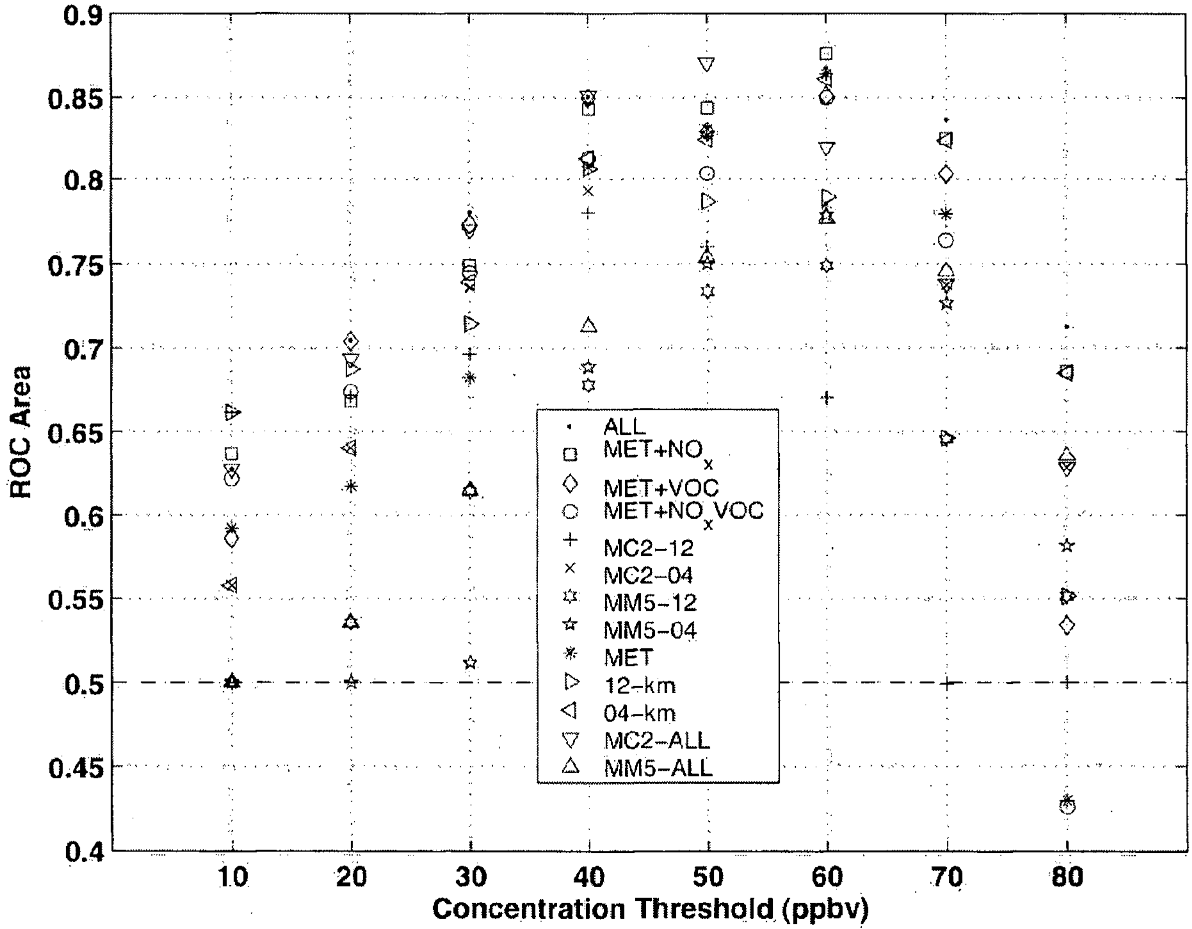

Figure 4.11: Similar to Figure 4.5, but for all the 13 ensemble groups considered in this study: all the forecasts available (ALL, 28 members), meteorology and NO$_x$ perturbations combined together (MET+NO$_x$, 12 members), meteorology and VOC perturbations (MET+VOC, 12 members), meteorology and NO$_x$ combined with VOC perturbations (MET+NO$_x$VOC, 12 members), all members driven by MC2 at 12 km (MC2-12, seven members), all members driven by MC2 at 4 km (MC2-04, seven members), all members driven by MM5 at 12 km (MM5-12, seven members), all members driven by MM5 at 4 km (MM5-04, seven members), and all the control runs (MET, four members), all the 12-km runs (12-km, 14 members), all the 4 km forecasts (04-km, 14 members), all members driven by MC2 (MC2-ALL, 14 members), and all members driven by MM5 (MM5-ALL, 14 members).

ALL appears to be the most useful probabilistic forecast, particularly because of its good resolution for high ozone concentrations, and because of its good reliability. Ensembles 04-km and MET+NO$_x$ closely follow. The choice of a particular PFS may be dictated by user needs, depending on which events are interesting (rare versus typical), the available computer power, and the importance of reliability versus resolution for a given situation.

4.5 Ensemble-mean Verification Statistics and Results

The ensemble mean of OEFSs is computed here as a linear average of the ensemble-member-predicted hourly concentrations. In this section the forecast skill of the ensemble means of the 13 OEFS groups are investigated. The ensemble means are analyzed because it has been found that they are the most skillful forecast when compared with the individual ensemble members against the observations, as shown in Chapter 2 and 3 and in McKeen et al. (2005).

The following subsections present and discuss the results by looking at correlation, RMSE, and unpaired peak prediction accuracy (UPPA). These discussions are then followed by a brief summary.

4.5.1 Correlation

Pearson product-moment coefficient of linear correlation (herein "correlation") can be computed as follows:

$$corr(s) = \frac{\sum_{t=1}^{N_{hour}} [C_o(t,s) - \overline{C_o(s)}][C_p(t,s) - \overline{C_p(s)}]}{\sqrt{\sum_{t=1}^{N_{hour}} [C_o(t,s) - \overline{C_o(s)}]^2 \sum_{t=1}^{N_{hour}} [C_p(t,s) - \overline{C_p(s)}]^2}} \tag{4.5}$$

where N_{hour} is the number of 1-h average concentrations over the 5-day period, $C_o(t, s)$ is the 1-h average observed concentration at a monitoring station s for hour t, $C_p(t, s)$ is the 1-h average predicted concentration at a monitoring station s for hour t , $\overline{C_o(s)}$ is the average of 1-h average observed concentrations at a monitoring station s over the 5-day period, $\overline{C_p(s)}$ is the average of 1-h average predicted concentrations at a monitoring station over the 5-day period.

We evaluate correlation to quantify timing errors of maximum and minimum concentrations at a specific location. The higher the correlation, the better is the match between the two signals; for example, the maximum ozone is predicted close to the right time of the day.

Figure 4.12 depicts correlation bar plots via five panels for each of the 13 OEFS groups presented in this study. Each panel shows the result for a different station, going from the west side of the LFV (CYVR), to the easternmost location (Hope). For comparison purposes, the ensemble means are listed on the abscissa following the same order they have been presented and grouped in Section 4.4. Moreover, the number at the bottom of each bar represents the ranking (1 being the best, 13 being the worst), computed for each station based on the individual correlation values.

Generally, correlation values tend to be lower moving towards the east side of the LFV, with all the ensembles having their poorest performance at Hope. Indeed Hope is located in a very steep narrow valley (less than 4 km wide), which none of the models are able to resolve. Since the 12 km runs do not see this valley, in the afternoon the ozone plume is advected past Hope (instead of being trapped there), resulting in decreasing values (after the plume passage) while in reality the concentration is increasing. Instead, when during the night a return flow (going westward) is established, the 12 km run tends to bring back the plume,

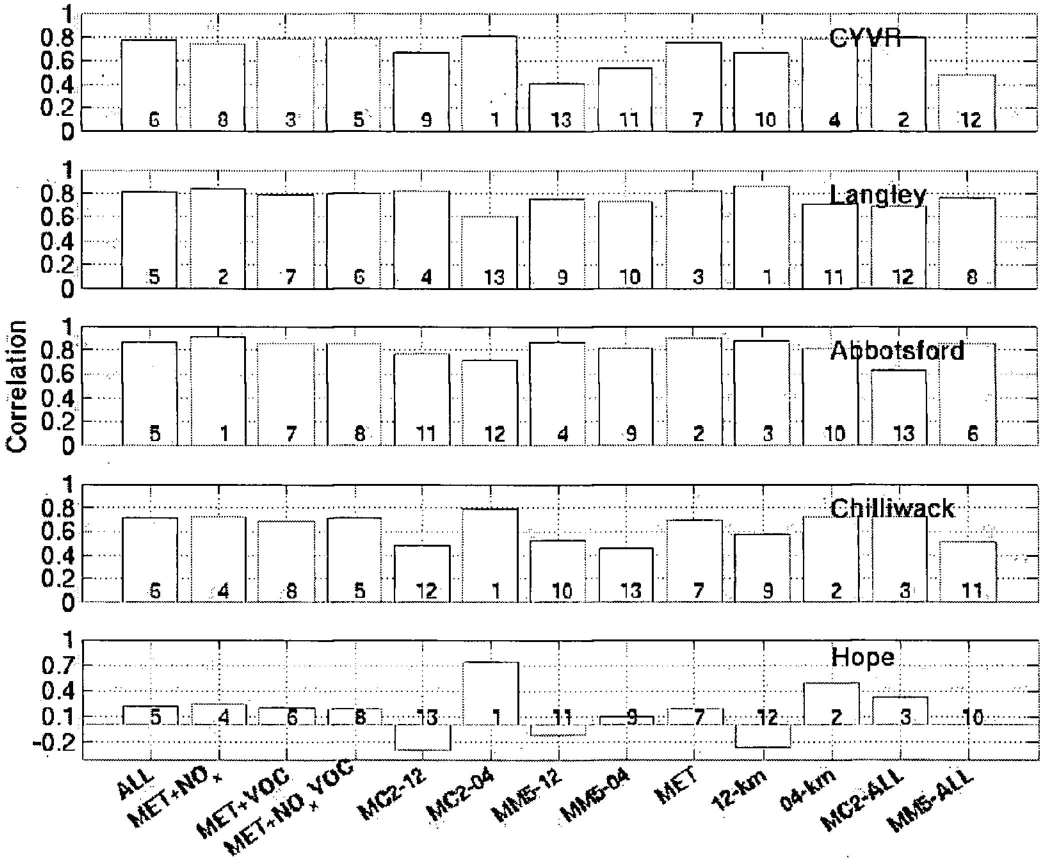

Figure 4.12: Correlation values between observed and predicted ozone 1-h average concentrations are plotted at five stations [Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope], for the 13 ensemble groups (as listed in Figure 4.11), for the 5-day period 11-15 August 2004. Values are within the interval [-1, 1], with correlation = 1 being the best possible value. The number at the bottom of each bar represents the ensemble ranking relative to the correlation values at each station.

resulting in increasing predicted concentrations when the observed ozone is decreasing. This results in negative correlation values for the 12 km runs, as shown in Figure 4.12. Because of that, the ensembles using finer resolution runs have better correlation values at Hope and Chilliwack, where the topography is more complex than at CYVR, Langley, and Abbotsford. Spatial resolutions even finer than 4 km would be needed to better capture these topographic effects.

Overall, MC2-04 shows the best correlation values, even though at Langley and Abbotsford it is the worst and second worst, but it still has a correlation of 0.61 and 0.71, respectively. The MC2-04 shows some utility at the challenging location of Hope, where its correlation is considerably higher than all the other ensembles. Conversely, MET+NO$_x$ and 12-km shows the best values at the central wider valley locations of Langley and Abbotsford. Also, 04-km is clearly better than 12-km at locations where the topography is complex (Chilliwack and Hope), or where the coastal settings (and the associated thermally driven circulations) are complicated, as at CYVR, which is located near the Georgia Strait waters.

MET, with only four ensemble members, has good correlation with the observations at Langley and Abbotsford, having a median (7^{th} rank) performance at the other locations. This means that meteorology plays an important role (as expected) in accurately predicting the location and timing of the ozone concentration. MM5-ALL is better at Langley and Abbotsford than MC2-ALL, but considerably worse elsewhere, underlying differences between the meteorological fields the two mesoscale models provide.

MET+NO$_x$ has the better correlation values among the PERT ensembles (the first four on the abscissa). ALL, despite the considerably higher number of ensemble members, is never the best among these four OEFSs, showing correlation values slightly better than the overall median values. In fact, more averaging in the larger ensemble smooths out the peaks and will, on average, lower the correlations.

4.5.2 RMSE

Root mean square error (RMSE) is expressed by:

$$RMSE(s) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C_p(t,s) - C_o(t,s)]^2} \tag{4.6}$$

RMSE gives important information about forecast skill in predicting the magnitude of ozone concentration, even though alone it does not draw a complete picture of a forecast value.

Figure 4.13 shows the 13 OEFS RMSE values, analogous to Figure 4.12. This metric clearly shows the difficulties of all the ensembles at Hope, and to a lesser degree at Chilliwack.

MC2-ALL shows the best performance with this metric, being among the first three ensembles everywhere. MM5-ALL has among the worst RMSE values. MC2-12 shows low RMSE values at CYVR, Langley and Abbotsford, while MC2-04 has low RMSE at Abbotsford, Chilliwack and Hope.

Again, mostly because the topographic complexity, 04-km is better than 12-km at Abbotsford, Chilliwack, and Hope, while the contrary is true at CYVR and Langley. Instead, all the PERT ensembles have similar RMSE at the five stations.

MET has very poor performance with RMSE. While the meteorology perturbation helps the ensemble mean to capture space and time variability of the ozone concentration field (as discussed with the correlation values), the same is not true for the magnitude of ozone concen-

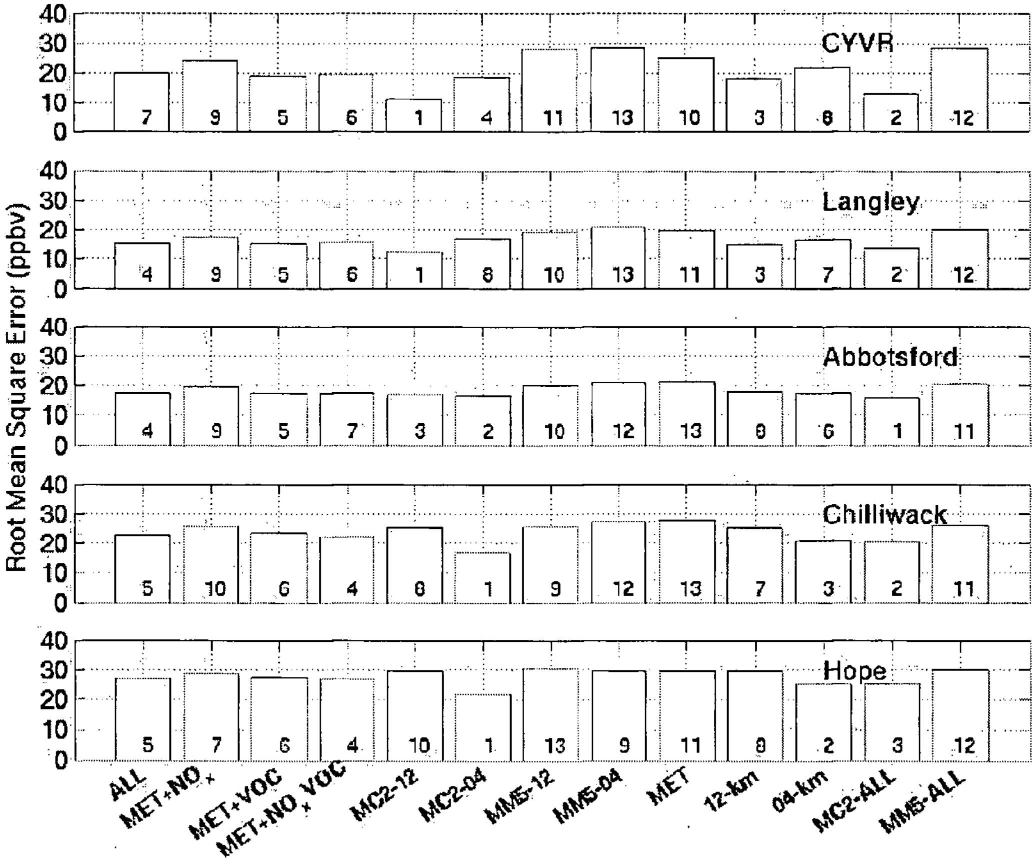

Figure 4.13: Root mean square error (RMSE) values (ppbv) at five stations [Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope] are plotted, for the 13 ensemble groups (as listed in Figure 4.11), for the 5-day period 11-15 August 2004. Values are within the interval [0, +∞], with a perfect forecast when RMSE = 0. The number at the bottom of each bar represents the ensemble ranking relative to the RMSE values at each station.

tration. This confirms the importance of emission perturbations in AQ ensemble forecasting, as already shown in Section 4.4.2, where it was needed to capture ozone concentrations above 80 ppbv.

RMSE can be separated into different components. One decomposition was proposed by Willmott (1981). First, an estimate of concentration $C^*(t,s)$ is defined as follows:

$$C^*(t,s) = a + bC_o(t,s) \tag{4.7}$$

where a and b are the least-square regression coefficients, and s is the observation station index. $C_p(t,s)$ and $C_o(t,s)$, are the predicted and observed ozone concentrations, respectively. Then the following two quantities can be defined:

$$RMSE_s(s) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C^*(t,s) - C_o(t,s)]^2} \tag{4.8}$$

$$RMSE_u(s) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C^*(t,s) - C_p(t,s)]^2} \tag{4.9}$$

where $RMSE_s(s)$ is the RMSE systematic component, while $RMSE_u(s)$ is the random one. $RMSE_s$ indicates the portion of error that depends on errors in the model, while $RMSE_u$ depends on random errors, on errors resulting from a model-skill deficiency in predicting a specific situation, and on initial-condition errors. The following relates RMSE to its components:

$$RMSE^2 = RMSE_s{}^2 + RMSE_u{}^2 \tag{4.10}$$

In Chapter 5 is discussed how $RMSE_s$ can be reduced with post-processing approaches such as the Kalman Filter bias-correction technique. Unfortunately, $RMSE_u$ reflects errors introduced by both model imperfections and initial-condition errors, and thus cannot be removed except by fundamental model improvements or better initial conditions.

Figure 4.14 shows $RMSE_s$ (bars bottom) and $RMSE_u$ (bars top) for the 13 OEFS groups at the five stations over the LFV. At Hope, there are the highest $RMSE_s$ values, meaning that all the ensembles can be improved with post-processing bias correction. CYVR shows instead among the highest $RMSE_u$ values, indicating an intrinsic lack of predictive skill at this location. Martilli and Steyn (2004) discuss the effects of the superimposed valley, slope, and thermal flows over the LFV. Often the pollution plume is transported during night over the Georgia Strait waters as a result of the combination of several transport processes. This makes it very challenging for the models to accurately predict the spatial and temporal evolution of ozone concentration in near water locations, such as CYVR, where the over-strait pool of pollutants can be re-advected over land by the daytime sea breeze.

Figure 4.14 also shows that the PERT ensemble means have a similar RMSE decomposition. $RMSE_u$ for 04-km is higher than for 12-km, and since these errors tend to grow more rapidly at smaller scales (i.e., high wavenumbers), the finer resolution could lose predictability faster than the coarser resolution due to rapid growth of the random errors. Also, MM5-ALL $RMSE_u$ is smaller than for MC2-ALL at CYVR, Abbotsford and Chilliwack. Finally, MC2-ALL, which has overall the best RMSE values, still can be considerably improved (via bias-removal techniques) at Chilliwack and Hope.

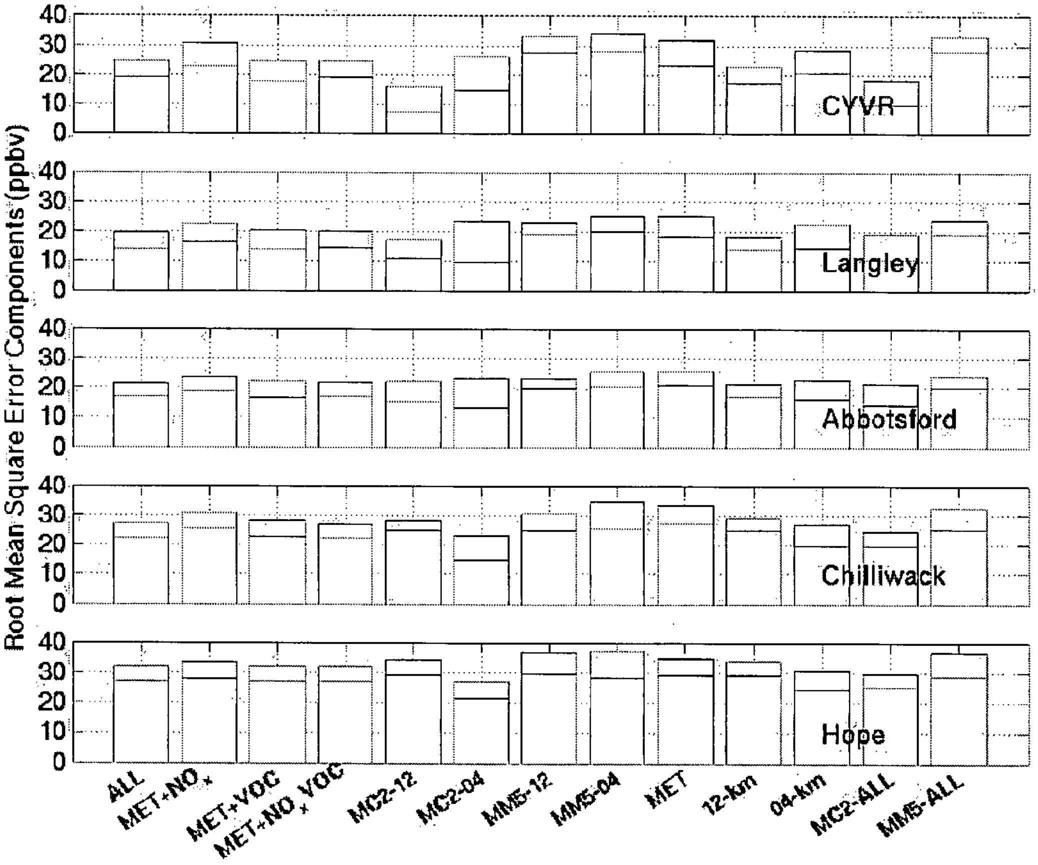

Figure 4.14: Similar to Figure 4.13, but for root mean square error systematic component (bottom bar) and unsystematic component (top bar).

4.5.3 UPPA

Unpaired peak prediction accuracy (UPPA) is computed as follows:

$$UPPA = \frac{1}{N_{day}} \sum_{day=1}^{N_{day}} \frac{|C_p(day,s)_{max} - C_o(day,s)_{max}|}{C_o(day,s)_{max}} \tag{4.11}$$

N_{day} is the number of days, $C_o(day,s)_{max}$ is the maximum 1-h average observed concentration at a monitoring station s observed during a one-day period, and $C_p(day,s)_{max}$ is the maximum 1-h average predicted concentration at a monitoring station s during the same day.

UPPA is included in the U.S. EPA guidelines EPA (1991) to analyze historical ozone episodes using photochemical grid models. The EPA acceptable performance upper-limit values are ± 20 %. UPPA is computed here as an average (over the five days available) of the absolute value of the normalized difference between the predicted and observed maximum at each station (Equation 4.11). UPPA is non-negative in our formulation, and only the + 20 % acceptance performance upper limit is used here.

UPPA has been chosen because it measures the ability of the forecasts to predict the ozone peak on a given day. In the past, peak concentrations have been the main concern for the public health, even though in recent years (over midlatitudes of the Northern Hemisphere) a rising trend has been observed in background ozone concentrations, while peak values are steadily decreasing (Vingarzan, 2004).

In Chapter 3 it has been discussed the possibility that ensemble averaging could cause excessive smoothing of the peak values. This has been improved in this Chapter by computing the ensemble-mean peak prediction as the average of the member predicted-ozone peaks. By doing an unpaired in time averaging for the peak values, the smoothing effect is avoided, and the ensemble-mean UPPA performance is improved.

Figure 4.15 shows the UPPA results. The solid line represents the EPA acceptance values for this parameter; forecasts below this line are desired. Only at Hope do all the ensembles have UPPA values above this limit (i.e., 20 ppbv), so this statistic confirms the difficulties that all the OEFSs have there.

MET+NO$_x$ has the best UPPA performance, confirming its good probabilistic predictive skill for high ozone concentration values, as shown in Section 4.4. MET+NO$_x$ is followed by MM5-04 and MET in the ranking based on UPPA. MET good performance is somewhat surprising because of its poor performance with RMSE. A comparison of the MET ensemble mean and the measured time series (not shown) confirms indeed that MET is accurate in replicating the maximum ozone (giving good UPPA, even if often the maximum is underestimated), it has reasonable timing of maxima and minima values with the observations (sufficiently good correlation), but it underestimates the other daylight observed values and largely overestimates the nighttime measured ozone (poor RMSE).

MC2-ALL has among the worse (higher) UPPA values, except at CYVR, and in fact MM5-ALL is clearly better with this parameter. MET+NO$_x$ is the best of the PERT ensembles, with good UPPA values at Langley, Abbotsford, and Chilliwack. Finally 04-km does a better job than 12-km (except at CYVR) in predicting the ozone peak magnitude.

4.5.4 Summary

In summary, the best performing ensemble-mean is MC2-04 for correlation, MC2-ALL with RMSE, and MET+NO$_x$ with UPPA. The ensemble mean computed with MC2-ALL also has a good performance with correlation, but performs poorly with UPPA. MC2-04 has good

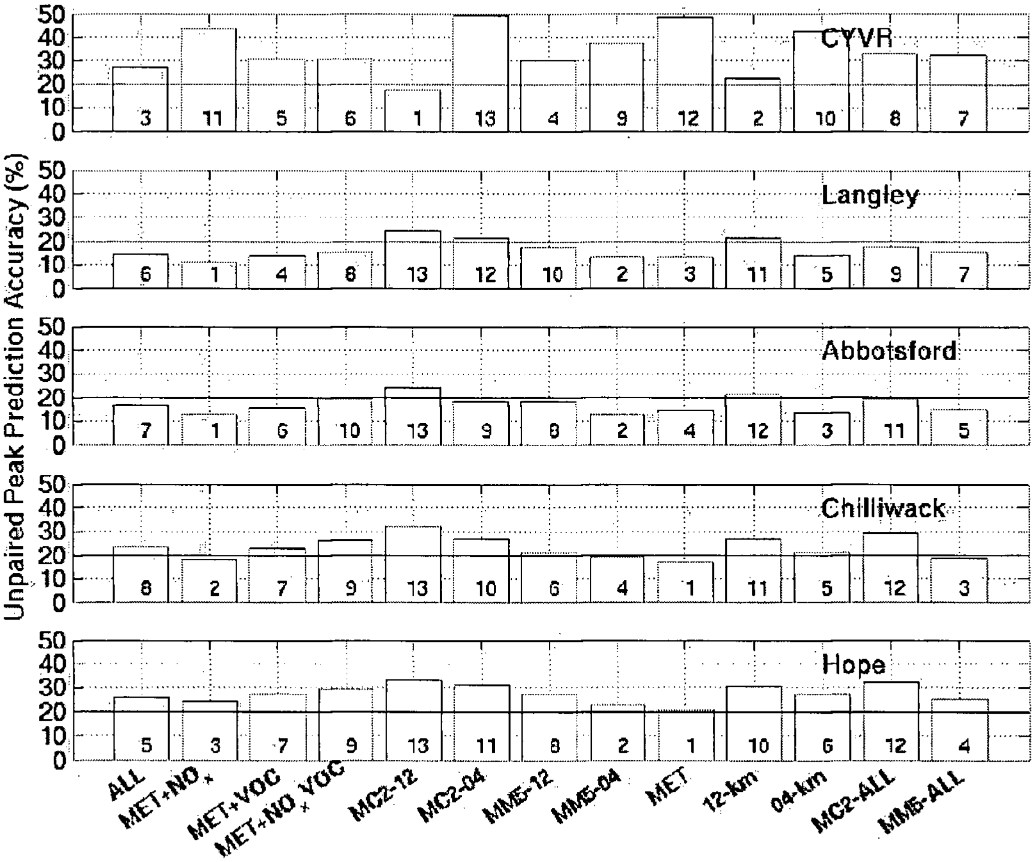

Figure 4.15: Unpaired peak prediction accuracy (UPPA) values (%) at five stations [Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope] are shown, for the 13 ensemble groups (as listed in Figure 4.11), for the 5-day period 11-15 August 2004. The ensemble peak prediction is computed as the average of the ensemble-member peak prediction. The continuous lines are the EPA acceptance values (+20 %). Values are within the interval [0, +∞], with a perfect peak forecast when UPPA = 0. The number at the bottom of each bar represents the ensemble ranking relative to the UPPA values at each station.

skill also with RMSE, but not with the remaining parameters. MET+NO_x has also a good performance with correlation, and is worse than average with RMSE.

Overall the MC2-ALL ensemble mean shows the best forecast skill by looking at the statistics presented here except for UPPA. This result would prevent its effectiveness for users most concerned with this metric. Different user needs can result in different relative importance among the statistical metrics. For example, if the main interest is in forecasting the ozone peak magnitude and timing, then MET+NO_x should be considered as the best ensemble-mean for this case study. If the computational resources are limited, then also MET (with only four ensemble members) has good skills at predicting the magnitude of the ozone peak.

Among MET+NO_x, MET+VOC, and MET+NO_xVOC, the MET+NO_x ensemble has the best resolution (Section 4.4.1). This agrees with the results presented in this Section, where MET+NO_x outperforms the others for unpaired peak prediction accuracy (i.e., for the highest values, which corresponds to its good ROC-area value at 80 ppbv) as shown in Figure 4.15. Moreover, its low reliability is confirmed by MET+NO_x having the highest bias, as shown in Figure 4.14 (systematic error).

4.6 Conclusions

This study is an analysis of the performance of 13 air-quality (AQ) ensemble groups, considering both probabilistic and ensemble-averaged ozone forecasts. Twenty-eight forecasts were generated over the Lower Fraser Valley (LFV), British Columbia (BC), Canada, for the 5-day period 11-15 August 2004, and compared with 1-h averaged measurements of ozone concentrations over five stations. The different forecasts are obtained by combining four driving meteorological input fields with seven emission scenarios: a control run, ± 50 % NO_x, ± 50 % VOC, and ± 50 % NO_x combined with VOC.

The driving meteorological fields are the output of two mesoscale models (run with 12 and 4 km horizontal spatial resolution): the Mesoscale Compressible Community (MC2) numerical weather prediction (NWP) model (Benoit et al., 1997) and the Penn State/NCAR mesoscale (MM5) model (Grell et al., 1994). The AQ forecasts are produced with the U.S. Environmental Protection Agency (EPA) Models-3/Community Multiscale Air Quality Model (CMAQ) Chemistry Transport Model (CTM) (Byun and Ching, 1991).

The following are the main findings of this study:

- Both meteorology and emission perturbations are needed to have a skillful probabilistic forecast system (PFS), and neither is sufficient alone to form a reliable PFS with a good resolution for the whole range of ozone concentrations.

- The meteorology perturbation is most important to capture the ozone temporal and spatial distribution.

- The emission perturbation is needed to accurately predict the ozone concentration magnitude.

- The emission perturbations are more important than the meteorology perturbations to capture high (and rarely measured) ozone concentrations, typically observed in the afternoon in areas such as the LFV where ozone production may be mainly attributed to local sources.

- Among the emission perturbations, NO_x perturbations resulted in more skillful probabilistic forecasts for the episode analyzed in this study.

- For all the emission perturbations, biases suggest the ± 50 % is not centered over an optimal estimate, and shifting the perturbations toward lower values could improve the forecasts by reducing the positive bias.

- Since NO_x has good (but positively biased) predictive skill, the ± 50 % limit appears to efficiently span the emission uncertainties space for this case.

- The ALL ensemble (formed by all the 28 ozone forecasts available) is the best probabilistic forecast, when considering both reliability and resolution.

- Ensemble averaging tends to smooth out the peak values (Chapter 3). However, this smoothing can be avoided if the ensemble-mean ozone peak is computed as the average of the ensemble-member peak predictions.

- The MC2 model has more variability than MM5 in the 5-day period analyzed in this study, and this resulted in MC2-ALL (formed by all the runs driven by MC2) being the most skillful ensemble-averaged ozone forecast. However, if the main interest is in forecasting the ozone peak magnitude and timing, then MET+NO_x should be considered as the best ensemble-averaged prediction.

- The root-mean-square-error random component for 04-km (formed by all the runs with 4 km horizontal spatial resolution) is higher than for 12-km (formed by all the 12 km runs). Since these errors tend to grow more rapidly at smaller scales (i.e., high wavenumbers), the finer resolution could lose predictability faster than the coarser resolution due to rapid random-error growth.

- With a hard limit on computational resources, the MET ensemble mean (with only four ensemble members; i.e., the control runs, where only meteorology is perturbed) is a viable option for predicting the magnitude of the ozone peak.

The results of this study suggest that future work should focus on OEFSs involving both meteorology and emissions perturbations. More specifically, the above findings suggest that the emission perturbations could be based on the time and spatial variability of different regimes. If (in a particular time of the day and on a subset of the spatial domain) a NO_x-sensitive regime is dominant, then a NO_x perturbation would be more useful than a VOC one to capture the ozone variability. Conversely, in VOC-sensitive regimes the VOC perturbations could be more efficient. In situations where neither of these two regimes is well defined, probably a combination of NO_x and VOC perturbations could be the best choice. These regimes could be identified in forecast mode by looking at the control-model forecasts, for example by evaluating the O_3/NO_y or H_2O_2/HNO_3 ratios (Sillman and He, 2002).

Ideally, each ensemble member should be an equally likely time evolution and space distribution of the ozone concentration, and they should all be equally good estimates of truth. With this in mind, the ensemble members should be "independent", in the sense that none of them should rely on other members for their realizations. This is not the case when nested grids are used, as for some of the PFSs used here (ALL, MET+NO_x, MET+VOC, MET+NO_xVOC, MC2-ALL, MM5-ALL, and MET). Namely, since CMAQ domains are linked using a 1-way nesting approach (similarly for MC2, but MM5 runs are implemented with 2-way nesting), all the 4 km runs cannot be considered independent of the runs where the driving meteorology or chemistry is their 12 km coarser domain.

The dependency among members of the same ensemble (no attempt has been done in this study to measure it) would result in an "effective" ensemble size smaller than the actual ensemble size. Moreover, a subset of the dependent members will span approximately the same subspace of the AQ modeling uncertainty space (or at least they should be closer to each other than to other members), resulting in both probabilistic and ensemble-averaged forecasts relying too heavily on the performances of these members than on others.

Finally, ensemble weather forecasts often provide information on the reliability of the forecasts; if the ensemble members have a large spread (defined as the standard deviation of the ensemble members about the ensemble mean), this implies less confidence in the forecast. However, similarly to Chapter 3 and 5, in this Chapter no correlation or relationship between ensemble spread and forecast error has been found.

4.7 References for Chapter 4

Ainslie, B., 2004: *A photochemical model based on a scaling analysis of ozone photochemistry*. Ph.D. thesis, 311 pp., University of British Columbia, Vancouver, Canada.

Benoit, R., M. Desgagne, P. Pellerin, S. Pellerin, Y. Chartier, and S. Desjardins, 1997: The Canadian MC2: A semi-Lagrangian, semi-implicit wide band atmospheric model suited for fine scale process studies and simulation. *Monthly Weather Review*, **125**, 2382–2415.

Brauer, M. and J. R. Brook, 1995: Personal and fixed-site ozone measurements with a passive sampler. *Journal of the Air and Waste Management Association*, **45**, 529–537.

Brown, R. P., T. Butler, and S. W. Hawley, 2001: *Ageing of Rubber - Accelerated Weathering and Ozone Test Results*. Rapra, Shawbury, United Kingdom.

Byun, D. W. and J. K. S. Ching, 1991: Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) modeling system. Technical Report EPA/600/R-99/030, U.S. Environmental Protection Agency.

Candille, G. and O. Talagrand, 2005: Evaluation of probabilistic prediction system for a scalar variable. *Quarterly Journal of Royal Meteorological Society*, **131**, 2131–2150.

Dabberdt, W. F., M. A. Carroll, D. Baumgardner, G. Carmichael, R. Cohen, T. Dye, J. Ellis, G. Grell, S. Grimmond, S. Hanna, J. Irwin, B. Lamb, S. Madronich, J. McQueen, J. Meagher, T. Odman, J. Pleim, H. P. Schmid, and D. Westphal, 2003: Meteorological research needs for improved air quality forecasting: report of the 11^{th} Prospectus Development Team of the U.S. Weather Research Program. Technical report, National Center for Atmospheric Research.

Dabberdt, W. F. and E. Miller, 2000: Uncertainty, ensembles and air quality dispersion modeling: applications and challenges. *Atmospheric Environment*, **34**, 4667–4673.

Delle Monache, L., X. Deng, Y. Zhou, H. Modzelewski, G. Hicks, T. Cannon, R. B. Stull, and C. di Cenzo, 2004: Air quality ensemble forecast over the Lower Fraser Valley, British Columbia, Canada. In *27th NATO/CCMS Conference on Air Pollution Modeling*. Banff, Alberta.

EPA, 1991: Guideline for regulatory application of the urban airshed model. Technical Report EPA-450/4-91-013, U.S. Environmental Protection Agency.

Grell, G. A., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation PENN State/NCAR Mesoscale Model (MM5). Technical Report NCAR/TN-398 +STR, National Center for Amospheric Research.

GVRD, 2002: 2000 emissions inventory for the Lower Fraser Valley airshed. Technical report, Greater Vancouver Regional District.

Hamill, T. M. and S. J. Colucci, 1997: Verification of ETA-RSM shortrange ensemble forecasts. *Monthly Weather Review*, **126**, 1322–1327.

Hanna, S. R., L. Zhigang, H. C. Frey, N. Wheeler, J. Vukovich, S. Arunachalam, M. Fernau, and D. Hansen, 2001: Uncertainties in predicted ozone concentrations due to input uncertainties for the UAM-V photochemical grid model applied to the July 1995 OTAG domain. *Atmospheric Environment*, **35**, 100–118.

Horvath, S. M. and D. J. McKee, 1994: Acute and chronic health effects of Ozone. In *Tropospheric ozone, human health and agricultural aspects*. Lewis Publisher, Boca Raton, Florida, pages 39–84.

Joliffe, I. T. and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley and Sons, West Sussex.

Martilli, A. and D. G. Steyn, 2004: A numerical study of recirculation processes in the Lower Fraser Valley (British Columbia, Canada). In *27th NATO/CCMS Conference on Air Pollution Modeling*. Banff, Alberta.

Mason, I., 1982: A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30**, 291–303.

Mason, S. J. and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operative levels. *Weather and Forecasting*, **14**, 713–725.

McHenry, J. N., W. F. Ryan, N. L. Seaman, C. J. Coats Jr., J. Pudykiewicz, S. Arunachalam, and J. M. Vukovich, 2004: A real-time Eulerian photochemical model forecast system. *Bulletin of the American Meteorological Society*, **85**, 525–548.

McKeen, S. A., J. M. Wilczak, G. A. Grell, I. Djalalova, S. Peckham, E.-Y. Hsie, W. Gong, V. Bouchet, S. Menard, R. Moffet, J. McHenry, J. McQueen, Y. Tang, G. R. Carmichael, M. Pagowski, A. Chan, T. Dye, G. Frost, P. Lee, and R. Mathur, 2005: Assessment of an ensemble of seven real-time ozone forecasts over Eastern North America during the summer of 2004. *To appear on Journal of Geophisical Research.*

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The new ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of Royal Meteorological Society*, **122**, 73–119.

Pryor, S. C., 1998: A case study of emission changes and ozone responses. *Atmospheric Environment*, **32**, 123–131.

Runeckles, V., 2002: Effects on vegetation and ecosystems. In Suzuki, D., editor, *A Citizen's guide to air pollution*. Vancouver, British Columbia, pages 177–216.

Sillman, S., 1999: The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments. *Atmospheric Environment*, **33**, 1821–1845.

Sillman, S. and D. He, 2002: Some theoretical results concerning O_3-NO_x-VOC chemistry and NO_x-VOC indicators. *Journal of Geophysical Research*, **107**, 1–15.

Talagrand, O. and R. Vautard, 1997: Evaluation of probabilistic prediction systems. In *Proceedings ECMWF Workshop on Predictability*. ECMWF, Reading, United Kingdom, pages 1–25.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bulletin of the American Meteorological Society*, **74**, 2317–2330.

Vaughan, J., B. Lamb, C. Frei, R. Wilson, C. Bowman, C. Figueroa-Kaminsky, S. Otterson, M. Boyer, C. Mass, M. Albright, J. Koenig, A. Collingwood, M. Gilroy, and N. Maykut, 2004: A numerical daily air quality forecast system for the Pacific Northwest. *Bulletin of the American Meteorological Society*, **85**, 549–561.

Vingarzan, R., 2004: A review of surface ozone background levels and trends. *Atmospheric Environment*, **38**, 3431–3442.

Willmott, C. J., 1981: On the validation of models. *Physical Geography*, **2**, 184–194.

Chapter 5

Ozone Forecasts Kalman-filter Predictor Bias Correction

5.1 Introduction

[1] Chapter 3 presented a new Ozone Ensemble Forecast System (OEFS), composed of 12 forecasts created using four different meteorological inputs and three different emission scenarios. The meteorological fields were obtained by running two mesoscale numerical weather prediction (NWP) models over two nested domains with 12 and 4 km horizontal grid spacing. The emission scenarios were a control run, a run with 50 % more NO_x emissions, and a run with 50 % less. The 12 combinations of the meteorological and emission fields were used to drive the U.S. Environmental Protection Agency (EPA) Models-3/Community Multiscale Air Quality Model (CMAQ) Chemistry Transport Model (CTM) (Byun and Ching, 1991).

[1] A version of this chapter has been accepted for publication. Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. B. Stull, 2005: Ozone ensemble forecasts. Part II: a Kalman-filter predictor bias correction, *accepted in August 2005 to be published in the Journal of Geophysical Research.*

This OEFS has been tested for the period 11-15 August 2004 using data from five stations across the Lower Fraser Valley (LFV), British Columbia (BC), Canada, a region where the ozone modeling is particular challenging because of the complex coastal mountain setting. The main finding in Chapter 3 is that, for the locations and days used to test this new OEFS, the ensemble-mean is the most skillful forecast when tested against the observations, and compared to any other ensemble member.

The results in Chapter 3 show that all the forecasts have systematic errors (e.g., nighttime over prediction). This is a problem common to all CTMs (Russell and Dennis, 2000), caused by a poor representation of the nightime ABL (e.g., vertical eddy diffusivity) and errors in the emissions. In this Chapter the Kalman filter predictor (KFP) post-processing bias-correction method (Bozic, 1994) has been applied to each ozone forecast (the 12 ensemble members and the ensemble-mean) to improve the individual forecast skill for all sites where ozone observations are available. The KFP correction is an automatic post-processing method that uses the recent past observations and forecasts to estimate the model bias in the forecast, where bias here is defined as the "difference of the central location of the forecasts and the observations" (Joliffe and Stephenson, 2003). This estimate can then be used to correct the raw model prediction. It is a recursive, adaptive method that takes into account the time-variation of forecast error at a specific location.

Details of the Kalman algorithm are given in Section 5.2. Section 5.3 describes the experiment and methodology. In Section 5.4, the performance of the raw (i.e., not corrected), the KFP bias-corrected forecasts, the ensemble-mean of the KFP bias-corrected forecasts (EK, is a linear average of the KFP bias-corrected ensemble-member predicted hourly concentrations), and the KFP bias-corrected EK (KEK) are compared using the same data set and statistical parameters as in Chapter 3. Moreover, EK and KEK performances are compared with two other bias-correction methods; namely, the additive and multiplicative methods (Section 5.5). In Section 5.6 those results are discussed and conclusions are drawn.

5.2 The Kalman-filter-predictor Bias Correction

The Kalman filter (KF) is a recursive algorithm to estimate a signal from noisy measurements. For NWP model forecasts, it has been mainly used in data-assimilation schemes to improve the accuracy of the initial conditions for both NWP (e.g., Burgers et al., 1998; Hamill and Snyder, 2000; Houtekamer and Mitchell, 2001; Houtekamer et al., 2005) and air quality (AQ) forecasts (e.g., van Loon et al., 2000; Segers et al., 2005). The KF has also been used for NWP model forecasts as a predictor bias-correction method during post-processing of short-term weather forecasts (Homleid, 1995; Roeger et al., 2003), an approach that is extended here for AQ forecasts (i.e., ozone).

In a post-processing predictor bias-correction method, the information (i.e., recent past forecasts and observations) is used to revise the estimate of the current raw forecast. Previous bias values are used as input to KF. The filter estimates the systematic component of the forecast errors, or bias, which is often present in AQ forecasts as shown in Chapter 3 and as reported in the literature (e.g., Russell and Dennis, 2000). Once the future bias has been estimated, it can be removed from the forecast to produce an improved forecast. Such a corrected forecast should be statistically more accurate in a least-squares sense.

The KF models the true (unknown) forecast bias x_t at time t, by the previous true bias plus a white noise η term (Bozic, 1994):

$$x_{t|t-\Delta t} = x_{t-\Delta t|t-2\Delta t} + \eta_{t-\Delta t} \tag{5.1}$$

where $\eta_{t-\Delta t}$ is assumed uncorrelated in time, and is normally distributed with zero-mean and variance σ_η^2, Δt is a time lag (see Figure 5.1), and $t|t-\Delta t$ means that the value of the variable at time t depends on values at time $t-\Delta t$. Because of unresolved terrain features, numerical noise, lack of accuracy in the physical parameterizations, and errors in the observations themselves, the KF approach further assumes that the the forecast error y_t (forecast minus observation at time t) differs from truth by a random error term ϵ_t:

$$y_t = x_t + \epsilon_t = x_{t-\Delta t} + \eta_{t-\Delta t} + \epsilon_t \tag{5.2}$$

where ϵ_t is assumed uncorrelated in time and normally distributed with zero-mean and variance σ_ϵ^2.

Kalman (1960) showed that the optimal recursive predictor of x_t (derived by minimizing the expected mean-square error) can be written as a combination of the previous predicted bias and the previous measurement of the bias:

$$\hat{x}_{t+\Delta t|t} = \hat{x}_{t|t-\Delta t} + \beta_{t|t-\Delta t}(y_t - \hat{x}_{t|t-\Delta t}) \tag{5.3}$$

where a hat $\hat{()}$ indicates the estimate. The weighting factor β, called Kalman gain, can be calculated from:

$$\beta_{t|t-\Delta t} = \frac{p_{t-\Delta t|t-2\Delta t} + \sigma_\eta^2}{(p_{t-\Delta t|t-2\Delta t} + \sigma_\eta^2 + \sigma_\epsilon^2)} \tag{5.4}$$

where p is the expected mean square error, which can be computed as follows:

$$p_{t|t-\Delta t} = (p_{t-\Delta t|t-2\Delta t} + \sigma_\eta^2)(1 - \beta_{t|t-\Delta t}) \tag{5.5}$$

It can be shown (Dempster et al., 1977) that the time series

$$z_t = y_{t+\Delta t} - y_t = \eta_t + \epsilon_{t+\Delta t} - \epsilon_t \tag{5.6}$$

has variance

$$\sigma_z^2 = \sigma_\eta^2 + 2\sigma_\epsilon^2 \tag{5.7}$$

Assuming $r = \sigma_\eta / \sigma_\epsilon$, Equation (5.7) become:

$$\sigma_z^2 = r\sigma_\epsilon^2 + 2\sigma_\epsilon^2 = (2 + r)\sigma_\epsilon^2 \tag{5.8}$$

σ_ϵ^2 (which is a time-varying quantity) can be estimated with the Kalman algorithm itself (i.e., by substituting $\hat{x}$ with σ_ϵ^2 in Equation (5.3) in combination with Equation(5.8). Further details on the filter implementation are given in Appendix A.

Since here a time lag of $\Delta t = 24$ hours is used, today's forecast bias is estimated using yesterday's bias, which in turn was estimated using the day-before-yesterday's bias, and so on. Figure 5.1 shows the flow diagram of the Kalman filter algorithm. The difference between today's forecast error (y_t) and the portion of today's bias that was estimated yesterday ($\hat{x}_{t|t-\Delta t}$), is weighted by the Kalman gain to give the correction that was "learned" from previous errors. This correction is applied to yesterday's estimate of today's bias ($\hat{x}_{t|t-\Delta t}$) to produce today's estimate of the bias for tomorrow ($\hat{x}_{t+\Delta t|t}$). Thus, real-time AQ forecasts are possible by taking the raw forecast from a model such as CMAQ, and correcting it with the bias forecast from KF.

The KF algorithm will quickly and optimally converge (after few time-step (Δt) iterations) for any reasonable initial estimate of p_0 and β_0. However, the filter performance is sensitive to the ratio $\sigma_\eta/\sigma_\epsilon$. If the ratio is too high, the filter will place excessive confidence on the past forecasts, and will therefore fail to remove any error. On the other hand, if the ratio is too low, the filter will be unable to respond to changes in bias. Thus, there exists an optimal value for the ratio that is given by the climatology of the forecast region, which can be estimated by evaluating the filter performance in different situations with different meteorology and different AQ scenarios (not only for AQ episodes).

The data set presented in this study is not extended enough to compute an optimal ratio value that can also be used for different AQ scenarios (i.e., non episodic). A ratio value of 0.01 is used in this study. This is the value from previous studies where the KF was used to bias-correct weather forecasts in the steep mountains of BC, Canada (Roeger et al., 2003), and close to the optimal value found in Homleid (1995); i.e., 0.06. With the availability of a longer data set (a full month or season), including both ozone forecasts and observations with a broader variability than just the AQ episode presented here, a different optimal value may result.

A period of two days (9-10 August 2004) is used to train the Kalman-gain coefficients. Kalman corrections are then applied to data for the subsequent five days (11-15 August 2004). Also, the filter algorithm is run on data for each hour of the day, using only values from previous days at the same hour of the day (corresponding to a $\Delta t = 24$ hours time delay in Figure 5.1).

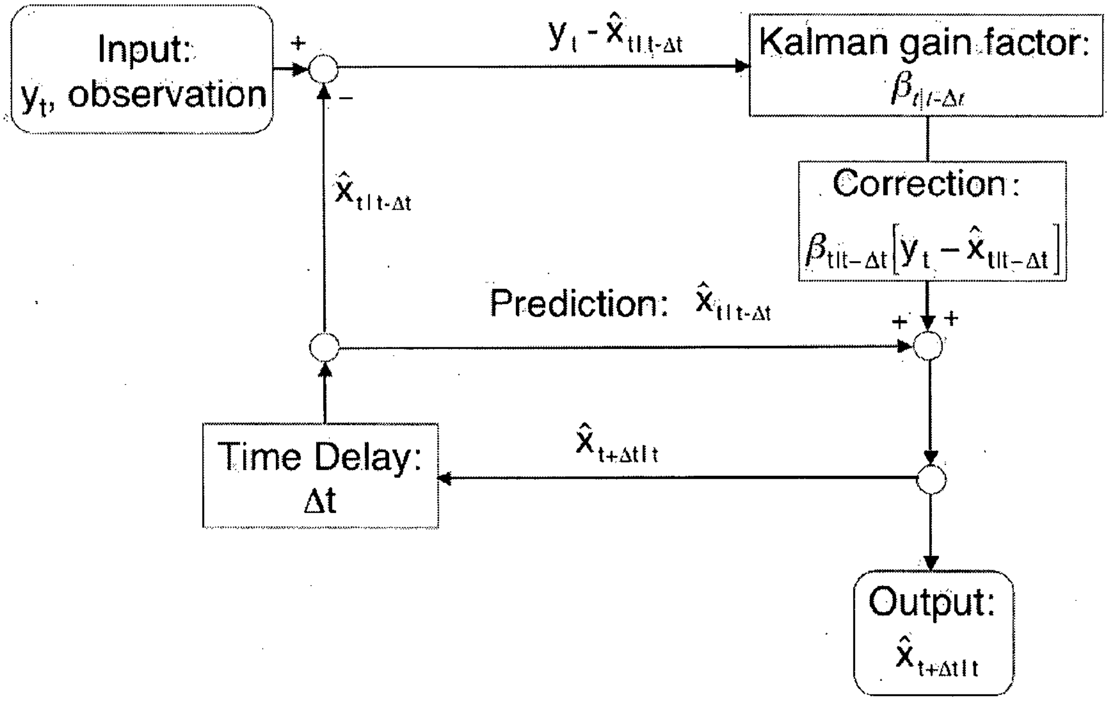

Figure 5.1: Flow diagram of the Kalman-filter bias estimator. It uses a predictor-corrector approach, starting with the previous estimate of the bias ($\hat{x}_{t|t-\Delta t}$) and correcting it by a fraction (β) of difference between the previous bias estimate and previous observed forecast error (y_t) to estimate the future bias ($\hat{x}_{t+\Delta t|t}$).

In this way, a given hour is corrected using only the past forecasts and observations at that same hour. This is to take into account the diurnally-varying behavior the bias may have at different times of the day (e.g., different ozone reactions during daytime versus nighttime). Thus, we compute and save different Kalman coefficients and variances for each hour of the day.

When observations are missing for an hour, the filter uses the last known bias for that same hour from an earlier day. In some cases, however, the true bias changes considerably in such a time period, causing the algorithm to use incorrect, old values. This creates spikes in the Kalman coefficients that can be smoothed by applying the following low-pass filter twice:

$$x_t = \frac{1}{2}\hat{x}_t + \frac{1}{4}[\hat{x}_{t-1} + \hat{x}_{t+1}] \qquad (5.9)$$

Since the bias correction is additive, the Kalman-filtered ozone concentrations were forced to a lower bound of 0 ppbv, in order to avoid negative forecast values.

The Kalman-filter predictor-corrector approach is:

- linear
- adaptive
- recursive
- optimal

Namely, it <u>predicts</u> the future bias as equal to the old bias, but corrected by a <u>linear</u> function of the difference between the previous prediction and the verifying bias. Contrast this to a neural-network approach, which is non-linear (e.g., Cannon and Lord, 2000). Contrary to a neural-network approach that requires a long training period and then behaves in a static manner, the KF approach adapts its coefficients during each time step. Advantages are a much shorter training period, and an ability to adapt to changing synoptic conditions, changing seasons, and even changing weather-forecast models or AQ models. A disadvantage is that it is less likely to predict extreme bias events; namely, it is unable to anticipate a large bias when all biases for the past few days have been smaller.

It is recursive because values of the KF coefficients at any one time step depend on the values at the previous time step. It is optimal in a least-square sense, since it minimizes the expected mean-square error. Finally it is easy to implement and fast running on the computer, requiring storage of a handful of the KF coefficients for each AQ site for each forecast hour.

5.3 Method

5.3.1 Experiments

Because each AQ ensemble member is a forecast based on a different meteorological model, different grid resolution, different emissions, or different initial specie concentrations, it is anticipated that each forecast will have a different bias. Some of these biases could be quite large. Also, this bias could vary depending on the hour of the day. To correct the individual AQ forecasts, we apply a separate Kalman filter for each ensemble member, for each hour. Individual Kalman-corrected AQ forecasts are denoted by K.

Next, if we ensemble (E) average all of the Kalman-corrected (K) forecasts for any hour, then the result is denoted by EK. This ensemble average could have a small residual bias, because the bias corrections that were applied to the individual members were only estimates of future biases (as is the case for true AQ forecasts, not for ex-post-facto calculations of actual biases). Hence, as a final fine-tuning, one can Kalman filter (K) the ensemble average (EK), with the result denoted by KEK.

Experiments are performed here for the same suite of case-study days, NWP models, and initial concentrations, as are described in Chapter 3, but this study tests and compares the performance of the raw, K, EK, and KEK forecasts. During the 5-day period of 11-15 August 2004 used in this case study, there were typical conditions that lead to high ground-level ozone concentrations in the LVF. Those conditions are associated with a northward progressing low-level thermal trough from Washington State, associated with a stationary upper-level ridge situated across southern British Columbia, as described by McKendry (1994).

The five AQ measurements sites for this study are in the complex terrain of the LFV, which is widest at its west terminus at the Georgia Strait. In the LFV sea-breeze circulations, valley and slope flows exist, and with the addition of the photochemistry, ozone modeling becomes quite challenging in this area (McKendry and Ludgren, 2000).

Roughly two million people in greater Vancouver live in this valley, causing significant anthropogenic emissions of NO_x that can mix with the volatile organic emissions from both anthropogenic sources and the surroundings evergreen forest. The Vancouver International Airport (CYVR) ozone monitoring site is at this western edge. The north and south walls of the valley are the steep Coast Range and Cascade Mountains. The valley width decreases considerably toward east, where the ozone site at the town of Hope is located in a very narrow, deep valley. See Chapter 3 for a map and site details. KF post-processing is particularly valuable at complex locations such as these, where both the NWP model and the AQ model can have difficulty.

5.3.2 Verification Statistics

The skill of the 14 forecasts (12 ensemble members plus EK and KEK) have been measured using the same statistical parameters as defined in Chapter 3:

- Pearson product-moment coefficient of linear correlation (herein "correlation"):

$$corr(station) = \frac{\sum_{t=1}^{N_{hour}} [C_o(t, station) - \overline{C_o(t, station)}][C_p(t, station) - \overline{C_p(t, station)}]}{\sqrt{\sum_{t=1}^{N_{hour}} [C_o(t, station) - \overline{C_o(t, station)}]^2 \sum_{t=1}^{N_{hour}} [C_p(t, station) - \overline{C_p(t, station)}]^2}} \quad (5.10)$$

- gross error (for hourly observed values of $O_3 > 30$ ppbv):

$$gross\ error(station) = \frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} \frac{|C_p(t, station) - C_o(t, station)|}{C_o(t, station)} \quad (5.11)$$

- root mean square error (RMSE):

$$RMSE(station) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C_p(t, station) - C_o(t, station)]^2} \quad (5.12)$$

- unpaired peak prediction accuracy (UPPA):

$$UPPA = \frac{1}{N_{day}} \sum_{day=1}^{N_{day}} \frac{|C_p(day, station)_{max} - C_o(day, station)_{max}|}{C_o(day, station)_{max}} \quad (5.13)$$

where N_{hour} is the number of 1-h average concentrations over the 5-day period, N_{day} is the number of days, $C_o(t, station)$ is the 1-h average observed concentration at a monitoring station for hour t, $C_p(t, station)$ is the 1-h average predicted concentration at a monitoring station for hour t, $\overline{C_o(t, station)}$ is the average of 1-h average observed concentrations at a monitoring station over the 5-day period, $\overline{C_p(t, station)}$ is the average of 1-h average predicted concentrations at a monitoring station over the 5-day period, $C_o(day, station)_{max}$ is the maximum 1-h average observed concentration at a monitoring station over one day, and $C_p(day, station)_{max}$ is the maximum 1-h average predicted concentration at a monitoring station over one day. Predicted values also include EK and KEK.

The gross error and UPPA are included in the U.S. EPA guidelines (EPA, 1991) to analyze historical ozone episodes using photochemical grid models. The EPA acceptable performance upper-limit values are + 35 % for gross error, and ± 20 % for unpaired peak prediction accuracy. UPPA is computed here as an average (over the five days available) of the absolute value of the normalized difference between the predicted and observed maximum at each station (Equation (5.13)). Thus, UPPA is non-negative; hence, only the + 20 % acceptance performance upper limit is used in the next sections.

The reasons for utilizing this set of statistics are as follows. We choose correlation to get an indirect indication of the phase differences between the predicted and measured ozone time series at a specific location. The closer the correlation is to one, the better is the correspondence of timing of ozone maxima and minima between the two signals.

RMSE (measured in ppbv) gives important information about the skill in predicting the magnitude of ozone concentration, even though alone it does not draw a complete picture of a forecast value. It is very useful also for understanding the filter behavior, because it can be decomposed into systematic and unsystematic components as discussed in detail in Section 5.4.3.

The gross-error statistic has been considered in this analysis because it is included in the U.S. EPA guidelines (EPA, 1991). Also, being computed for hourly observed values of $O_3 >$

30 ppbv, it gives useful information about the forecast skill for higher concentration values, which are important for health-related issues. It gives information about the error magnitude (as RMSE), but as a portion of the observed ozone concentration (i.e., is measured in %).

UPPA (%) is also used because it measures the ability of the forecasts to predict the ozone peak maximum on a given day. In the past, peak concentrations have been the main concern for the public health. However, in recent years over midlatitudes of the Northern Hemisphere, a rising trend for background ozone concentrations has been observed, while peak values are steadily decreasing (Vingarzan, 2004).

5.4 Results

Figure 5.2 shows a typical example of the KFP bias-correction behavior. In the top panel, the time series include the observations (circles), the ensemble-mean of the raw forecasts (continuous black line), EK (black dashed line), and KEK (black dotted line), for the 7-day period of 09-15 August 2004, at Abbotsford. The first two days on the left side of the vertical dashed line represent the training period, when the coefficients start to be computed, but no correction is applied to the forecast.

Even though the CMAQ model has been spun-up the four days before the start of training (i.e., in the period 05-08 August 2004), the poor first day (August 09) prediction suggests that the forecast did not yet recovered from the cold start initialization. Therefore, a longer CMAQ spin-up period would improve the filter performance as well.

Nevertheless, KFP preserves the good performance of the raw ensemble-mean for the peak concentration, except for the first day. The underestimated peak the first day is not adequately corrected by the KFP because the bias was much smaller for the previous training day. The overnight over prediction (that is indeed common to all the forecasts and the raw ensemble-mean) is improved, with KEK closer to the observations than EK.

The bottom panel of Figure 5.2 shows the behavior of the Fractional Relative Improvement (FRI), defined as follows:

$$FRI = \frac{|RawFcsts - KEK|}{|RawFcsts - Obs|} \tag{5.14}$$

where $RawFcsts$ is the ensemble-mean of the raw forecasts, and Obs is the observation. FRI is computed in Figure 5.2 at 4:00 am (PDT), each day, when the nighttime over prediction is more evident. The fact that FRI, after the training period, almost steadily increases towards its optimal value (FRI $= 1$; i.e., when KEK $= Obs$) it means that the filter, day after day, keeps learning about the over prediction at that hour, and progressively improves its performance. This also confirms what was said in Section 5.2, that the filter quickly and optimally converges after few time-step iterations. It also means that, with a slightly longer training period, the results presented here could be improved, particularly for statistical parameters such as gross error and RMSE.

FRI is not shown here for daytime because the forecasts are already good then. The following subsections present and discuss the results by looking at correlation, gross error, RMSE, and UPPA.

5.4.1 Correlation

Figure 5.3 shows the correlation results for the KFP bias-corrected 12 ensemble members and the ensemble-mean for the 5-day period of 11-15 August 2004, at the five stations (CYVR,

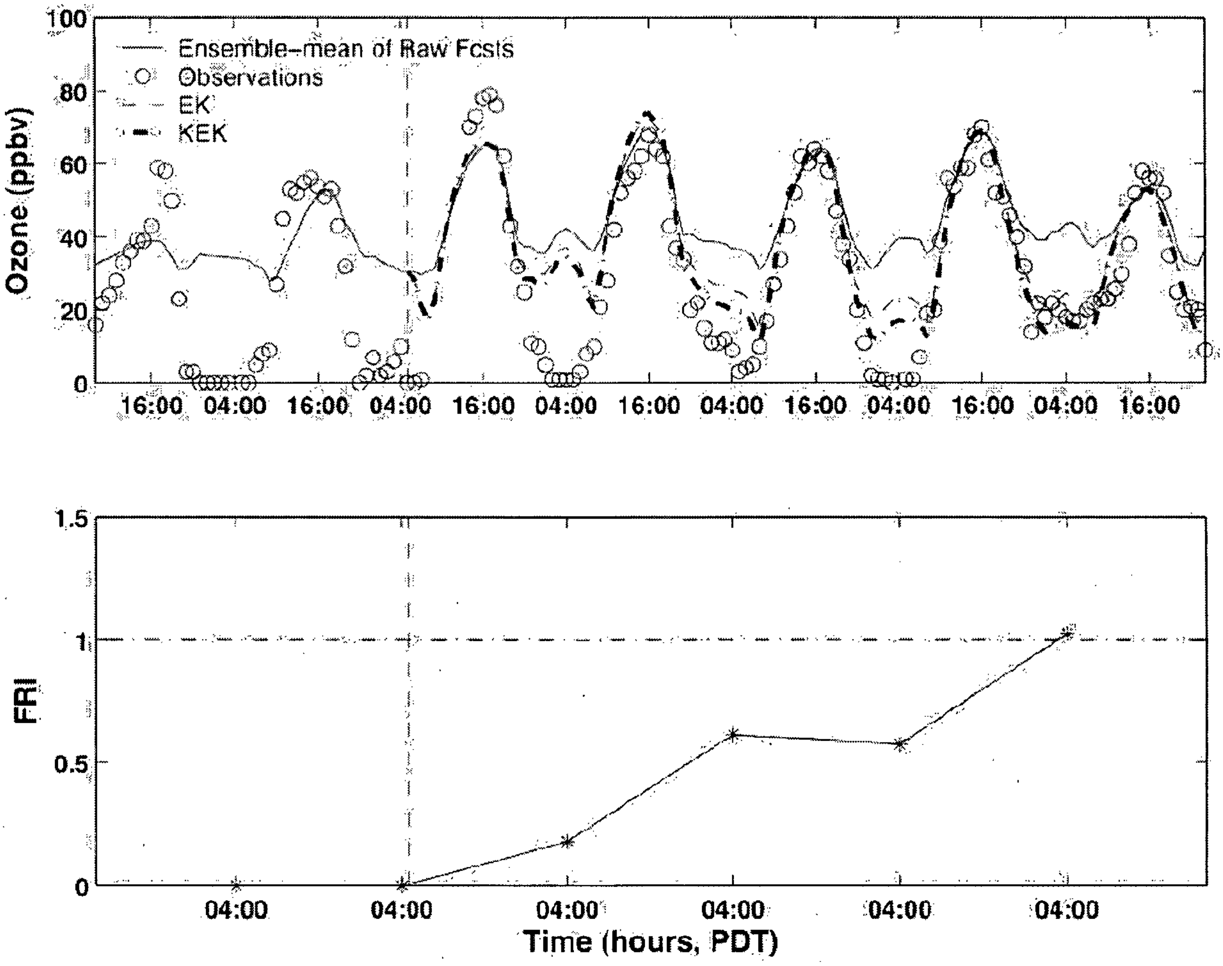

Figure 5.2: Ozone ensemble-mean forecasts and observations at Abbotsford, for the 7-day period 09-15 August 2004 are shown. Top panel: the continuous line is the raw ensemble-mean, the dashed line represents the ensemble-mean of the KFP bias-corrected forecasts (EK), and the dashed-dotted line represents the KFP bias-corrected EK (KEK). The circles are the observations. The vertical dashed lined separates the training period (two days, left) from the filter application (five days, right). Bottom panel: Fractional Relative Improvement (FRI) at 4:00 am for each day. Vertical dashed line as in the top panel, and the dashed-dotted line represents the optimal FRI value (one). Local Pacific Daylight Time (PDT) is UTC - 7 h.

Langley, Abbotsford, Chilliwack and Hope). The black bars are the values for the raw forecasts and raw ensemble-mean (as in Figure 3.6), the grey bars are the values for the KFP bias-corrected forecasts and EK, while the white bars in the last column represent the KEK correlation values. There are improvements (higher correlation between forecast and observations) in most of the cases, except at CYVR where forecasts 10, 11 and 12 (MM5, 4 km) have slightly lower correlation after the KF. The EK improvements are up to a factor of six and they are larger for correlation values below 0.5. At Hope, six ensemble members have negative correlation before the KF bias-correction, but have positive correlation (with values between 0.3 and 0.5) after the correction.

The EK correlation is slightly lower than the raw ensemble-mean at CYVR, slightly higher at Abbotsford and Langley, better at Chilliwack, and significantly improved at Hope. The KEK correlation values are slightly lower than the EK values at CYVR and Abbotsford (but still very high correlation there), while they are higher at the other stations. Notably, after the KFP bias-correction, the differences between the correlation values of the forecasts are lower, meaning that the filter brings all of them closer to the same point – the observations.

Table 5.1 shows for each station the ranking (from 1 to 14) of each ensemble member, EK, and KEK, where the highest correlation value has a ranking of 1, and the lowest has 14. Forecast 08 has similar rankings when compared to EK, while forecasts 08 and 09 (MC2, 4 km) have a slightly worse performance. KEK rankings are the best when compared to any other forecast.

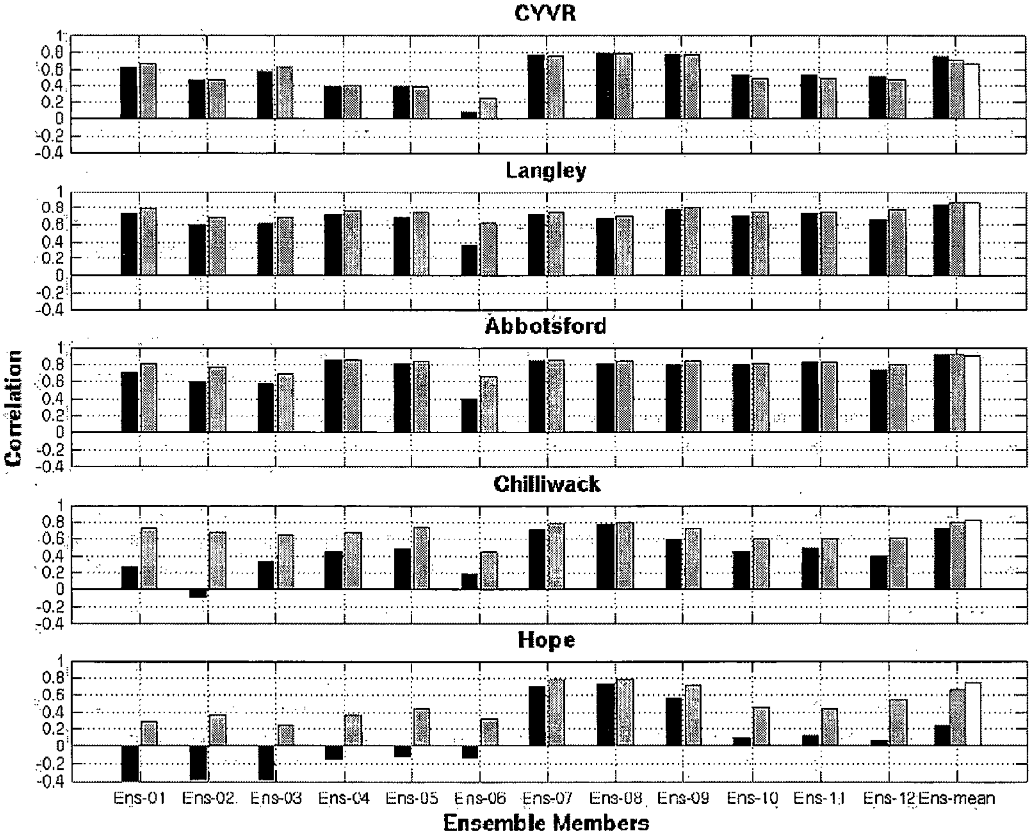

Figure 5.3: Correlation values between observed and predicted ozone 1-h average concentrations are plotted for the 12-member Ozone Ensemble Forecast System (01, 02, ···, 12) and the ensemble-mean (E-mean). The black bars are the values for the raw forecasts and raw ensemble-mean, the grey bars are the values for the Kalman filter predictor (KFP) bias-corrected forecasts and their ensemble-mean (EK), and the white bar represents the KFP bias-corrected ensemble of the KFP members (KEK). Results are plotted at five stations [Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope], for the 5-day period 11-15 August 2004. Values are within the interval [−1, 1], with correlation = 1 being the best possible value.

Table 5.1: Ranking for correlation of KFP bias-corrected 12 ensemble members (01, 02, ···, 12), the ensemble-mean of the KFP bias-corrected forecasts (EK), and the KFP bias-corrected EK (KEK) at the Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack and Hope stations.

	01	02	03	04	05	06	07	08	09	10	11	12	EK	KEK
CYVR	6	11	7	12	13	14	3	1	2	9	8	10	4	5
Langley	4	12	13	6	10	14	9	11	3	7	8	5	2	1
Abbotsford	9	12	13	4	6	14	3	5	7	10	8	11	1	2
Chilliwack	6	9	10	8	5	14	4	2	7	13	12	11	3	1
Hope	13	10	14	11	8	12	2	1	4	7	9	6	5	3

Table 5.2: Ranking for gross error of the KFP bias-corrected 12 ensemble members (01, 02, ···, 12), the ensemble-mean of the KFP bias-corrected forecasts (EK), and the KFP bias-corrected EK (KEK) at the Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack and Hope stations.

	01	02	03	04	05	06	07	08	09	10	11	12	EK	KEK
CYVR	1	9	2	6	10	4	14	13	12	8	11	3	7	5
Langley	4	10	11	6	5	8	14	13	12	7	9	3	1	2
Abbotsford	4	6	12	3	5	11	13	14	10	7	9	8	2	1
Chilliwack	10	7	2	5	8	13	12	14	11	6	9	4	3	1
Hope	12	13	10	5	8	14	3	7	11	6	4	9	2	1

5.4.2 Gross Error

The KFP bias-corrected forecasts have better (lower) gross-error values than the raw forecasts, except at CYVR for forecasts 01 and 06 (Figure 5.4), with improvements roughly between 10 and 20 %. KEK is always better than EK, which in turn is always better than the raw ensemble-mean. The gross-error computation (Equation (5.11)) has a lower ozone concentration limit (observed 30 ppbv). Those improved gross-error values after the KF correction means that the KFP bias-correction is improving not only the forecast nighttime over-prediction, but also efficiently remove bias throughout the time series, regardless of the time of the day.

Table 5.2 summarizes the rankings computed by looking at the gross error. KEK is clearly the best, while EK is the best when compared to the single deterministic forecasts. Here, as

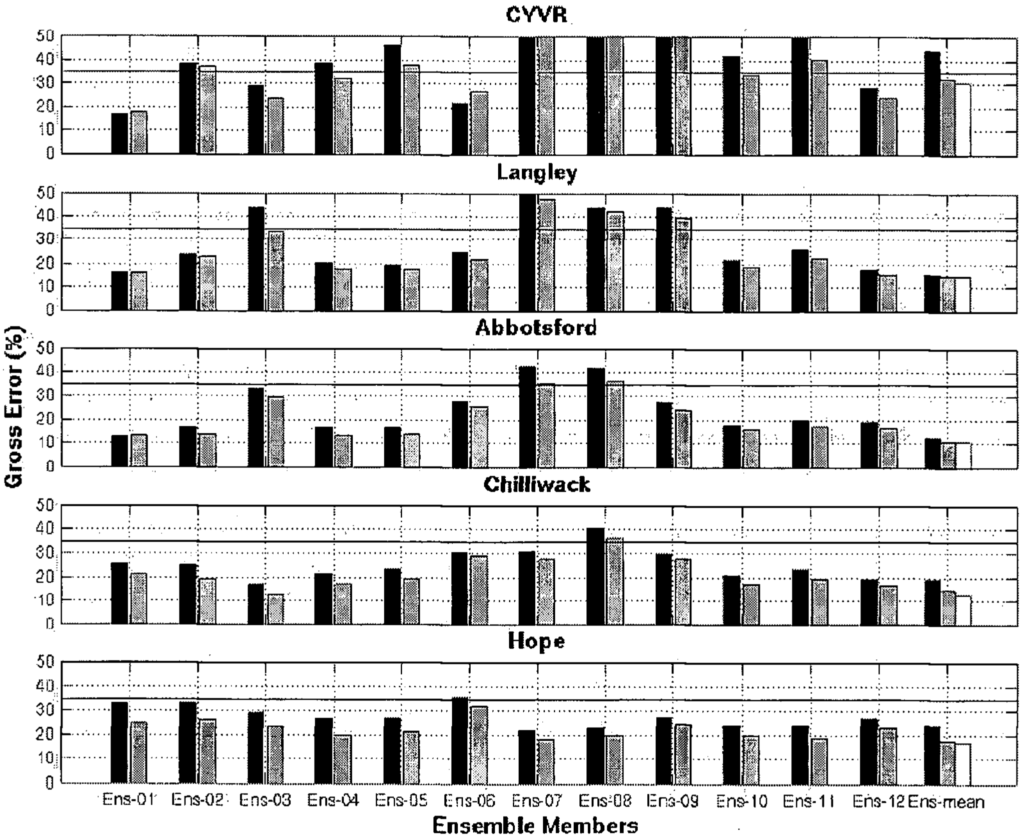

Figure 5.4: Similar to Figure 5.3, but for gross-error values (%). The continuous line is the EPA acceptance value (+ 35 %). Values are within the interval [0, + ∞], with a perfect forecast having gross error = 0.

Table 5.3: Ranking for root mean square error of the KFP bias-corrected 12 ensemble members (01, 02, $\cdots$, 12), the ensemble-mean of the KFP bias-corrected forecasts (EK), and the KFP bias-corrected EK (KEK) at the Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack and Hope stations.

	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**EK**	**KEK**
CYVR	2	7	1	9	11	4	14	12	13	8	10	6	5	3
Langley	2	11	6	8	9	4	14	12	13	7	10	5	3	1
Abbotsford	4	11	7	6	5	8	13	14	12	9	10	3	2	1
Chilliwack	14	6	9	7	2	10	8	13	4	11	12	5	3	1
Hope	12	7	13	14	9	10	3	4	2	8	11	6	5	1

well as for the correlation (Table 5.1), the KFP forecast shows the same problem as the raw ones at CYVR, but not at Hope. The overall poor skill of the raw forecasts at CYVR and Hope are due to the fact that both stations are located in areas where all the individual ensemble members have difficulties, as explained in Section 3.4.2. The KFP is able to considerably improve the raw ensemble-mean at Hope (where it was 4^{th}), with EK being 2^{nd} and KEK 1^{st}. Moreover, both EK and KEK gross error are always well within the EPA acceptance limit (+ 35 %).

5.4.3 RMSE

The RMSE results are shown in Figure 5.5. With this parameter there is an improvement after the KFP bias-correction for all the forecasts, with values improved (decreased) up to 20-25 %. The raw ensemble-mean RMSE is considerably improved at each location, with further improvements (decreases) between 17 and 21 % with EK, and between 29 and 36 % with KEK. Table 5.3 shows the RMSE rankings. KEK is always the best except at CYVR where it is 3^{rd}. EK is 3^{rd} at Langley and Chilliwack, and second at Abbotsford, therefore it is the second best forecast when compared with the other 13.

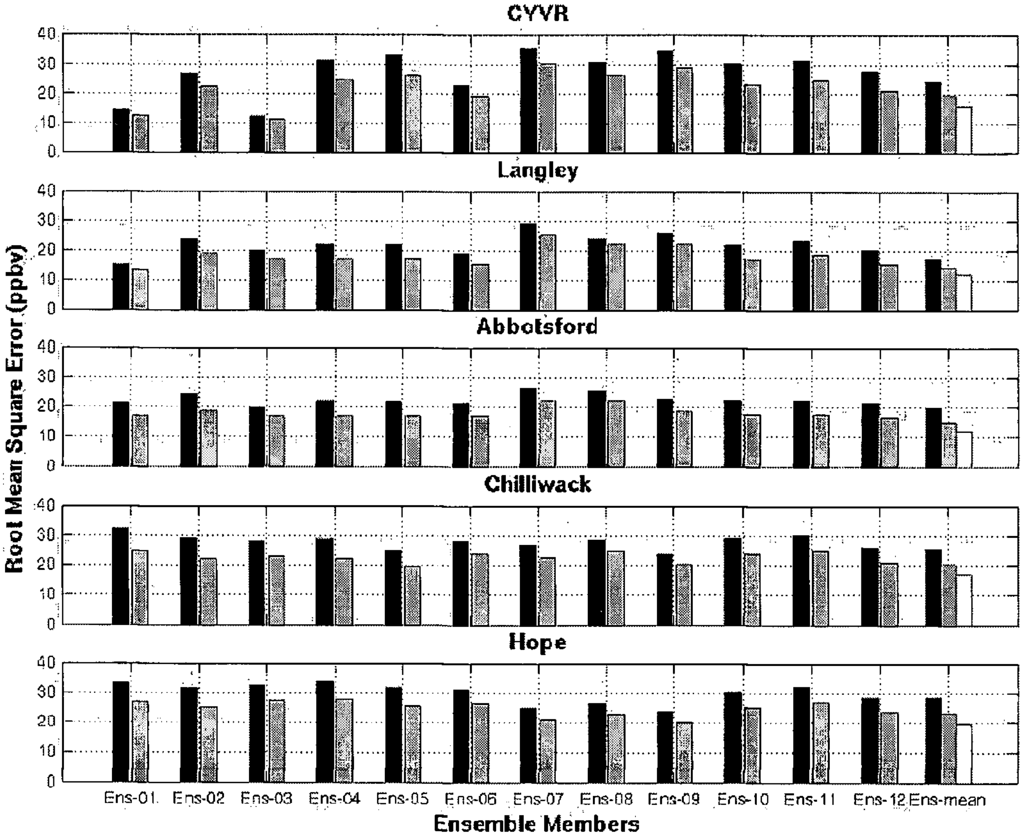

Figure 5.5: Similar to Figure 5.3, but for root mean square error (RMSE) values (ppbv). Values are within the interval [0, + ∞], with a perfect forecast when RMSE = 0.

RMSE can be separated into different components. One decomposition was proposed by Willmott (1981). First, an estimate of concentration $C^*(t, station)$ is defined as follows:

$$C^*(t, station) = a + bC_o(t, station) \tag{5.15}$$

where a and b are the least-square regression coefficients of $C_p(t, station)$ and $C_o(t, station)$ (the predicted and observed ozone concentrations, respectively, as defined in Section 5.3.2). Then the following two quantities can be defined:

$$RMSE_s(station) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C^*(t, station) - C_o(t, station)]^2} \tag{5.16}$$

$$RMSE_u(station) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C^*(t, station) - C_p(t, station)]^2} \tag{5.17}$$

where $RMSE_s(station)$ is the RMSE systematic component, while $RMSE_u(station)$ is the unsystematic one. $RMSE_s$ indicates the portion of error that depends on errors in the model, while $RMSE_u$ depends on random errors, on errors resulting by a model skill deficiency in predicting a specific situation, and on initial-condition errors. The following is an interesting relationship between RMSE and its components:

$$RMSE^2 = {RMSE_s}^2 + {RMSE_u}^2 \tag{5.18}$$

The KF is expected to correct some of the systematic components of the errors (i.e., $RMSE_s$), while the unsystematic component ($RMSE_u$) on average (over the different forecasts) should be affected little by the filter correction. In fact, if $RMSE_u$ reflects errors introduced by model imperfections and initial-condition errors, then it cannot be removed except by fundamental model improvements or improvements in initial conditions.

Figure 5.6 shows the results for $RMSE_s$. The filter is correcting some of the forecast systematic errors, as expected, meaning that the algorithm is properly designed. There is an improvement even when the filter is applied twice (with KEK), meaning that successive applications of the filter correction will decrease further the systematic errors of all the forecasts. The 12-km runs (forecasts 01-06) have their highest systematic error at Hope. All these model runs poorly reproduce the real topography effects at this location, and this lead to systematic misrepresentations of ozone temporal and spatial distribution. Conversely, the 4-km runs have their highest systematic error at CYVR (in particular for MC2 driven runs, forecasts 07-09), where their ability to capture complex terrain more accurately than the 12-km runs is not an advantage, since at CYVR the terrain is flat.

The results for $RMSE_u$ are shown in Figure 5.7. The filter does not decrease the unsystematic errors, and often increases them for this AQ episode. CYVR shows among the highest $RMSE_u$ values (particularly for MC2 driven runs, forecasts 01-03 and 07-09), indicating an intrinsic lack of predictive skill at this location. Martilli and Steyn (2004) discuss the effects of the superimposed valley, slope, and thermal flows over the LFV. Often the pollution plume is transported during night over the Georgia Strait waters, as a result of the combination of several transport processes. This makes it very challenging for the models to accurately predict the spatial and temporal evolution of ozone concentration near water locations, such as CYVR, where the over-strait pool of pollutants can be re-advected over land during daytime sea breeze.

For the ensemble mean, $RMSE_u$ keeps growing after successive filter applications, the

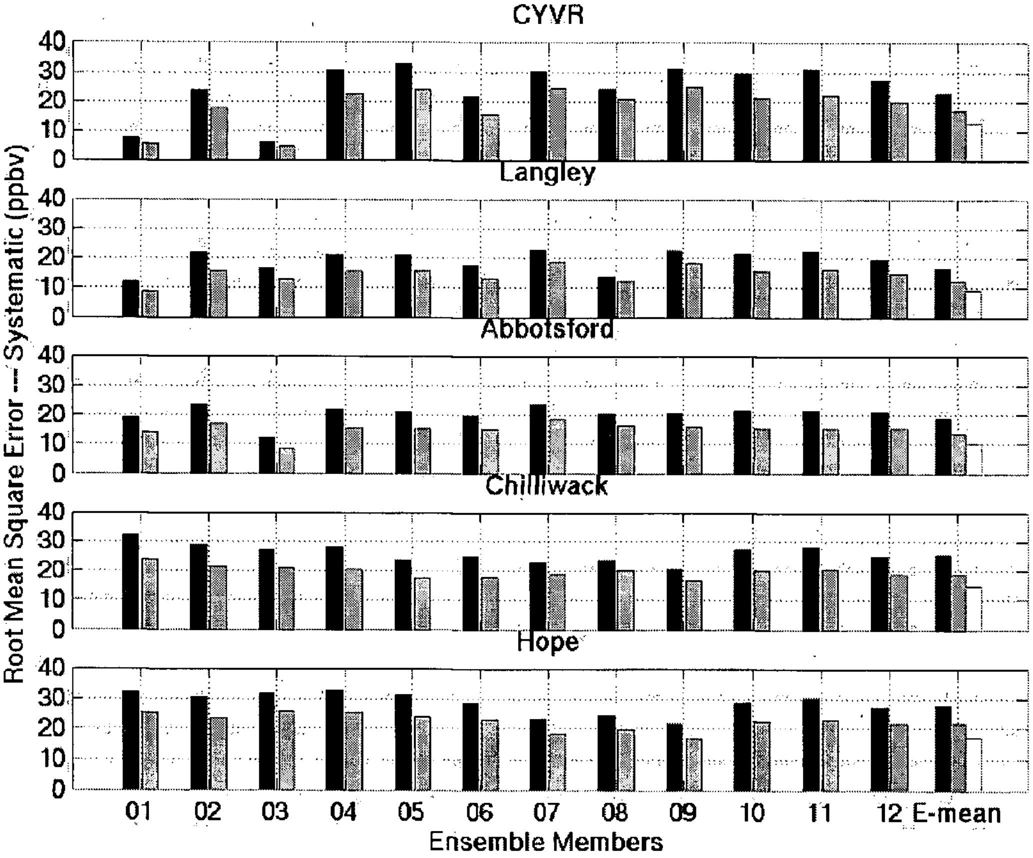

Figure 5.6: Similar to Figure 5.5, but for root mean square error (RMSE) systematic component values (ppbv). Values are within the interval [0, + ∞], with a perfect forecast when RMSE = 0.

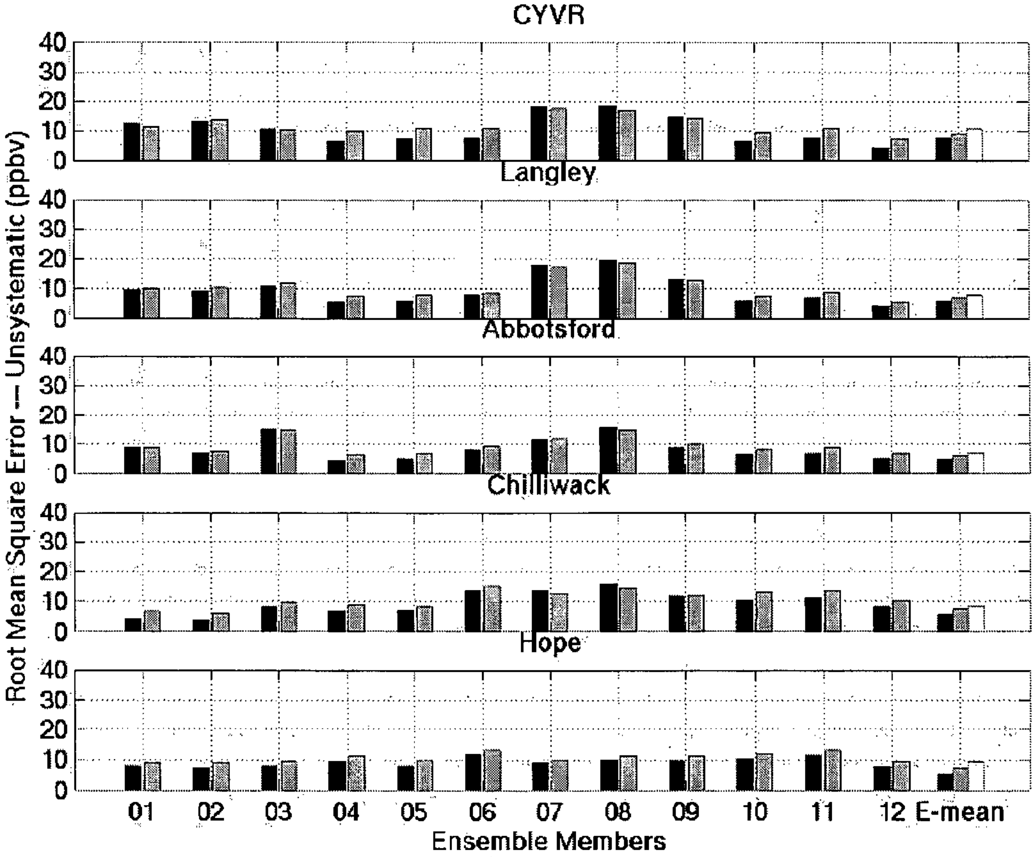

Figure 5.7: Similar to Figure 5.5, but for root mean square error (RMSE) unsystematic component values (ppbv). Values are within the interval [0, + ∞], with a perfect forecast when RMSE = 0.

opposite of what is observed for $RMSE_s$. This means that there is a finite upper limit on the number of useful corrections that can be obtained by successive KF applications. Here, for the ensemble mean, RMSE decreased until the fourth iteration, and grew considerably afterward (not shown).

5.4.4 UPPA

Figure 5.8 shows the results for UPPA. There are improvements (values closer to zero) in the majority of cases; however, in one, three, six, five and three cases out of 14 at CYVR, Langley, Abbotsford, Chilliwack and Hope, respectively, there is no improvement or the KF forecasts are slightly higher. The improvements of the UPPA KFP forecasts with respect to the raw forecasts are modest if compared with the improvements shown with the previous statistical parameters. EK is always better than the raw ensemble-mean, except at Chilliwack, where it is slightly higher. The same can be said for KEK when compared to EK, with the larger improvements for both EK and KEK at Hope. EK and KEK have UPPA values within the EPA acceptance limit (+ 20 %) at Langley, Abbotsford and Chilliwack, while they are close to this limit at Hope and above 30 % at CYVR.

UPPA is the only parameter where the ensemble-mean does not have the best overall ranking, even after the forecasts are KFP bias-corrected. Both EK and KEK have approximately an average performance for UPPA, when compared with the other forecasts (Table 5.4).

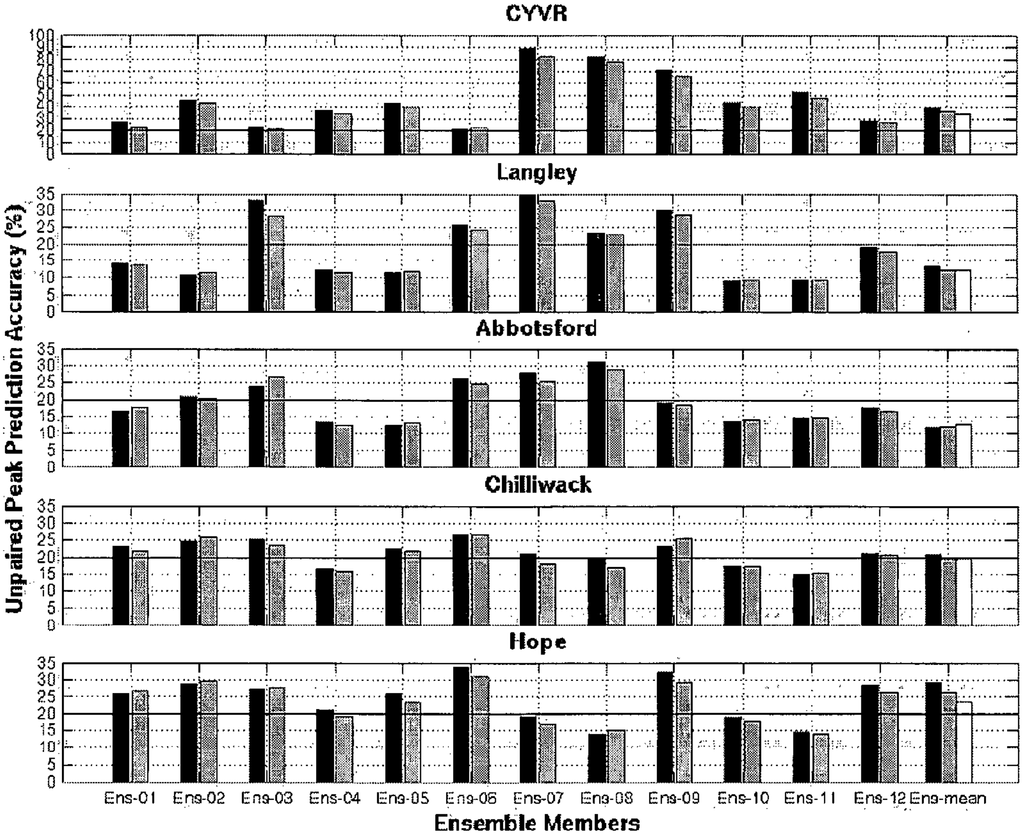

Figure 5.8: Similar to Figure 5.3, but for unpaired peak prediction accuracy (UPPA) values. The continuous lines are the EPA acceptance values (+ 20 %). Values are within the interval [0, + ∞], with a perfect peak forecast when UPPA = 0.

Table 5.4: Ranking for unpaired peak prediction accuracy of KFP bias-corrected 12 ensemble members (01, 02, ···, 12), the ensemble-mean of the KFP bias-corrected forecasts (EK), and the KFP bias-corrected EK (KEK) at the Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack and Hope stations.

	01	02	03	04	05	06	07	08	09	10	11	12	EK	KEK
CYVR	3	10	1	5	9	2	14	13	12	8	11	4	7	6
Langley	8	4	12	3	5	11	14	10	13	2	1	9	7	6
Abbotsford	8	10	13	2	4	11	12	14	9	5	6	7	1	3
Chilliwack	10	13	11	2	9	14	5	3	12	4	1	8	6	7
Hope	10	13	11	5	6	14	3	2	12	4	1	8	9	7

5.5 Comparison with other Bias-correction Methods

Figure 5.9 shows the ensemble-mean RMSE values for the five stations (CYVR, Langley, Abbotsford, Chilliwack and Hope), for the 5-day period 11-15 August 2004. On the abscissa are KEK, EK, the additive bias-correction (AC), the multiplicative bias-correction (MC), and the raw ensemble-mean for comparison purposes.

The additive bias-corrected concentration is computed as follows:

$$C_{AC}(t, station) = C_p(t, station) - \frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C_p(t, station) - C_o(t, station)] \tag{5.19}$$

whereas the multiplicative bias-corrected concentration is given by

$$C_{MC}(t, station) = \frac{\sum_{t=1}^{N_{hour}} C_o(t, station)}{\sum_{t=1}^{N_{hour}} C_p(t, station)} C_p(t, station) \tag{5.20}$$

Both AC and MC use observations throughout the experiment period, so the ozone time series corrected with these methods cannot be considered forecasts, since they cannot be computed in a predictor mode. Contrast this with both KEK and EK that are predictor post-processing procedures of the forecasts, which use only observations available before the time for which the forecast verify. In this sense, this is a stringent test for the KFP bias correction.

Nevertheless, at every station (except CYVR) KEK is the best, while EK in general is better than MC, but has higher (worse) RMSE values than AC (except at Hope). Finally, at CYVR, KEK is third while EK is better only than the raw ensemble-mean.

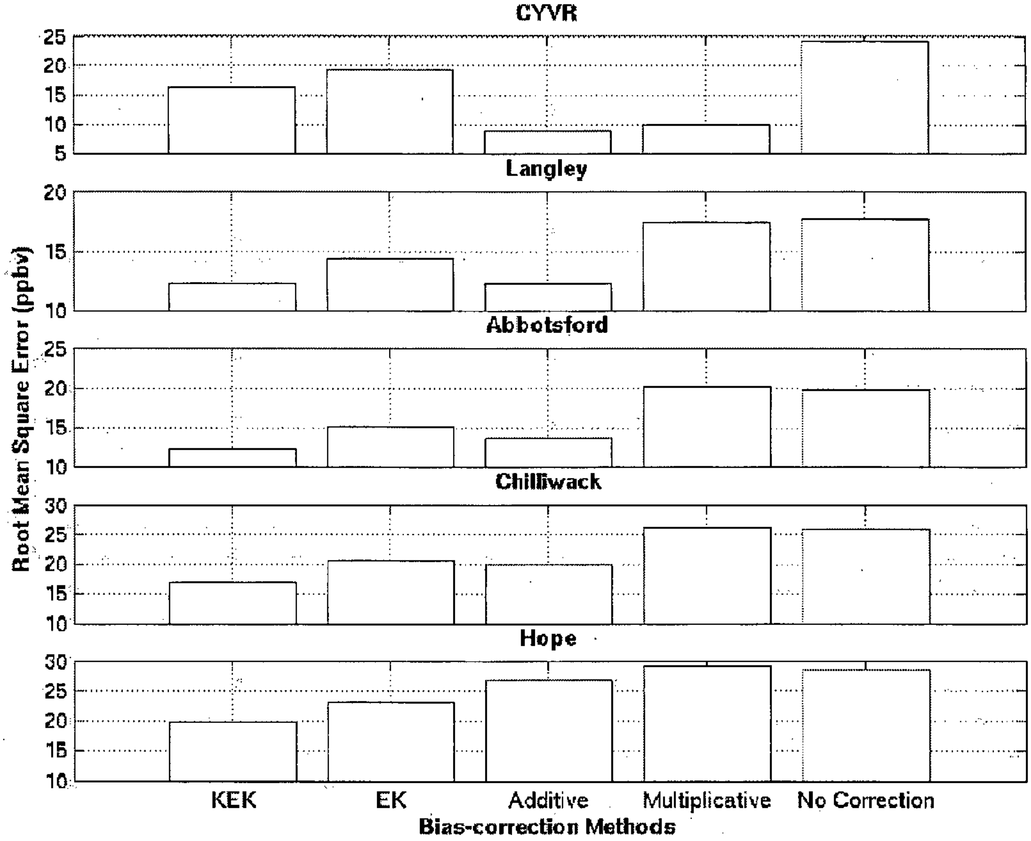

Figure 5.9: Root mean square error (RMSE) values (ppbv) are shown for four different bias-correction methods applied to the ensemble-mean. These methods are: the Kalman filter predictor (KFP) bias-corrected ensemble-mean of the KFP bias-corrected forecasts (KEK), the ensemble-mean of the KFP bias-corrected forecasts (EK), the Additive correction (AC), and the Multiplicative correction (MC). The last values on the abscissa are for the raw ensemble-mean with no corrections. Results are plotted at five stations [Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope], for the 5-day period 11-15 August 2004. Smaller values are better.

5.6 Discussion and Conclusions

In summary, the Kalman-filter predictor (KFP) bias-corrected forecasts and their ensemble-mean have better forecast skill than the raw forecast, for the locations and days used here to test their performance. The corrected forecasts are improved for correlation, gross error, root mean square error (RMSE), and unpaired peak prediction accuracy (UPPA), the latter being the statistical parameter showing the less pronounced improvement after the KFP bias-correction. In general, the ensemble-mean forecast benefits from the improvement of each single Kalman-corrected ensemble member. In fact, the ensemble-mean of the KFP bias-corrected forecasts (EK) and the KFP bias-corrected EK (KEK) are the second best and the best forecasts overall when compared with the other 12 individual forecasts members and their raw ensemble-mean. The results in Section 5.4.3 showed also that only a limited number of successive KF application to the same forecast would result in an improvement.

Those results indicate that the filter improves the forecast timing of maxima and minima concentrations with respect to the observations, because the correlation is closer to one. From the improved (decreased) RMSE and gross-error values, we infer that the KF improves the forecast accuracy in reproducing the magnitude of ozone concentrations. Better (closer to zero) UPPA and gross-error values indicate that the filter improves the forecast ability to capture rare (but important for health-related issues) events, such as the occurrence of ozone concentration peaks. Moreover, the KF reduced systematic errors such as can be induced by model error, as for example the poor representation of topographic complexity. Ensemble averaging tended to remove the unsystematic errors, as showed in Chapter 3. This is why the combination of Kalman filtering and ensemble averaging results in the best forecasts; i.e., EK and KEK.

EK and KEK performances have been compared also with the performances of two other bias-correction (not in predictor mode) techniques, the additive bias-correction (AC), the multiplicative bias-correction (MC). At every station (except CYVR) KEK is the best, while EK is better than MC, but has higher (worse) RMSE values than AC (except at Hope). Finally, at CYVR, KEK is third while EK is better only than the raw ensemble-mean.

A concise way to summarize the results from Section 5.4 is given in Figures 5.10, 5.11, 5.12, 5.13, and 5.14. A Taylor's diagram (Taylor, 2001) is used to create a multi-statistic plot of correlation, centered RMSE (CRMSE: RMSE computed after the overall bias is removed), and standard deviation. CRMSE is the distance on the diagram between the point representing the forecast and the one representing the observations. For each forecast (smaller arrows) and for EK and KEK (bigger arrows, with different arrowhead), the arrow tail gives the standard deviation and the correlation of a raw forecast, while the arrowhead represents the same values for the KFP bias-corrected version of the same forecast. If the arrow points toward to the observation (circle) it means that the KFP is correcting the forecast statistically in the right direction. The arrows representing EK and KEK are consecutive; i.e., the EK arrowhead is also the KEK arrow tail, because EK is the raw version of KEK. The three concentric lines centered over the point representing the observation indicate the CRMSE for the raw ensemble-mean (dotted line), EK (thick dashed line), and KEK (thick continuous line).

At CYVR (Figure 5.10) the majority of arrows point away from the observation (including the arrows with different arrowhead for EK and KEK), indicating that the KFP in those cases degraded the raw forecasts. This is caused by the dominance of unsystematic errors at this location (as discussed in Section 5.4.3), that prevent the filter to being able to do a successful correction.

At Langley (Figure 5.11) the forecasts tend to be improved, as indicated by the arrows pointing closer to the observation. EK is better than the raw-ensemble-mean (which in turn is better than all the individual deterministic forecasts), since the thick dashed line passing through its arrowhead is closer to the observations than the dotted line passing through the tail. KEK is the best being the closest to the observations (thick continuous line).

The same conclusions can be drawn for Abbotsford (Figure 5.12), with even larger improvements after the correction. At this location, the forecast standard deviations after the correction are much more similar to the observation standard deviations (but the same can be said also at the other stations).

Figure 5.13 shows the same diagram for Chilliwack. The forecasts are improved, since the arrows point toward the observations. At this location, EK is fourth best, while KEK is still the best.

The results for Hope are shown in Figure 5.14. All the forecasts are improved, with EK and KEK being the fifth, and third best, respectively. In this case (as well as for Chilliwack) the benefit of applying the KFP bias correction is even higher than at the other locations, demonstrating that the KF correction is particularly efficient if the raw forecast shows high systematic errors, as discussed in Section 5.4.3. This is evident since the arrows are on average longer than at the other locations. At Hope, forecasts 07 and 08 are the first and second best forecasts (by comparison with Figure 3.19), while they were among the worst at other locations, particularly at CYVR, Langley at Abbotsford.

The KFP bias-correction approach for the locations and days used in this study successfully removes part of the forecast bias. The filter is able to recognize systematic errors in the

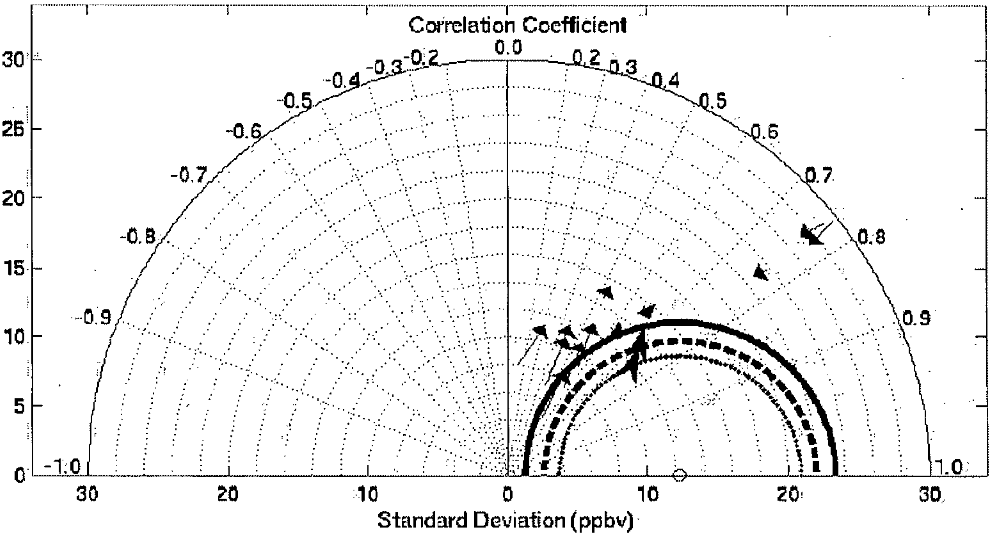

Figure 5.10: Taylor's diagram is plotted for Vancouver International Airport (CYVR). The azimuthal position gives the correlation, while the radial distance from the origin is proportional to the standard deviation (ppbv). The smaller arrows represent the 12 ensemble members, and the bigger arrows (with different arrowhead) represent the ensemble-mean of the Kalman filter predictor (KFP) bias-corrected forecasts (EK) and the KFP bias-corrected EK (KEK). Each arrow tail represents the forecast statistics of a raw forecast, and the arrowhead indicates KFP-corrected values. If the arrow points closer to the observation point (circle) it means that the KFP is correcting the forecast in the right direction. The arrows representing EK and KEK are consecutive; i.e., the EK arrowhead is also the KEK arrow tail, because EK is the raw version of KEK. The distance between the observation and a given point is proportional to the centered root mean square error (CRMSE) between the observation and the forecast. The three concentric lines centered over the point representing the observation indicate the CRMSE for the raw ensemble-mean (dotted line), EK (thick dashed line), and KEK (thick continuous line). If the line passing through the arrowhead is closer to the observation than the one passing through the tail, it means that that the KFP is improving (reducing) the CRMSE.

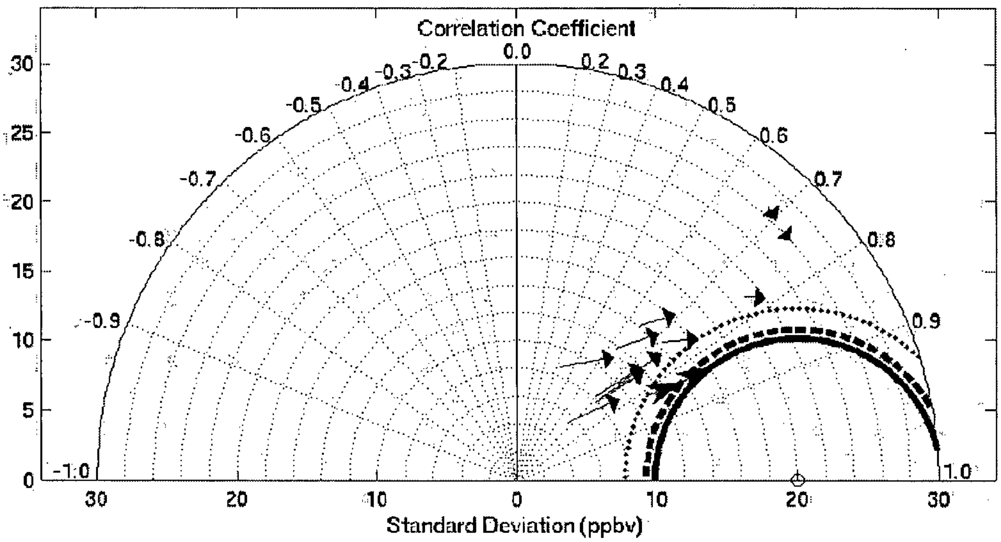

Figure 5.11: Taylor's diagram for Langley (similar to Figure 5.10).

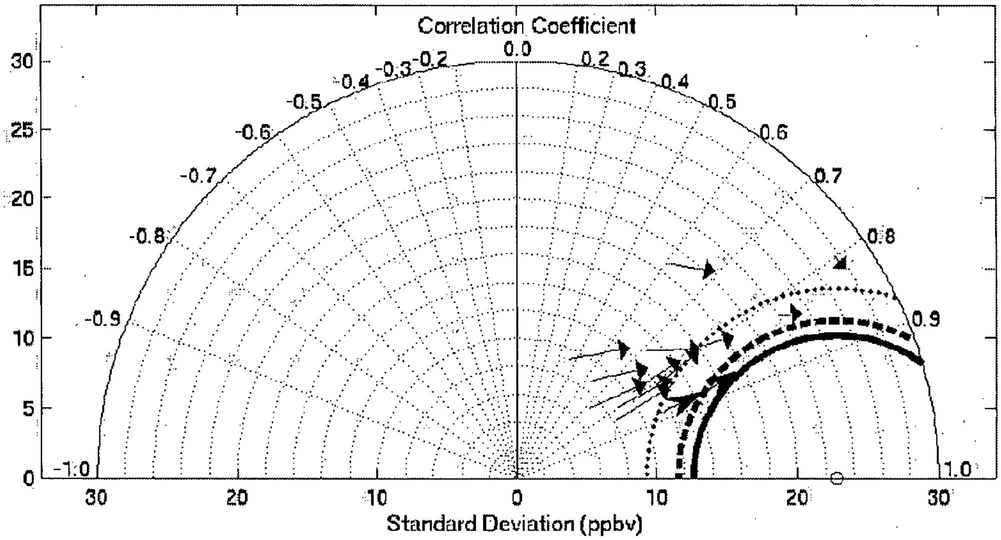

Figure 5.12: Taylor's diagram for Abbotsford (similar to Figure 5.10).

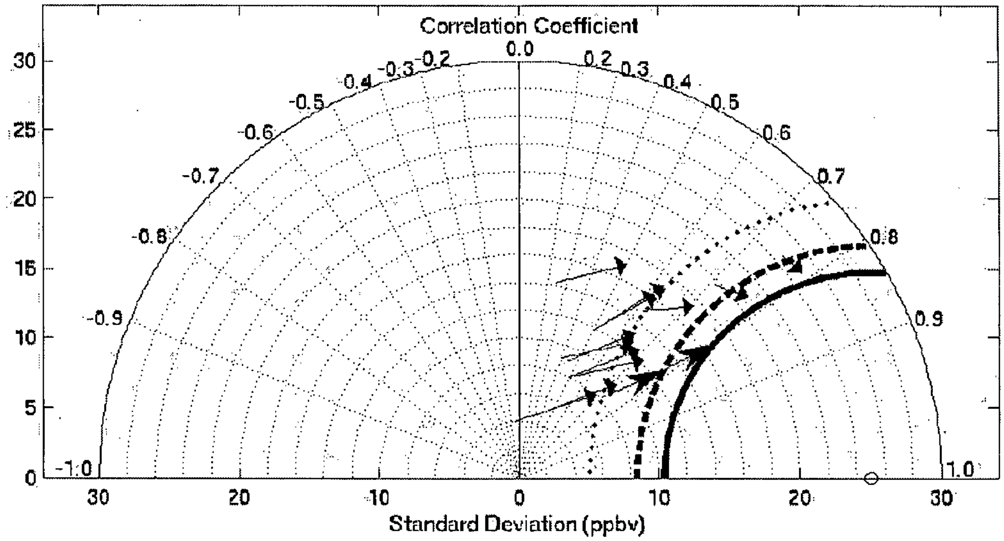

Figure 5.13: Taylor's diagram for Chilliwack (similar to Figure 5.10).

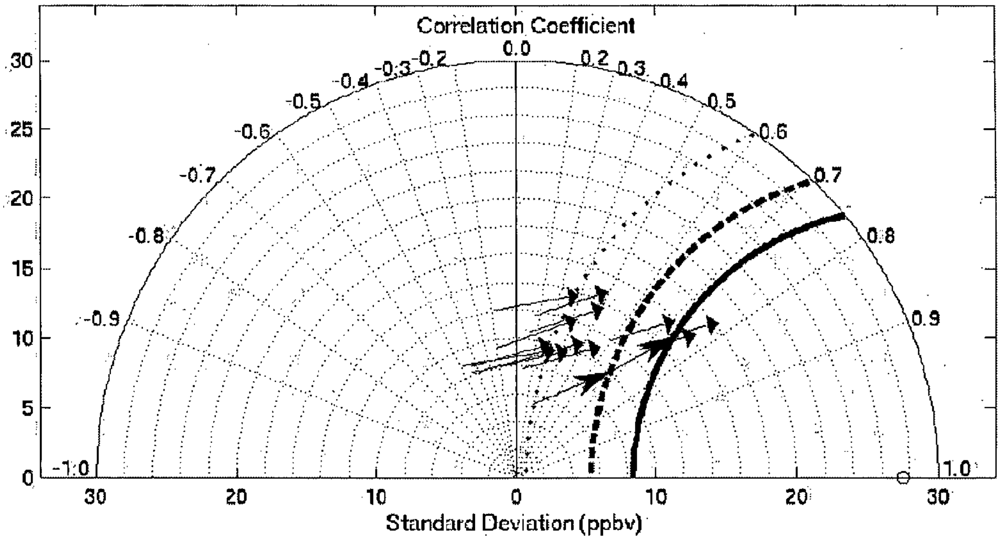

Figure 5.14: Taylor's diagram for Hope (similar to Figure 5.10).

forecast, as for example the nighttime over-prediction of ozone concentration induced by a poor representation of the nighttime boundary layer, or the errors at Chilliwack and Hope induced by the systematic misrepresentation of topographic complexity in the model. As a consequence of the improved nighttime over prediction, the ozone distribution low-concentration tail is better represented after the KF correction, resulting in forecasts having a variance that resembles more closely the observation variance, as discussed above.

The experiments performed in this study suggest that better forecasts can be made with a longer KF training period (such as 5 days), and with a longer CMAQ model spin-up. Moreover, with the availability of a longer data set (a full month or season), including ozone forecasts and observations with a broader variability of low and high ozone events, an optimal value for the sigma ratio (as discussed in Section 5.2) could be found.

KEK, which combines the beneficial effects of ensemble averaging and KFP post-processing, is overall the most skillful forecast for the locations and days tested here, where the ozone modeling is particular challenging because the complex coastal mountain setting. For this reason the approach used here to improve ozone forecasts might be equally successful when implemented in other regions with similar or less complex topographical settings.

Finally, ensemble weather forecasts often provide information on the reliability of the forecast: if the ensemble members have a large spread (defined as the standard deviation of the ensemble members about the ensemble mean), this implies less confidence in the forecast. Perhaps a similar spread-skill relationship exists for air-quality forecasts. However, in Chapter 3, neither a correlation nor a relationship between the raw ensemble spread and the raw forecast error has been found. Similarly, a spread-skill relationship has not been found for the Kalman-filtered AQ forecasts in this study.

5.7 References for Chapter 5

Bozic, S. M., 1994: *Digital and Kalman Filtering, Second Ed.*. John Wiley & Sons, New York.

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, **126**, 1719–1724.

Byun, D. W. and J. K. S. Ching, 1991: Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) modeling system. Technical Report EPA/600/R-99/030, U.S. Environmental Protection Agency.

Cannon, A. J. and E. R. Lord, 2000: Forecasting summertime surface level ozone concentrations in the Lower Fraser Valley of British Columbia: An ensemble neural network approach. *Journal of the Air and Waste Management Association*, **50**, 322–339.

Dempster, A., N. Laird, and D. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39**, 1–38.

EPA, 1991: Guideline for regulatory application of the urban airshed model. Technical Report EPA-450/4-91-013, U.S. Environmental Protection Agency.

Hamill, T. M. and C. Snyder, 2000: A hybrid ensemble Kalman filter-3D variational analysis scheme. *Monthly Weather Review*, **128**, 2905–2919.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using Kalman filter. *Weather and Forecasting*, **10**, 987–707.

Houtekamer, P. L. and H. L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **129**, 123–137.

Houtekamer, P. L., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with an Ensemble Kalman Filter: results with real observations. *Monthly Weather Review*, **133**, 604–620.

Joliffe, I. T. and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley and Sons, West Sussex.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **133**, 35–45.

Martilli, A. and D. G. Steyn, 2004: A numerical study of recirculation processes in the Lower Fraser Valley (British Columbia, Canada). In *27th NATO/CCMS Conference on Air Pollution Modeling*. Banff, Alberta.

McKendry, I. G., 1994: Synoptic circulation and summertime ground-level ozone concentrations at Vancouver, British Columbia. *Journal of Applied Meteorology*, **33**, 627–641.

McKendry, I. G. and J. Ludgren, 2000: Tropospheric layering of ozone in regions of urbanized complex and/or coastal terrain: a review. *Progress in Physical Geography*, **24**, 329–354.

Roeger, C., R. B. Stull, D. McClung, J. Hacker, X. Deng, and H. Modzelewski, 2003: Verification of mesoscale numerical weather forecasts in mountainous terrain for application to avalanche predicition. *Weather and Forecasting*, **18**, 1140–1160.

Russell, A. and R. Dennis, 2000: NARSTO critical review of photochemical models and modeling. *Atmospheric Environment*, **34**, 2283–2324.

Segers, A., H. J. Eskes, R. J. van der A, R. F. van Oss, and P. F. J. van Velthoven, 2005: Assimilation of gome ozone profiles and a global chemistry-transport model using a kalman filter with anisotropic covariance. *Quarterly Journal of Royal Meteorological Society*, **131**, 477–502.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, **106**, 7183–7192.

van Loon, M., P. J. H. Builtjes, and A. J. Segers, 2000: Data assimilation applied to LOTOS: first experiences. *Environmental Modeling Software*, **15**, 603–609.

Vingarzan, R., 2004: A review of surface ozone background levels and trends. *Atmospheric Environment*, **38**, 3431–3442.

Willmott, C. J., 1981: On the validation of models. *Physical Geography*, **2**, 184–194.

Chapter 6

Conclusions

The goal of this research was to improve real-time short-term forecasts of tropospheric ozone measured at near-surface receptor sites. This goal was achieved.

6.1 Summary of Methods and Procedures

This research was based on the hypothesis that the ensemble technique and Kalman-filter postprocessing can be transferred to air-quality modeling, and can potentially yield similar benefits as for NWP. The method used here was 3-D mesoscale NWP modeling coupled with 3-D chemical numerical modeling. The procedure was to run these models using emission inventories for actual ozone episodes, and to calibrate and verify the results against near-surface ozone observations.

This dissertation summarizes the results of an immense amount of numerical computations:

- Nine days run with the 3-D Eulerian NWP model MC2 with four (108, 36, 12, and 4 km horizontal grid spacing) nested grids.

- Nine days run with the 3-D Eulerian NWP model MM5 with four (108, 36, 12, and 4 km horizontal grid spacing) nested grids.

- 36 (two NWP model by nine days by two grids) 1-day runs with the meteorological pre-processor MCIP.

- 56 (four control runs by nine days plus four lagged runs by five days) 1-day runs with the SMOKE emission pre-processor.

- 186 (four spin-up days by four control runs plus five days by 28 forecasts plus five days by six lagged forecasts) 1-day run with the 3-D Eulerian CMAQ model (to perform the 12 AQ forecasts).

All the above resulted in few hundreds Gigabytes of data, and several hundreds of computational hours on processors of a high-performance computing Linux super-cluster.

Chaos theory has been applied through the ensemble approach to improve our ability to predict the spatial and temporal distribution of tropospheric ozone concentration, and to estimate in advance its magnitude. The ensemble approach is one method to represent the time evolution of the probability density function describing the atmosphere's initial state and its uncertainty. The probability density function is represented by a limited set of points. The evolution of each of these points would be a member of the ensemble. Each of these members should ideally represents an equally likely evolution of the dynamical system.

Kalman-filter theory has been applied in this dissertation to remove ozone forecast bias; i.e., systematic errors. The filter was applied as a post-processing procedure, in a predictor mode. Previous bias values were used as input to Kalman filter. Once the future bias has been estimated, it was removed from the raw forecast to produce an improved forecast. Such a corrected forecast should be statistically more accurate in a least-squares sense.

To accomplish this goal, the following research work was conducted:

- The realization and test of an air-quality ensemble built on a previous photochemical model intercomparison study (see Chapter 2). This preliminary work demonstrated the value of ensemble air-quality forecasts, and lead to more-detailed research.

- The realization and test of a new air-quality ensemble design, created by perturbating the input fields that most affect the uncertainty of the air-quality photochemical models; i.e., the meteorological and the emissions fields (see Chapters 3 and 4).

- The realization and test (introducing a new reliability index) of probabilistic forecasts resulting from ensemble methods (see Chapter 4).

- The realization and test of a new method to remove systematic errors from air-quality forecasts, based on the Kalman-filter-predictor algorithm (see Chapter 5).

6.2 Summary of Findings

The findings of this dissertation can be summarized as follows:

- An ensemble average computed from the ozone prediction of different photochemical models is a more skillful ozone forecast than the one from a single deterministic model (Delle Monache and Stull, 2003).

- An average of ensembles created by both meteorology and emission perturbations has better-forecast performance than any individual ozone prediction, being able to filter out unpredictable components of the transport, diffusion, and chemical reactions governing the ozone spatial and temporal distribution evolution (Delle Monache et al., 2005a).

- Twenty-eight forecasts (grouped in 13 different ensembles) have been generated over the Lower Fraser Valley, British Columbia, Canada, for the 5-day period 11-15 August 2004, and they have been compared with 1-hour averaged measurements of ozone concentrations over five stations. The different forecasts are obtained by combining four driving meteorological input fields with seven emission scenarios: a control run, ± 50 % NO_x, ± 50 % VOC, and ± 50 % NO_x combined with VOC (Delle Monache et al., 2005c).
    - Both meteorology and emission perturbations are needed to have a skillful probabilistic forecast system, and neither of two is sufficient alone to form a reliable probabilistic forecast system with a good resolution for the whole spectrum of ozone concentrations.
    - The meteorology perturbation is important to capture the ozone temporal and spatial distribution.
    - The emission perturbation is needed to accurately predict the ozone concentration magnitude. In particular, the emission perturbations are more important than the meteorology ones to capture high (and rarely measured) ozone concentrations, typically observed in the afternoon in areas such as the Lower Fraser Valley where ozone production may be mainly attributed to local sources.
    - Among the emission perturbations, the ones involving NO_x resulted in more skillful probabilistic forecasts for the episode analyzed in this study.
    - The ± 50 % emission perturbations appeared to be not centered over an optimal

estimate, and shifting the perturbations toward lower values could improve the forecasts by reducing the positive bias.

  - Since NO_x has good (but positively biased) predictive skill, the ± 50 % limit seems to efficiently span the emission uncertainties space for this case.
  - The ensemble formed by all the 28 ozone forecasts available is the best probabilistic forecast, when considering both reliability and resolution.
  - The smoothing of peak values caused by ensemble averaging (Delle Monache et al., 2005a) can be avoided if the ensemble-mean ozone peak is computed as the average of the ensemble-member peak predictions (Delle Monache et al., 2005c).
  - The MC2 model has more variability than MM5 in the 5-day period analyzed in this study, and this resulted in the ensemble formed by all the runs driven by MC2 forming the more skillful ensemble-averaged ozone forecast. However, the 12-member ensemble based on meteorology and NO_x perturbations provided the best ensemble-averaged prediction of the magnitude and timing of peak ozone.
  - The root-mean-square-error random component for the ensemble formed with all the runs with 4 km horizontal spatial resolution is higher than the one formed with the 12 km resolution runs.
  - With a hard limit on computational resources, the ensemble mean computed with only the four control runs, where only meteorology is perturbed, has good skill at predicting the magnitude of the ozone peak.

- The Kalman-filter predictor bias-corrected forecasts and their ensemble-mean have better forecast skill than the raw forecasts, for the locations and days used here to test their performance (Delle Monache et al., 2005b). The corrected forecasts are improved for correlation, gross error, root mean square error, and unpaired peak prediction accuracy, the latter being the statistical parameter showing the least pronounced improvement after the Kalman-filter predictor bias correction. Furthermore:

  - The Kalman-filter predictor bias-correction approach successfully removes part of the forecast bias for the locations and days used in this study.
  - Only a limited number of successive Kalman-filter applications to the same forecast would result in an improvement, since while the filter removes systematic errors, it tends to amplify random errors.
  - As a consequence of the raw-forecast nighttime over prediction, the ozone distribution low-concentration tail is better represented after the Kalman filter correction, resulting in forecasts having a variance that resembles more closely the observed variance.

- Ensemble averaging tends to remove the unsystematic errors. Its combination with Kalman filtering (which removes part of the systematic errors) results in the best ozone forecasts.

6.3 Discussion and Recommendations

This dissertation proved the necessity of considering the chaotic behaviour of the atmsphere (associated with the nonlinearity of the chemistry) in any attempt to describe the evolution of such a system. Any deterministic prediction of this evolution would most likely misrepresent the nature of the problem. Ensemble and Kalman-filter methods can indeed significantly improve near-surface ozone forecasts, even in the complex coastal mountain setting of the Lower Fraser Valley. There are no intrinsic limitations to these methods that would prevent their application in real time to other geographic settings.

The results of this dissertation suggest that future ozone-forecast work should focus on ensemble forecast systems involving both meteorology and emission perturbations. More specifically, the above findings suggest that the emission perturbations could be based on the temporal and spatial variability of different regimes. If (during a particular time of the day and in a subset of the spatial domain) a NO_x-sensitive regime is dominant, then a NO_x perturbation would be more useful than a VOC one to capture the ozone variability. Conversely, in VOC-sensitive regimes the VOC perturbations could be more efficient. In situations where neither of these two regimes is well defined, probably a combination of NO_x and VOC perturbations could be the best choice. These regimes could be identified in forecast mode by looking at the control-model forecasts, for example by evaluating the O_3/NO_y or H_2O_2/HNO_3 ratios (Sillman and He, 2002).

One of the findings of this dissertation is to shift the emission perturbations toward lower values (for both NO_x and VOC), to improve the forecasts by reducing their overall positive bias. This correction will improve the forecasts on the west side of the spatial domain considered in this dissertation, while for the eastern-most locations (i.e., Chilliwack and Hope) such a shift will not improve the ozone forecasts, or may deteriorate them slightly.

Ideally, each ensemble member should be an equally likely time evolution and space distribution of the ozone concentration, and they should all be equally good estimates of truth. With this in mind, the ensemble members should be "independent", in the sense that none of them should rely on other members for their realizations. This is not the case when nested grids are used, as for some of the probability forecast systems created in this dissertation. Namely, CMAQ domains are linked using a 1-way nesting approach (similarly for MC2, but MM5 runs are implemented with 2-way nesting), and all the 4 km runs cannot be considered independent of the coarser 12 km runs providing the driving meteorology and/or the initial and boundary chemistry. The dependency among members of the same ensemble (no attempt has been made in this study to measure it) would result in an "effective" ensemble size smaller than the actual ensemble size. Moreover, a subset of the dependent members will span approximately the same subspace of the air-quality modeling uncertainty space (or at least they should be closer to each other than to other members), resulting in both probabilistic and ensemble-averaged forecasts relying too heavily on the performances of these members than on others.

Ensemble weather forecasts often provide information on the reliability of the forecasts; if the ensemble members have a large spread (defined as the standard deviation of the ensemble members about the ensemble mean), this implies less confidence in the forecast. However, in this research no correlation or relationship between ensemble spread and forecast error has been found.

Finally, the methodology developed in this study to improve ozone regional forecasts could be implemented also to improve forecasts of particulate-matter and other pollutants.

6.4 References for Chapter 6

Delle Monache, L., X. Deng, Y. Zhou, and R. B. Stull, 2005a: Ozone ensemble forecasts. Part I: a new ensemble design. *Accepted in November 2005 to be published in the Journal of Geophysical Research.*

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2005c: On probabilistic and ensemble-averaged ozone forecasts. *Manuscript submitted in November 2005 to the Journal of Geophysical Research.*

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. B. Stull, 2005b: Ozone ensemble forecasts. Part II: a Kalman-filter predictor bias correction. *Accepted in August 2005 to be published in the Journal of Geophysical Research.*

Delle Monache, L. and R. B. Stull, 2003: An ensemble air quality forecast over western Europe during an ozone forecast. *Atmospheric Environment*, **37**, 3469–3474.

Sillman, S. and D. He, 2002: Some theoretical results concerning O_3-NO_x-VOC chemistry and NO_x-VOC indicators. *Journal of Geophysical Research*, **107**, 1–15.

Appendix A

Kalman-filter Implementation

Here a step-by-step description of the filter implementation is given. First, σ_ϵ^2 is estimated via the Kalman filter algorithm as follows (by applying Equation (5.5)):

$$p^{\sigma_\epsilon^2}_{t|t-\Delta t} = (p^{\sigma_\epsilon^2}_{t-\Delta t|t-2\Delta t} + \sigma^2_{\sigma^2_\eta})(1 - \beta^{\sigma^2_\epsilon}_{t|t-\Delta t}) \tag{A.1}$$

where $p^{\sigma^2_\epsilon}$ is the expected mean-square-error in the σ^2_ϵ estimate, $\sigma^2_{\sigma^2_\eta}$ is the variance of σ^2_η, and $\beta^{\sigma^2_\epsilon}$ is the Kalman gain when the filter is used to estimate σ^2_ϵ. Next, the new Kalman gain can be computed, similarly to Equation(5.4):

$$\beta^{\sigma^2_\epsilon}_{t+\Delta t|t} = \frac{p^{\sigma^2_\epsilon}_{t|t-\Delta t} + \sigma^2_{\sigma^2_\eta}}{p^{\sigma^2_\epsilon}_{t|t-\Delta t} + \sigma^2_{\sigma^2_\eta} + \sigma^2_{\sigma^2_\epsilon}} \tag{A.2}$$

where $\sigma^2_{\sigma^2_\epsilon}$ is the variance of σ^2_ϵ. Finally, σ^2_ϵ can be estimated by combining Equations (5.3) and (5.8):

$$\sigma^2_{\epsilon,t+\Delta t|t} = \sigma^2_{\epsilon,t|t-\Delta t} + \beta^{\sigma^2_\epsilon}_{t|t-\Delta t}[\frac{(y_t - y_{t-\Delta t})^2}{2+r} - \sigma^2_{\epsilon,t|t-\Delta t}] \tag{A.3}$$

$\sigma^2_{\sigma^2_\epsilon}$ and $\sigma^2_{\sigma^2_\eta}$ are assumed constant, with values of 1 and 0.0005, respectively, as determined from previous works (e.g., Roeger et al., 2003).

Once σ^2_ϵ is estimated, σ^2_η can be computed as $\sigma^2_\eta = r\sigma^2_\epsilon$. Then, Equations (5.5), (5.4), and (5.3) can be applied in sequence, resulting in the final estimate of the bias ($\hat{x}$). This process is iterated trough different Δt, and for the first step, given initial values are used as discussed in Section 5.2.