Ozone ensemble forecasts: 1. A new ensemble design

Luca Delle Monache,[1,2] Xingxiu Deng,[1,3] Yongmei Zhou,[1,4] and Roland Stull[1]

Received 1 June 2005; revised 10 September 2005; accepted 2 December 2005; published 7 March 2006.

[1] A new Ozone Ensemble Forecast System (OEFS) is tested as a technique to improve the accuracy of real-time photochemical air quality modeling. The performance of 12 different forecasts along with their ensemble mean is tested against the observations during 11–15 August 2004, over five monitoring stations in the Lower Fraser Valley, British Columbia, Canada, a population center in a complex coastal mountain setting. The 12 ensemble members are obtained by driving the U.S. Environmental Protection Agency (EPA) Models-3/Community Multiscale Air Quality Model (CMAQ) with two mesoscale meteorological models, each run at two resolutions (12- and 4-km): the Mesoscale Compressible Community (MC2) model and the Penn State/NCAR mesoscale (MM5) model. Moreover, CMAQ is run for three emission scenarios: a control run, a run with 50% more NO_x emissions, and a run with 50% fewer. For the locations and days used to test this new OEFS, the ensemble mean is the best forecast if ranked using correlation, gross error, and root mean square error and has average performance when evaluated with the unpaired peak prediction accuracy. Ensemble averaging removes part of the unpredictable components of the physical and chemical processes involved in the ozone fate, resulting in a more skilful forecast when compared to any deterministic ensemble member. There is not one of the 12 individual forecasts that clearly outperforms the others on the basis of the four statistical parameters considered here. A lagged-averaged OEFS is also tested as follows. The 12-member OEFS is expanded to an 18-member OEFS by adding the second day from the six 12-km "yesterday" forecasts to the "today" ensemble forecast. The 18-member ensemble does not improve the ensemble mean forecast skill. Neither correlation nor a relationship between ensemble spread and forecast error is evident.

Citation: Delle Monache, L., X. Deng, Y. Zhou, and R. Stull (2006), Ozone ensemble forecasts: 1. A new ensemble design, *J. Geophys. Res.*, *111*, D05307, doi:10.1029/2005JD006310.

1. Introduction

[2] 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. 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 these numerical forecast systems are described by *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 Forecasting [*Dabberdt et al.*, 2003] because of the chaotic nature of the atmosphere and chemistry nonlinearity.

[3] 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 do not know a priori which of the two solutions is closest to the true evolution of the system.

[4] The atmosphere exhibits this behavior, and is thus a chaotic system. We are not able to accurately measure the initial state of the atmosphere, because of 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; that is, 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

[1]Atmospheric Science Programme, Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, British Columbia, Canada.
[2]Now at Lawrence Livermore National Laboratory, Livermore, California, USA.
[3]Now at Meteorological Service of Canada, Environment Canada, Montreal, Quebec, Canada.
[4]Now at Meteorological Service of Canada, Environment Canada, Edmonton, Alberta, Canada.

0148-0227/06/2005JD006310$09.00

(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.

[5] 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 initial conditions [*Toth and Kalnay*, 1993, 1997; *Molteni et al.*, 1996], 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.

[6] The ensemble technique can yield similar benefits to real-time 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.

[7] For NWP ensembles, errors typically grow linearly at first, and nonlinearly later [*Kalnay*, 2003]. 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.

[8] *Delle Monache and Stull* [2003] 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.

[9] Another successful implementation of the ensemble approach is given by *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 multimodel 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.

[10] *McKeen et al.* [2005] present results for a multimodel (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.

[11] Recently, *O'Neill and Lamb* [2005] presented an interesting intercomparison of the U.S. Environmental Protection Agency (EPA) Models-3/Community Multiscale Air Quality Model (CMAQ) [*Byun and Ching*, 1999] 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 show 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.

[12] 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 the observations from five different stations over the Lower Fraser Valley (LFV), British Columbia (BC), Canada (see Figure 1). This is a region where ozone modeling is particularly challenging, because of the complex coastal mountain setting [*McKendry and Lundgren*, 2000]. OEFS performance is compared with the performance of each single forecast for a 5-day period (11–15 August 2004).

[13] *Galmarini et al.* [2004b] showed that the ensemble median (the median of the ensemble 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.

[14] For situations where ensemble outliers might be problem, there are some solutions. One is to build a record of error variances for each members based on past forecasts, and then weight each member inversely with its error to compute a weighted ensemble mean (similarly to what is presented by *Pagowski et al.* [2005]). Another is to reduce their systematic errors, and then combine these corrected forecasts into an uniformly weighted average. This is the approach used in the companion paper [*Delle Monache et al.*, 2006, hereinafter referred to as DM2].

[15] Section 2 describes the case study and the data, while a detailed description of the OEFS is given in section 3. Section 4 presents the results and their analysis, and a discussion followed by a summary and conclusions can be found in sections 5 and 6, respectively.

2. Case Study and Data

[16] The LFV lies across the western edge of the Canada/United States border (Figure 1). The main metropolitan area

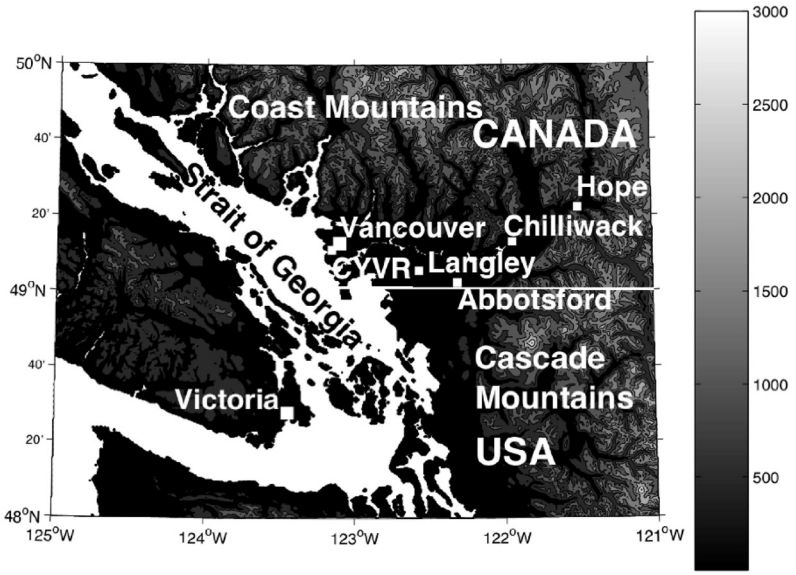

Figure 1. Lower Fraser Valley, which is the floodplain region spanning the stations of Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope. Shading (vertical bar at right) indicates terrain elevation above sea level.

is located at the northwest end of the Valley, where the Greater Vancouver region is located, with a population slightly greater than two million. 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 Mountain Range limiting the Valley's southeastern side.

[17] 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 Lundgren*, 2000].

[18] 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 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, because of 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-hour average concentration) during the first 4 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].

[19] Studies of ozone photochemistry with a scaling-level model 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 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. Ensemble Design

[20] 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, nonhydrostatic 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, nonhydrostatic, primitive equation meteorological model that uses a terrain-following sigma (nondimensionalized pressure) vertical coordinate. The MM5 model is initialized

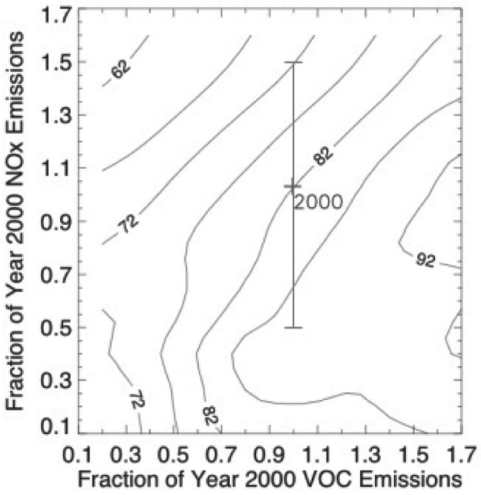

Figure 2. Isopleths of maximum ozone concentration (ppbv) given as a function of year 2000 VOC and NO_x emissions over the Lower Fraser Valley (adapted from *Ainslie* [2004]). The total annual VOC and NO_x emissions are 111,196 and 99,897 tonnes, respectively [*Greater Vancouver Regional District*, 2002]. The vertical bar shows the ±50% NO_x used for the ensemble perturbations.

from the same analysis and for the same five nested grids as MC2, but with 2-way nesting.

[21] Both MC2 and MM5 produce meteorological fields that are used in this study to drive the CMAQ Chemistry Transport Model (CTM) [*Byun and Ching*, 1999]. CMAQ has been run at UBC daily real time for 3 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. CMAQ emissions 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 LFV summertime background ozone concentrations), while the finer grids are initialized each day with the previous day's prediction.

[22] 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 and boundary conditions, meteorological and emission fields, numerical schemes, chemical mechanisms, etc. Unfortunately, to consider all those modeling aspects would require an ensemble with an unfeasibly large number of 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].

[23] A related question is what ensemble size and perturbed attributes are necessary for capturing most of the forecast uncertainty, on the basis of 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.

[24] A flowchart of the OEFS tested in this paper is shown in Figure 3. CMAQ is run with a 12-km horizontal resolution domain covering southern BC, Washington State, and the 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

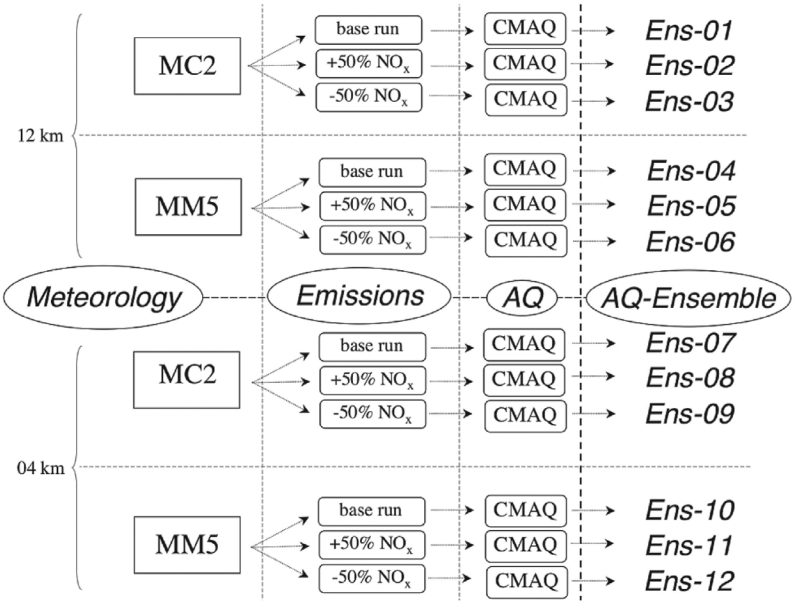

Figure 3. Twelve-member (01, 02, .., 12) Ozone Ensemble Forecast System. It is formed with four different meteorological fields (MC2 at 4 and 12 km and MM5 at 4 and 12 km) and three different emission scenarios: a control run (CTRL), a run with plus 50% NO_x (NOXP), and a run with minus 50% NO_x (NOXN).

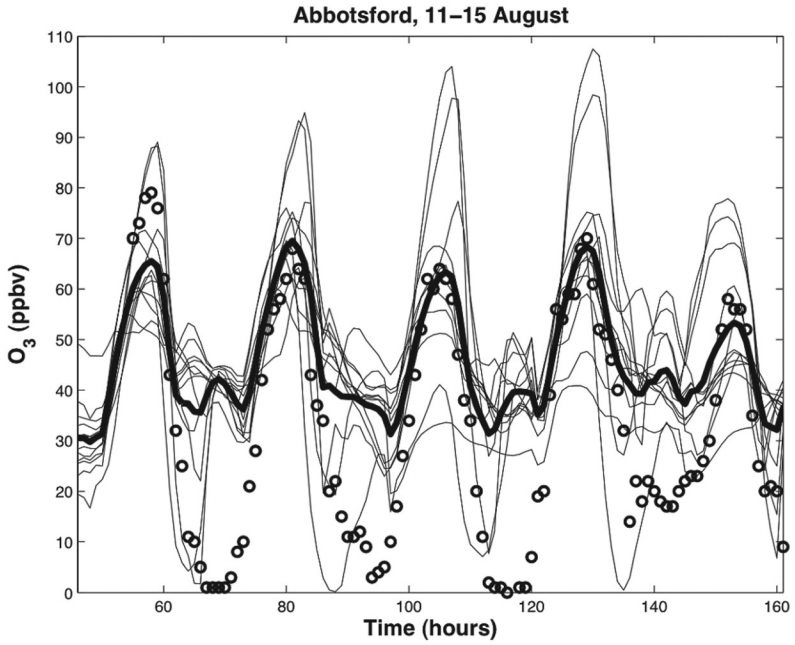

Figure 4. Twelve ensemble members (solid lines) and the ensemble mean (thick solid line) predictions along with the observations (circles), at Abbotsford, 11–15 August 2004.

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 2). These scenarios were chosen because NO_x emissions are mainly anthropogenic [*Jacobson*, 1998] and strongly influence ground-level ozone concentrations [*Steyn et al.*, 1997]. This leads to a system with 12 ensemble members (01, 02, . . ., 12), as shown in Figure 3. An example (Abbotsford, 11–15 August) of the ensemble members (solid lines) and their ensemble mean (thick solid line) temporal evolution, compared with the observed ozone concentrations (circles), can be found in Figure 4.

[25] 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 Nth forecast day ensemble forecast. Figure 5 depicts the resulting 18-member OEFS

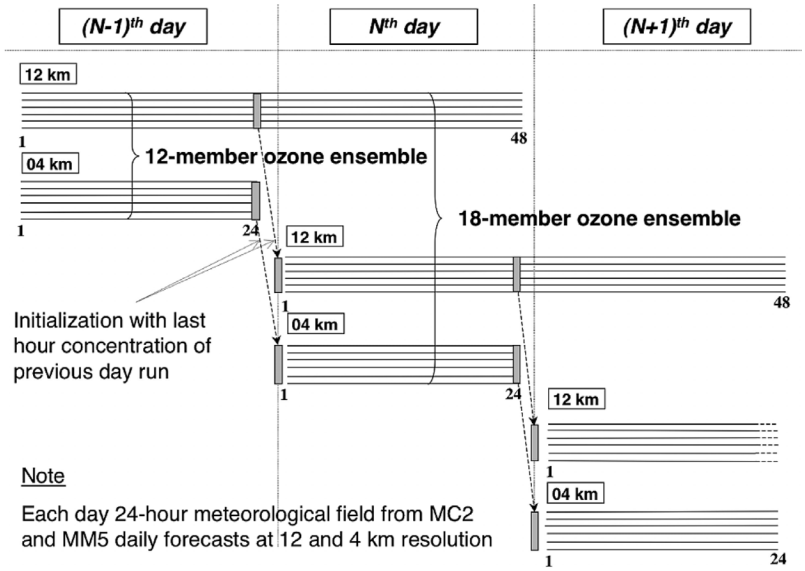

Figure 5. Eighteen-member Ozone Ensemble Forecast System (OEFS). The six 12-km resolution ensemble members are run for 48 hours. The second half of the $(N-1)$th forecast day can be added the Nth day 12-member OEFS to form a lagged averaged ozone 18-member ensemble.

tested in this study, built as a lagged-averaged ozone ensemble (see section 4.4).

4. Results and Analysis

4.1. Verification Statistics

[26] The forecast skill of each ensemble member and the ensemble mean has been evaluated using the following statistical parameters: (1) Pearson product-moment coefficient of linear correlation (herein "correlation"):

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

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

$$gross\ error(station) = \frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} \frac{\left| C_p(t,station) - C_o(t,station) \right|}{C_o(t,station)} \quad (2)$$

(3) root mean square error (RMSE):

$$RMSE(station) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} \left[C_p(t,station) - C_o(t,station) \right]^2} \quad (3)$$

and (4) unpaired peak prediction accuracy (UPPA):

$$UPPA(station) = \frac{1}{N_{day}} \sum_{day=1}^{N_{day}} \cdot \frac{\left| C_p(day,station)_{\max} - C_o(day,station)_{\max} \right|}{C_o(day,station)_{\max}} \quad (4)$$

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

[27] The gross error and UPPA are included in the U.S. EPA guidelines [*U.S. Environmental Protection Agency* (*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 5 days available) of the absolute value of the normalized difference between the predicted and observed maximum at each station (equation (4)). Thus UPPA is nonnegative; hence only the +20% acceptance performance upper limit is used in the next sections.

[28] We selected this set of statistics for the following reasons. We choose correlation to obtain 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.

[29] 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 4.2.3.

[30] 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 %).

[31] UPPA (%) is also used because it measures the ability of the forecasts to predict the ozone peak maximum on a given day. Traditionally, 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 have been steadily decreasing [*Vingarzan*, 2004].

4.2. Twelve-Member OEFS Results

[32] The performance of the OEFS presented in section 3 has been tested by computing the statistical parameters introduced in section 4.1, using the data described in section 2.

4.2.1. Correlation

[33] Figure 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.

[34] Generally, correlation values tend to be lower moving toward 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 night a returning flow (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

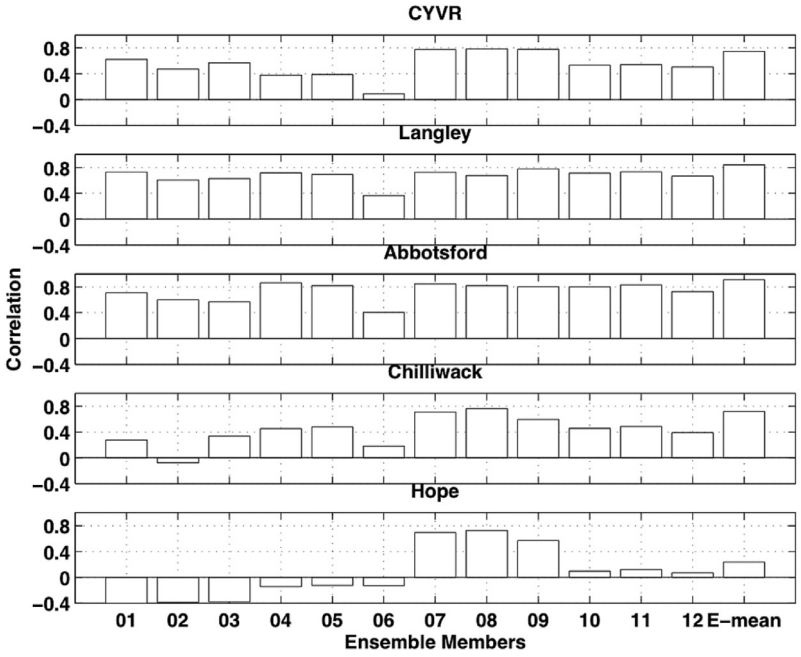

Figure 6. Correlation values between observed and predicted ozone 1-hour average concentrations plotted at five stations (Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack, and Hope) for the 12-member Ozone Ensemble Forecast System (01, 02, .., 12) and the ensemble mean (E-mean) for the 5-day period 11–15 August 2004. Values are within the interval [−1, 1], with correlation = 1 being the best possible value.

causes negative correlation values for the 12 km runs, as shown in Figure 6. Thus the ensembles using finer resolution runs have better correlation values at Hope and 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.

[35] 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.

[36] Table 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 member 08 having a number of first rankings.

[37] 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 above. The correlation values are significantly improved (closer to one) with Kalman filter (KF) postprocessing, as shown in DM2.

4.2.2. Gross Error

[38] The gross error results are shown in Figure 7, and the rankings are summarized in Table 1. Overall the ensemble

Table 1. Ranking of the 12 Ensemble Members (01, 02, …, 12) and the Ensemble Mean (E-Mean) at the Vancouver International Airport (CYVR), Langley, Abbotsford, Chilliwack and Hope Stations Tabulated for Correlation, Gross Error, Root Mean Square Error (RMSE), and Unpaired Peak Prediction Accuracy (UPPA)[a]

	01	02	03	04	05	06	07	08	09	10	11	12	E-Mean
Correlation													
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
Gross Error													
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
RMSE													
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
UPPA													
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

[a]The lowest sum of rankings indicates the best overall performance.

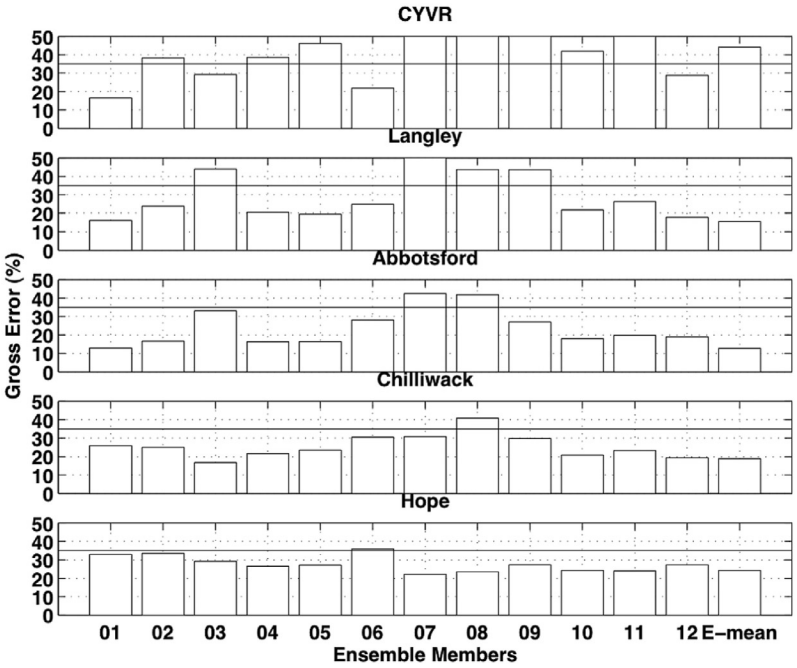

Figure 7. Similar to Figure 6 but for gross error (%). The solid horizontal line is the EPA acceptance value (+35%). Values are within the interval [0, $+\infty$), with a perfect forecast having gross error = 0.

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.

[39] The ensemble mean is well within the 35% EPA acceptance value at Langley, Abbotsford, Chilliwack and Hope. At CYVR the ensemble mean has the highest gross error values, confirming the difficulties for all the ensemble members at this location. In DM2 it is shown that application of the KF postprocessing improves (brings closer to zero) the gross error performance of most forecasts, with an improvement up to 20%.

4.2.3. RMSE

[40] The RMSE results are shown in Figure 8 and summarized in Table 1. In general, the values of this statistical parameter are between 20 and 30 ppbv. However, the KF correction presented in DM2 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 worse than average for gross error. Again, the ranking sum shows that the ensemble mean is the best.

[41] 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:

$$C^*(t, station) = a + bC_o(t, station) \tag{5}$$

where a and b are the least squares regression coefficients of $C_p(t, station)$ and $C_o(t, station)$ (the predicted and observed ozone concentrations, respectively, as defined in section 4.1). Then the following two quantities can be defined:

$$RMSE_s(station) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} [C^*(t, station) - C_o(t, station)]^2} \tag{6}$$

$$RMSE_u(station) = \sqrt{\frac{1}{N_{hour}} \sum_{t=1}^{N_{hour}} \left[C^*(t, station) - C_p(t, station)\right]^2} \tag{7}$$

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:

$$RMSE^2 = RMSE_s^2 + RMSE_u^2 \tag{8}$$

[42] 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 affected little 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.

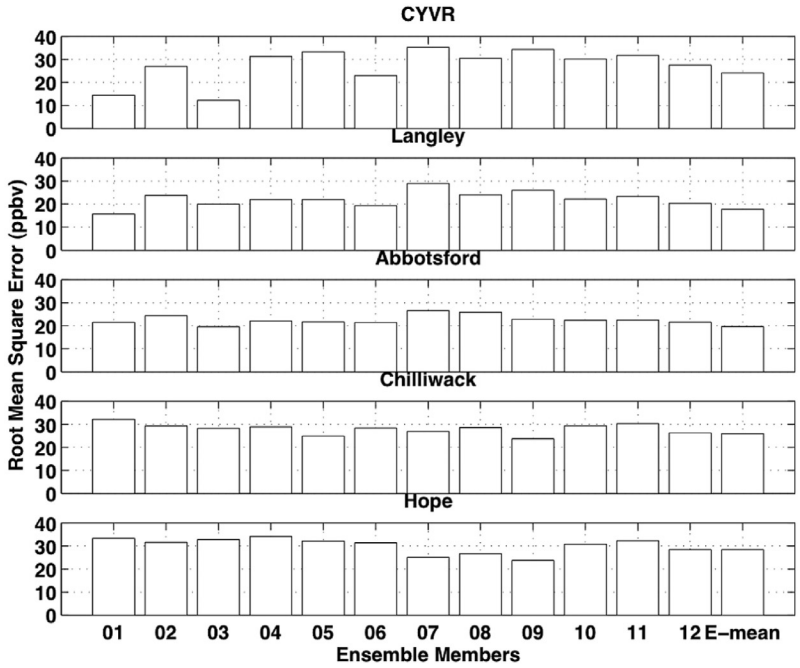

Figure 8. Similar to Figure 6 but for root mean square error (RMSE) (ppbv). Values are within the interval [0, $+\infty$), with a perfect forecast when RMSE = 0.

[43] Figure 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 4.2.1. *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 overstrait pool of pollutants can be readvected over land by the daytime sea breeze.

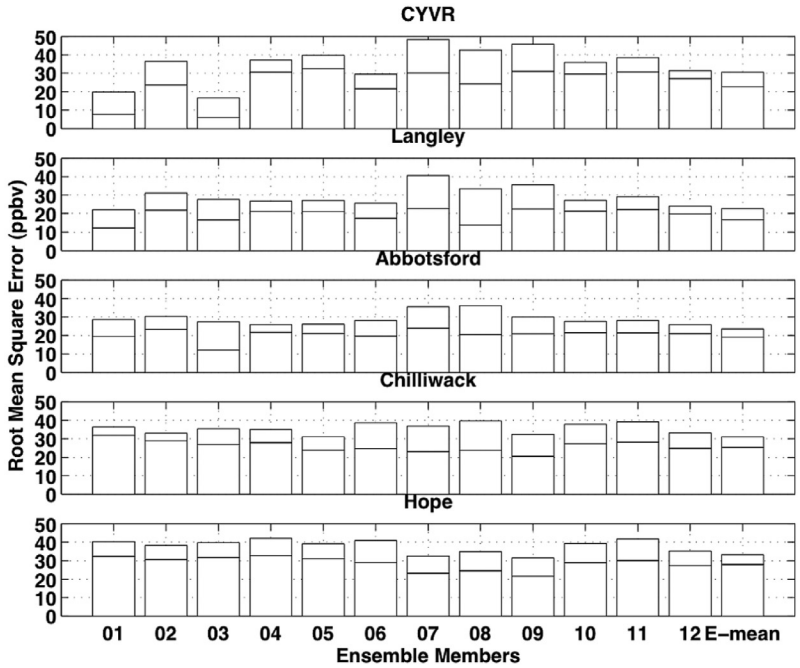

Figure 9. Similar to Figure 8 but segregating the root mean square error into its systematic (bottom bar) and unsystematic components (top bar).

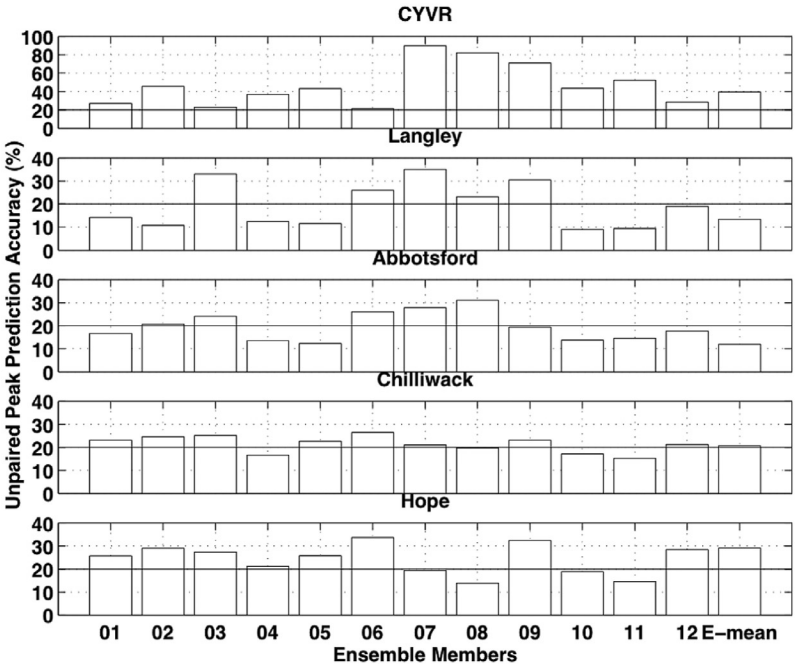

Figure 10. Similar to Figure 6 but for unpaired peak prediction accuracy (UPPA) (%). The solid horizontal lines are the EPA acceptance values (+20%). Values are within the interval [0, $+\infty$), with a perfect peak forecast when UPPA = 0.

[44] 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 CYVR the terrain is flat.

[45] 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 averaging: 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 skilful forecast when compared to any deterministic ensemble member.

4.2.4. UPPA

[46] Figure 10 shows the UPPA results. At CYVR, forecasts 07, 08 and 09 largely overestimate the observed ozone peak concentration, even though at this station they have a good correlation value (close to 0.8). The UPPA rankings in Table 1 are computed using absolute values, so that under and overprediction 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., low 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.

[47] 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 DM2 is shown that application of the KF postprocessing modestly improves (brings closer to zero) the UPPA performance.

4.3. Eleven-Member OEFS Results

[48] 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 recompute 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.

[49] Figure 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.

[50] 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

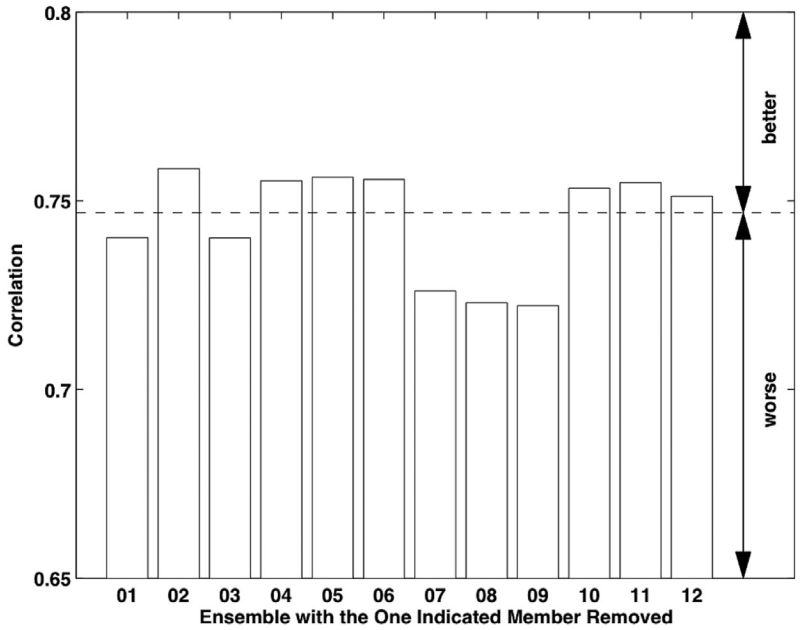

Figure 11. Median (over the five stations) of the correlation of the 11-member ensemble mean, given for the 5-day period 11–15 August 2004. Each bar represents the correlation value for the ensemble mean without the corresponding ensemble member (the label below the bar). The dashed line is the correlation 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.

09) removed give generally worse correlation values, while 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.

[51] Figure 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.

[52] Similar results for RMSE are shown in Figure 13. If the value is below the dashed line, it implies that the

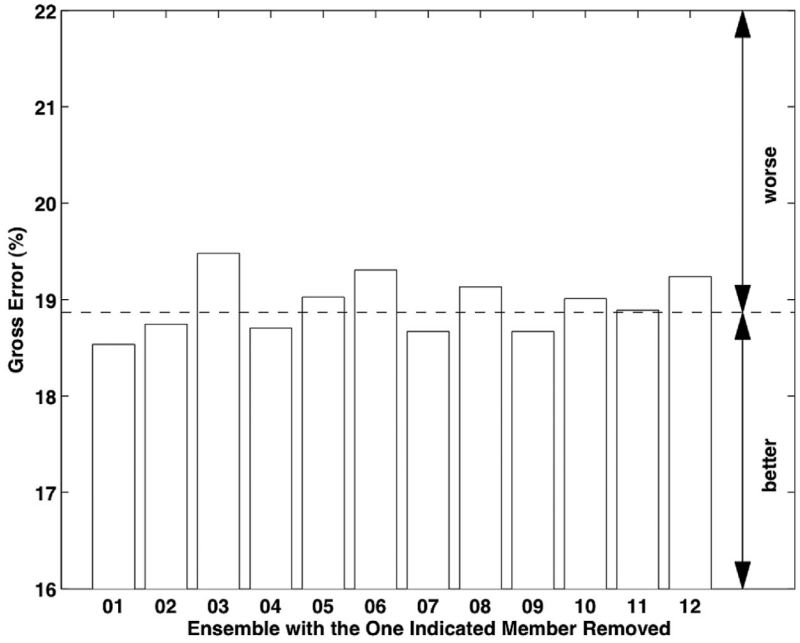

Figure 12. Similar to Figure 11 but for gross error (%). Values are within the interval [0, +∞), with perfect forecast when gross error = 0.

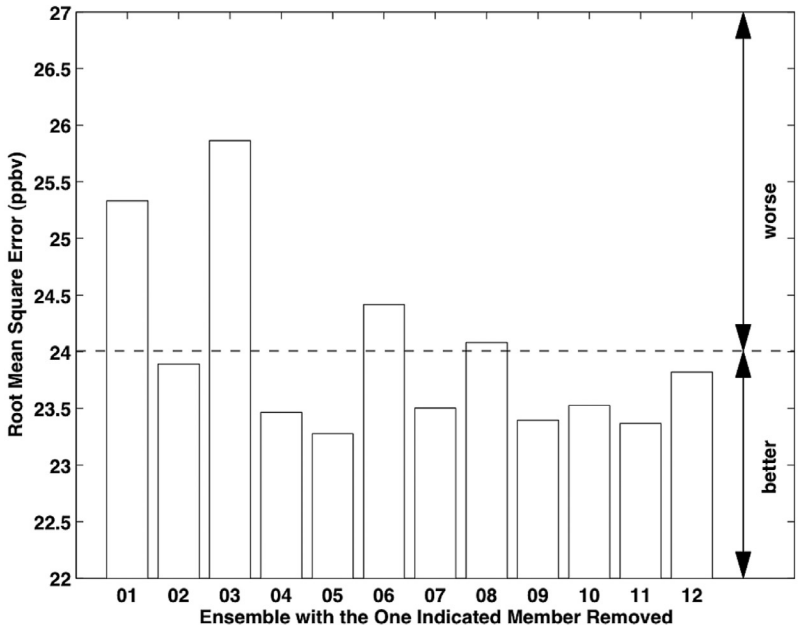

Figure 13. Similar to Figure 11 but for the root mean square error (RMSE) (ppbv). Values are within the interval [0, $+\infty$), with a perfect forecast when RMSE = 0.

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

[53] UPPA results are shown in Figure 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 parameter worse when they are retained, when they contribute to the ensemble.

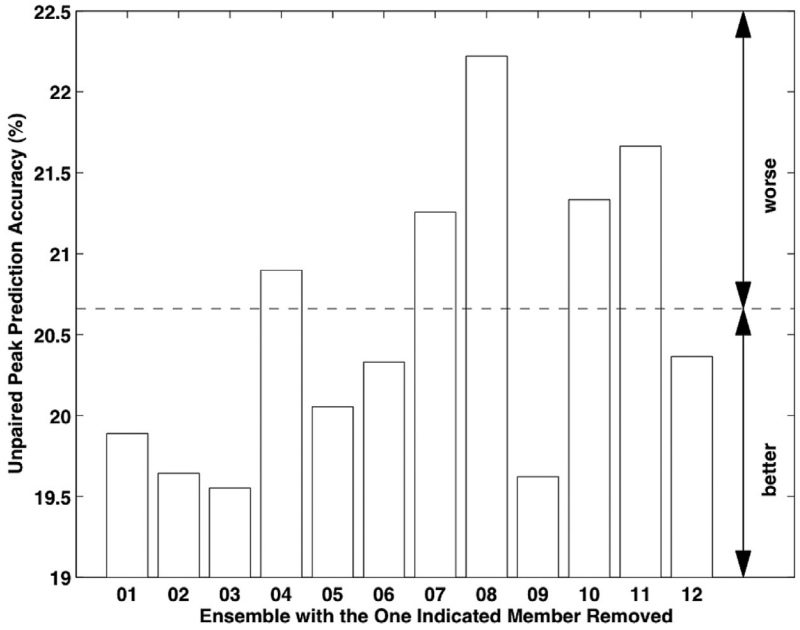

Figure 14. Similar to Figure 11 but for the unpaired peak prediction accuracy (UPPA) (%). Values are within the interval [0, $+\infty$), with a perfect peak forecast when UPPA = 0.

4.4. Eighteen-Member OEFS Results

[54] *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, $t = -\tau, -2\tau, \ldots, -(N-1)\tau$ 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 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 obtained because the lagged average forecast perturbations are not randomly chosen, but better capture the "error of the day." Other applications of this ensemble approach can be found in the literature, as for example in the work by *Dalcher et al.* [1988].

[55] 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 5.

[56] Table 2 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 six lagged members to the original system does not provide valuable results.

[57] Ideally, each ensemble member should give an equally likely time evolution and space distribution of the 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 one-way nesting approach (similarly for MC2, but MM5 runs are implemented with two-way nesting), all the 4 km runs cannot be considered independent of the runs where the driving meteorology and chemistry 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.

5. Discussion

5.1. Taylor Diagrams

[58] A concise way to display and study these results is to use a Taylor diagram [*Taylor*, 2001]. It can be used to create a multistatistics 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.

Table 2. 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 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

[59] At the Vancouver International Airport (CYVR) (Figure 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 16) the ensemble mean is the closest, while forecasts 07 and 08 are the worst, and 09 has an average performance. At Abbotsford (Figure 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 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 19) forecasts 07, 08, and 09 are all closer to the observations than the ensemble mean.

[60] 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 skilful forecast when tested against the observations, and compared to any other individual ensemble member. The key point in favor of the ensemble mean is that is not possible to establish a priori which specific ensemble member will outperform the ensemble mean in any specific situation.

5.2. Meteorology Versus Emission Perturbations

[61] 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 nonperturbed 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 utility of meteorology versus emission perturbations.

[62] 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 appear to be necessary to better predict ozone concentration magnitude.

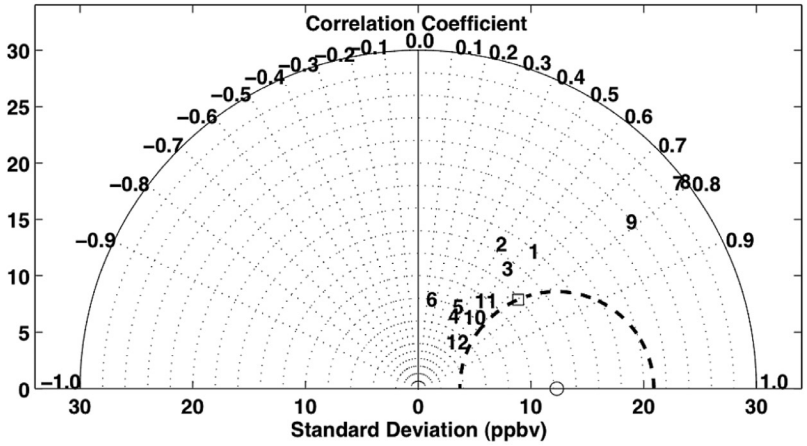

Figure 15. Taylor diagram plotted for Vancouver International Airport (CYVR). The azimuthal position gives the correlation, while the radial distance from the origin is proportional to the standard deviation (ppbv). The circle represents the observations, and the square is the ensemble mean. The numbers 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 observations.

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

5.3. Spread Versus Skill

[64] 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.

[65] In this study no correlation or relationship between ozone ensemble spread and forecast error has been found. This could be 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 (e.g., overnight (see Figure 4)). This could cause most of the forecasts to be close to each other, resulting in a small spread. At the same time those forecasts might be far from the observations, and this could result 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 occasions would be at least partially mitigated by what occurs in those specific circumstances.

6. Summary and Conclusions

[66] A new Ozone Ensemble Forecast System (OEFS) has been tested as a technique to improve the accuracy of real-time air quality forecasts. 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

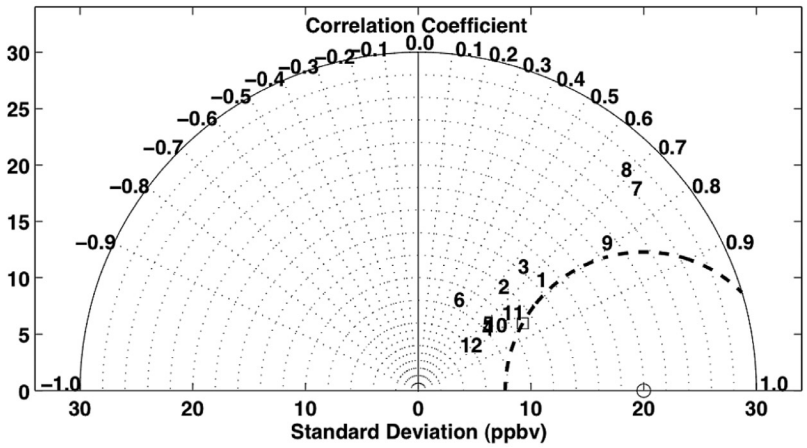

Figure 16. Taylor diagram for Langley (similar to Figure 15).

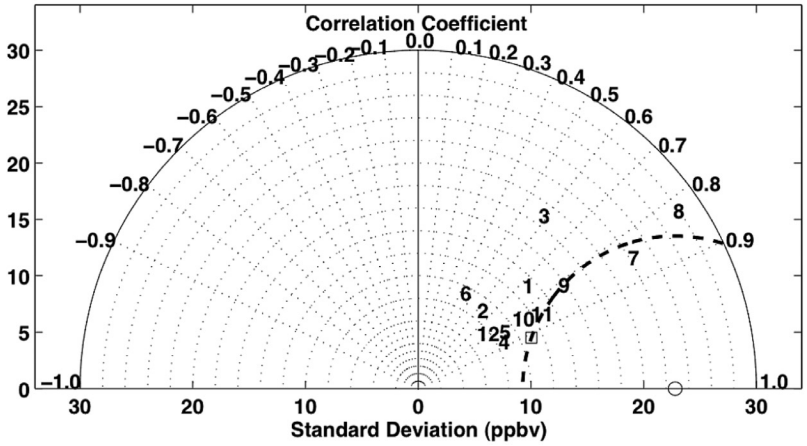

Figure 17. Taylor diagram for Abbotsford (similar to Figure 15).

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

[67] The performance of the ensemble mean and 12 different forecasts is compared with the individual forecasts and tested against the 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:

[68] 1. The ensemble mean is usually the best ozone forecast if ranked using correlation, gross error, or root mean square error (RMSE).

[69] 2. The ensemble mean has an average performance with the unpaired peak prediction accuracy (UPPA). One possible reason could be that ensemble averaging could cause excessive smoothing of the peak values.

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

[71] 4. The ranking sum is useful for comparing overall performance.

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

[73] 6. 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.

[74] 7. 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, on the basis of the four statistical parameters considered here.

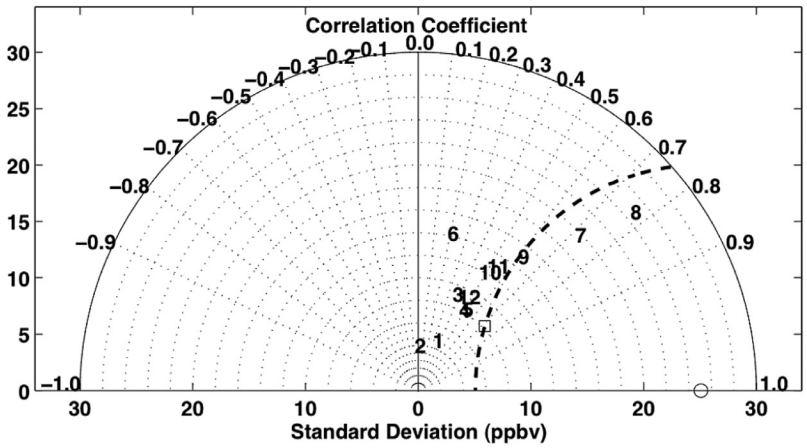

Figure 18. Taylor diagram for Chilliwack (similar to Figure 15).

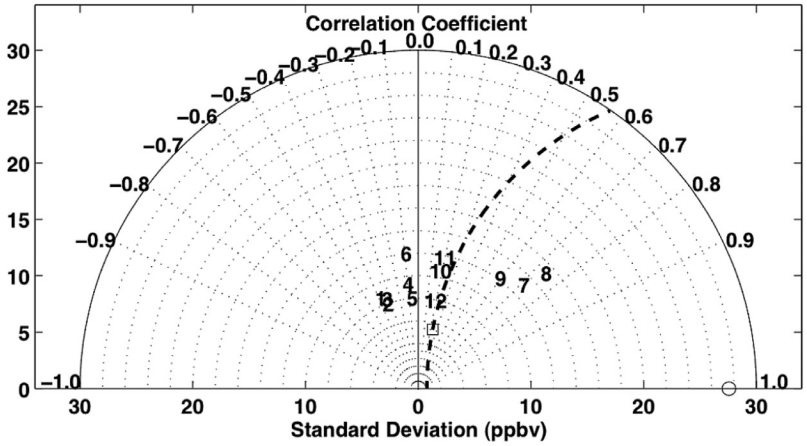

Figure 19. Taylor diagram for Hope (similar to Figure 15).

[75] 8. 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.

[76] These results indicate that ensemble averaging 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 ensemble averaging does improve the forecast accuracy by reproducing the magnitude of ozone concentrations. The ensemble mean has an average performance in predicting the daily ozone maximum, as shown with the UPPA results.

[77] 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 multimodel 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 multimodel approach is an efficient way to perturb the meteorological input in an AQ ensemble design as well.

[78] 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 multiNWP model perturbation) gives good results. Further investigation is needed to clarify this point.

[79] A refinement of the system could focus on the emission perturbations. Ideally, a multimodel approach, using the Sparse Matrix Operator Kernel Emission (SMOKE) model and other state-of-the-art emission preprocessors, 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 preprocessor (e.g., SMOKE) with different configurations, and starting from different emission inventories to generate different (but equally likely) emission fields.

[80] 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 by generating ensemble members by also perturbing other phases of the AQ modeling process, such as the chemistry. For instance, *Hanna et al.* [2001, p. 899] 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 heterogeneous chemistry when compared to gas-phase chemistry.

[81] 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.

[82] 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 skilful forecast when compared to any deterministic ensemble member. In the companion paper [*Delle Monache et al.*, 2006], 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.

[83] **Acknowledgments.** We thank George Hicks, Henryk Modzelewski and Trina Cannon for maintaining the computing system used to perform the simulations presented here. We also thank Todd Plessel (EPA) for providing very useful tools to handle Models-3 formatted data. We are 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. Bruce Ainslie (University of British Columbia) kindly provided Figure 2. We are thankful

to Judi Krzyzanowski and Bruce Thomson (University of British Columbia) for carefully reviewing the first manuscript draft, and to Phil Austin, Allan Bertram, Ian McKendry, Douw Steyn (University of British Columbia), and Brian Lamb (Washington State University) for providing useful suggestions to further improve the paper. Grant support came from the Canadian Natural Science and Engineering Research Council, the BC Forest Investment Account, the British Columbia Ministry of Water Land and Air Protection (formerly the Ministry of Environment), Environment Canada (Colin di Cenzo), 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. Thanks are also due to two anonymous reviewers for their valuable comments and suggestions.

References

Ainslie, B. (2004), A photochemical model based on a scaling analysis of ozone photochemistry, Ph.D. thesis, 311 pp., Univ. of B. C., Vancouver, B. C., 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, *Mon. Weather Rev.*, *125*, 2382–2415.

Brauer, M., and J. R. Brook (1995), Personal and fixed-site ozone measurements with a passive sampler, *J. Air Waste Manage. Assoc.*, *45*, 529–537.

Brown, R. P., T. Butler, and S. W. Hawley (2001), *Ageing of Rubber—Accelerated Weathering and Ozone Test Results*, 192 pp., Rapra, Shawbury, U. K.

Byun, D. W., , and J. K. S. Ching (Eds.) (1999), Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) modeling system, *EPA/600/R-99/030*, Off. of Res. and Dev., U.S. Environ. Prot. Agency, Washington, D. C.

Carmichael, G. R., Y. S. Chang, J. S. Scire, and R. J. Yamartino (1992), The CALGRID mesoscale photochemical grid model—I. Model formulation, *Atmos. Environ.*, *26*, 1493–1512.

Coats, C. J., Jr. (1996), High-performance algorithms in the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system, paper presented at 9th AMS Joint Conference on Applications of Air Pollution Meteorology with Air and Waste Management Association, Am. Meteorol. Soc., Atlanta, Ga., 28 Jan. to 2 Feb.

Dabberdt, W. F., and E. Miller (2000), Uncertainty, ensembles and air quality dispersion modeling: Applications and challenges, *Atmos. Environ.*, *34*, 4667–4673.

Dabberdt, W. F., et al. (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, Natl. Cent. for Atmos. Res., Boulder, Colo.

Dalcher, A., E. Kalnay, and R. N. Hoffmann (1988), Medium range lagged average forecasts, *Mon. Weather Rev.*, *116*, 402–416.

Delle Monache, L., and R. Stull (2003), An ensemble air quality forecast over western Europe during an ozone forecast, *Atmos. Environ.*, *37*, 3469–3474.

Delle Monache, L., X. Deng, Y. Zhou, H. Modzelewski, G. Hicks, T. Cannon, R. Stull, and C. di Cenzo (2004), Air quality ensemble forecast over the Lower Fraser Valley, British Columbia, Canada, paper presented at 27th NATO/CCMS Conference on Air Pollution Modeling 2004, NATO, Banff, Alberta, Canada, 25–29 Oct.

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. Stull (2006), Ozone ensemble forecasts: 2: A Kalman-filter predictor bias correction, *J. Geophys. Res.*, *111*, D05308, doi:10.1029/2005JD006311.

Galmarini, S., et al. (2004a), Ensemble dispersion forecasting—Part I: Concept, approach and indicators, *Atmos. Environ.*, *38*, 4607–4617.

Galmarini, S., et al. (2004b), Ensemble dispersion forecasting—Part II: Application and evaluations, *Atmos. Environ.*, *38*, 4619–4632.

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, *J. Geophys. Res.*, *94*, 12,925–12,956.

Greater Vancouver Regional District (2002), 2000 emissions inventory for the Lower Fraser valley airshed, technical report, GVRD Policy and Plann. Dep., Burnaby, B. C., Canada.

Grell, G., J. Dudhia, and D. Satuffer (1994), A description of the fifth-generation Penn State/NCAR mesoscale model (MM5), *NCAR Tech. Note, NCAR/TN-398+STR*, Natl. Cent. for Atmos. Res., Boulder, Colo.

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, *Atmos. Environ.*, *35*, 891–903.

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*, pp. 39–84, A. F. Lewis, New York.

Hou, D., E. Kalnay, and K. K. Droegemeier (2001), Objective verification of the SAMEX'98 ensemble forecasts, *Mon. Weather Rev.*, *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, *J. Geophys. Res.*, *106*, 20,175–20,188.

Jacobson, M. Z. (1998), *Fundamentals of Atmospheric Modeling*, 656 pp., Cambridge Univ. Press, New York.

Kalnay, E. (2003), *Atmospheric Modeling, Data Assimilation and Predictability*, 341 pp., Cambridge Univ. Press, New York.

Leith, C. E. (1974), Theoretical skill of Monte Carlo forecasts, *Mon. Weather Rev.*, *102*, 409–418.

Lorenz, E. N. (1963), Deterministic non-periodic flow, *J. Atmos. Sci.*, *20*, 130–141.

Martilli, A., and D. G. Steyn (2004), A numerical study of recirculation processes in the Lower Fraser Valley (British Columbia, Canada), paper presented at 27th NATO/CCMS Conference on Air Pollution Modeling 2004, NATO, Banff, Alberta, Canada, 25–29 Oct.

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, *Bull. Am. Meteorol. Soc.*, *85*, 525–548.

McKeen, S. A., et al. (2005), Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, *J. Geophys. Res.*, *110*, D21307, doi:10.1029/2005JD005858.

McKendry, I. G. (1994), Synoptic circulation and summertime ground-level ozone concentrations at Vancouver, British Columbia, *J. Appl. Meteorol.*, *33*, 627–641.

McKendry, I. G., and J. Lundgren (2000), Tropospheric layering of ozone in regions of urbanized complex and/or coastal terrain: A review, *Prog. Phys. Geogr.*, *24*, 329–354.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis (1996), The new ECMWF ensemble prediction system: Methodology and validation, *Q. J. R. Meteorol. Soc.*, *122*, 73–119.

Nodop, K., R. Connolly, and G. Girardi (1998), The field campaigns of the European Tracer Experiment (ETEX): Overview and results, *Atmos. Environ.*, *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, *Environ. Sci. Technol.*, *39*, 5742–5753.

Pagowski, M., et al. (2005), A simple method to improve ensemble-based ozone forecasts, *Geophys. Res. Lett.*, *32*, L07814, doi:10.1029/2004GL022305.

Runeckles, V. (2002), Effects on vegetation and ecosystems, in *A Citizen's Guide to Air Pollution*, pp. 177–216, Bates and Caton, Vancouver, B. C., Canada.

Russell, A., and R. Dennis (2000), NARSTO critical review of photochemical models and modeling, *Atmos. Environ.*, *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, *Atmos. Environ.*, *36*, 5771–5782.

Stensrud, D. J., J.-W. Bao, and T. T. Warner (1998), Ensemble forecasting of mesoscale convective systems, paper presented at 12th Conference on Numerical Weather Prediction, Am. Meteorol. Soc., Phoenix, Ariz., 11–16 Jan.

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, *Atmos. Environ.*, *31*, 2025–2035.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.*, *106*, 7183–7192.

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

Toth, Z., and E. Kalnay (1993), Ensemble forecasting at NMC: The generation of perturbations, *Bull. Am. Meteorol. Soc.*, *74*, 2317–2330.

Toth, Z., and E. Kalnay (1997), Ensemble forecasting at NCEP: The breeding method, *Mon. Weather Rev.*, *125*, 3297–3319.

U.S. Environmental Protection Agency (1991), Guideline for regulatory application of the Urban Airshed Model, *USEPA Rep. EPA-450/4-91-013*, Research Triangle Park, N. C.

Vaughan, J., et al. (2004), A numerical daily air quality forecast system for the Pacific Northwest, *Bull. Am. Meteorol. Soc.*, *85*, 549–561.

Vingarzan, R. (2004), A review of surface ozone background levels and trends, *Atmos. Environ.*, *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, *Mon. Weather Rev.*, *129*, 729–747.

Willmott, C. J. (1981), On the validation of models, *Phys. Geogr.*, *2*, 184–194.

L. Delle Monache, Lawrence Livermore National Laboratory, L-103, Livermore, CA 94550, USA. (ldm@llnl.gov)

X. Deng, Meteorological Service of Canada, Environment Canada, Montreal, Quebec, Canada.

R. Stull, Department of Earth and Ocean Science, University of British Columbia, 6339 Stores Road, Vancouver, BC, Canada V6T 1Z4.

Y. Zhou, Meteorological Service of Canada, Environment Canada, Edmonton, Alberta, Canada.