JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 111, D24307, doi:10.1029/2005JD006917, 2006

Probabilistic aspects of meteorological and ozone regional ensemble forecasts

Luca Delle Monache,[1,2] Joshua P. Hacker,[3] Yongmei Zhou,[1,4] Xingxiu Deng,[1,5] and Roland B. Stull[1]

Received 23 November 2005; revised 7 June 2006; accepted 17 July 2006; published 29 December 2006.

[1] This study investigates whether probabilistic ozone forecasts from an ensemble can be made with skill: i.e., high verification resolution and reliability. Twenty-eight ozone forecasts were generated over the Lower Fraser Valley, British Columbia, Canada, for the 5-day period 11–15 August 2004 and compared with 1-hour averaged measurements of ozone concentrations at five stations. The forecasts were obtained by driving the Community Multiscale Air Quality Model (CMAQ) model with four meteorological forecasts and seven emission scenarios: a control run, ±50% NO_x, ±50% volatile organic compounds (VOC), and ±50% NO_x combined with VOC. Probabilistic forecast quality is verified using relative operating characteristic curves, Talagrand diagrams, and a new reliability index. Results show that both meteorology and emission perturbations are needed to have a skillful probabilistic forecast system: the meteorology perturbation is important to capture the ozone temporal and spatial distribution and the emission perturbation is needed to span the range of ozone concentration magnitudes. Emission perturbations are more important than meteorology perturbations for capturing the likelihood of high ozone concentrations. Perturbations involving NO_x resulted in a more skillful probabilistic forecast for the episode analyzed, and therefore the 50% perturbation values appear to span much of the emission uncertainty for this case. All of the ensembles analyzed show a high ozone concentration bias in the Talagrand diagrams, even when the biases from the unperturbed emissions forecasts are removed from all ensemble members. This result indicates nonlinearity in the ensemble, which arises from both ozone chemistry and its interaction with input from particular meteorological models.

Citation: Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull (2006), Probabilistic aspects of meteorological and ozone regional ensemble forecasts, *J. Geophys. Res.*, *111*, D24307, doi:10.1029/2005JD006917.

1. Introduction

[2] Exposure to ozone concentration 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 by *Delle Monache et al.* [2004], *McHenry et al.* [2004] and *Vaughan et al.* [2004]. The U.S. Weather Research Program and its Prospectus Development Team on Air Quality Forecasting [*Dabberdt et al.*, 2003] recommended a probabilistic approach to AQ forecasting because of the chaotic nature of the atmosphere and chemistry nonlinearity.

[3] It has been found for regional, mesoscale numerical weather prediction (NWP) that the ensemble mean is more accurate that an individual model realization [e.g., *Hou et al.*, 1998; *Gneiting et al.*, 2005]. Recent studies have shown that the ensemble average yields similar benefits for AQ prediction, because there are similar model complexities and constraints (e.g., *Delle Monache and Stull* [2003], *McKeen et al.* [2005], and *Delle Monache et al.* [2006a], hereinafter referred to as DM06a). Moreover, through probabilistic forecasts, NWP ensembles have been useful for providing information about the likelihood of possible future evolutions of the atmosphere, an approach that is extended here for AQ forecasts (i.e., ozone). Similar to NWP, AQ ensembles may be able to provide reliable probabilistic information about possible AQ scenarios.

[4] Given the nonlinear nature of both NWP and AQ models, the differences among ensemble members of an Ozone Ensemble Forecast System (OEFS) may be able to

[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]Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado, USA.
[4]Now at Meteorological Service of Canada, Environment Canada, Edmonton, Alberta, Canada.
[5]Now at Meteorological Service of Canada, Environment Canada, Montreal, Quebec, Canada.

0148-0227/06/2005JD006917$09.00

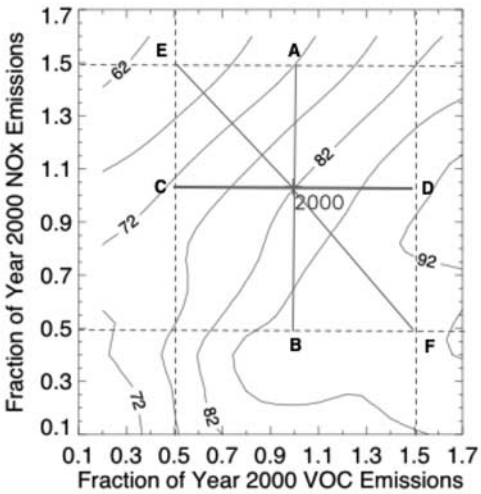

Figure 1. Isopleths of maximum ozone concentration (ppbv) are given as a function of year 2000 volatile organic compounds (VOC) and NO_x emissions over the Lower Fraser Valley (adapted from *Ainslie* [2004]). The year 2000 total annual VOC and NO_x emissions were 111,196 and 99,897 metric tonnes, respectively [*Greater Vancouver Regional District*, 2002]. Vertical bar spans plus (point A) and minus (point B) 50% NO_x perturbations. Horizontal bar spans plus (point D) and minus (point C) 50% VOC perturbations. Diagonal bar extends from plus 50% NO_x and minus 50% VOC perturbation (point E) to the minus 50% NO_x and plus 50% VOC perturbation (point F).

account for some of the uncertainties associated with each component of the modeling process. It has been observed that NWP ensemble error growth typically has two distinct phases: an initial period of linear growth, followed by a nonlinear period [*Kalnay*, 2003] that extends to the limits of predictability. In AQ ensembles the linear growth period might be shorter because of the strong nonlinear nature of the chemistry. Additionally, complex interactions between ozone chemistry and a driving meteorological model may introduce further nonlinearity in the ensemble error growth. Therefore the differences among AQ ensemble members may account for the uncertainties associated with each component of the AQ process more rapidly than for NWP ensembles. These effects have not been systematically studied. This work is one step toward better understanding AQ forecast uncertainties.

[5] *Ainslie* [2004] shows that AQ in the LFV depends nearly equally on NO_x and volatile organic compounds (VOC) emission variations (Figure 1). If the maximum ozone concentration is plotted as a function of NO_x and VOC emissions, the state of the LFV is above the ridgeline of ozone relative maxima. DM06a experimented with emission perturbations having 50% more NO_x emissions (point A in Figure 1), and 50% less (point B in Figure 1). They 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.

[6] The successful experiments in DM06a prompted the work presented here, where VOC perturbations are also considered, and the 12-member ensemble has been expanded to 28 members.

[7] The different forecasts are grouped in 13 different OEFS categories, as described in Section 2. The performance of these OEFS groups are investigated here by comparing their forecast skill as probabilistic forecasts, using the probabilistic forecast skill metrics described in Section 3. The effects of different perturbations, resolutions, and driving models on the ensemble skill are analyzed in Section 4. In Section 5 conclusions are summarized.

2. Ozone Ensembles

[8] Here we briefly describe the composition of the ensembles. For more detailed information, the reader is referred to DM06a. The ensembles used four meteorological forecasts and seven emissions scenarios, yielding a total of 28 members that can be subsampled to understand their overall contributions. Meteorological forecasts were generated by running two different mesoscale NWP models, the Mesoscale Compressible Community (MC2) NWP model [*Benoit et al.*, 1997] and the Penn State/NCAR mesoscale (MM5) model [*Grell et al.*, 1994], each with horizontal grid spacing of 4 and 12 km. These models have been running daily for a decade at the University of British Columbia (UBC), (http://weather.eos.ubc.ca/wxfcst/). Forecasts were initialized at 0000 UTC and run for 48 hours, with initial and boundary conditions from the NCEP North American Mesoscale (NAM) model.

[9] 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*, 1999], which used the NWP model runs and the Sparse Matrix Operator Kernel Emission (SMOKE) system [*Coats*, 1996] emissions estimates as input. Emissions uncertainty is considered by perturbing both NO_x and VOC emissions. Each ozone precursor is independently perturbed ±50% about the control, resulting in four additional forecasts (points A–D in Figure 1). The precursors are also perturbed together, resulting in two additional forecasts (points E and F in Figure 1). Including the control leads to seven emissions scenarios. The primary difference between the experimental data sets in this study and in DM06a is the addition of VOC perturbations, which were not considered before.

[10] The 28 AQ forecasts resulting from the above perturbation combinations are tested here for the same AQ episode analyzed in DM06a, with 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 2). The study period is 11–15 August 2004, and further details about the data and episode can be found in Section 2 of DM06a.

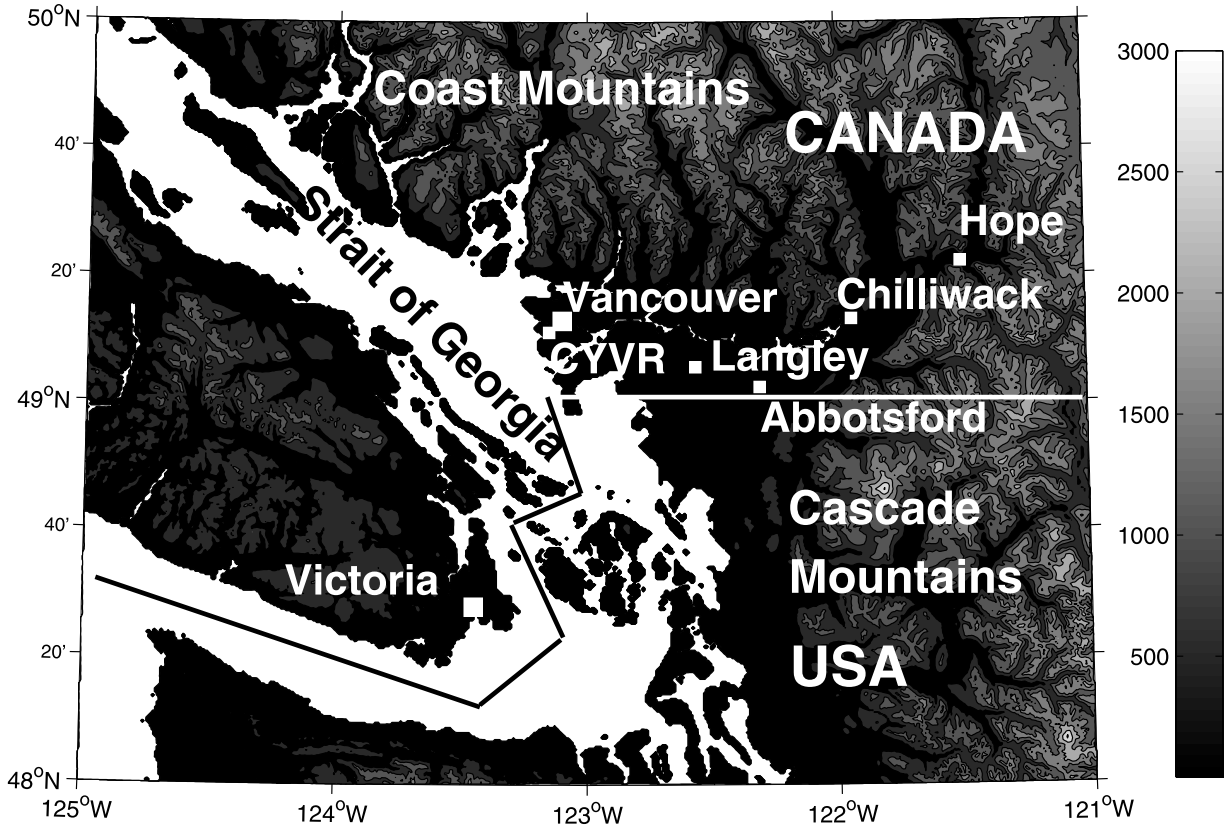

Figure 2. The Lower Fraser Valley is a floodplain spanning the ozone stations of Vancouver International Airport (CYVR) (urban), Langley (suburban), Abbotsford (urban), Chilliwack (suburban), and Hope (rural). The triangular valley is widest near CYVR along the coast of the Georgia Strait and tapers to a narrow gorge between steep mountain walls near Hope. Shading (vertical bar at right) indicates terrain elevation (m) above sea level.

[11] The 28 ensemble members are subsampled to form 13 different ensembles, as also summarized in Table 1 (in parenthesis are the ensemble name and number of members): (1) All the forecasts available (ALL, 28 members). (2) Meteorology and NO_x perturbations combined together (MET+NO_x, 12 members). (3) Meteorology and VOC perturbations (MET+VOC, 12 members). (4) Meteorology and NO_x combine with VOC perturbations (MET+NO_x-VOC, 12 members). (5) All the ensemble members driven by MC2 at 12 km (MC2-12, seven members). (6) All the

Table 1. Ensemble Members Included in Each of the 13 Ensemble Groups[a]

	MC2-CMAQ								MM5-CMAQ								
	12 km				4 km				12 km				4 km				
Ensemble	Base	NO_x	VOC	NO_x VOC	Base	NO_x	VOC	NO_x VOC	Base	NO_x	VOC	NO_x VOC	Base	NO_x	VOC	NO_x VOC	Ensemble Members
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**

[a]"Base" is the forecast obtained by running the Community Multiscale Air Quality Model (CMAQ) with the base emissions at one of the two possible resolutions (12 or 4 km) driven by numerical weather prediction (NWP) models (MC2 or MM5). NO_x indicates runs with perturbations of ±50% NO_x, volatile organic compounds (VOC) includes the ±50% VOC runs, and NO_x VOC 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.

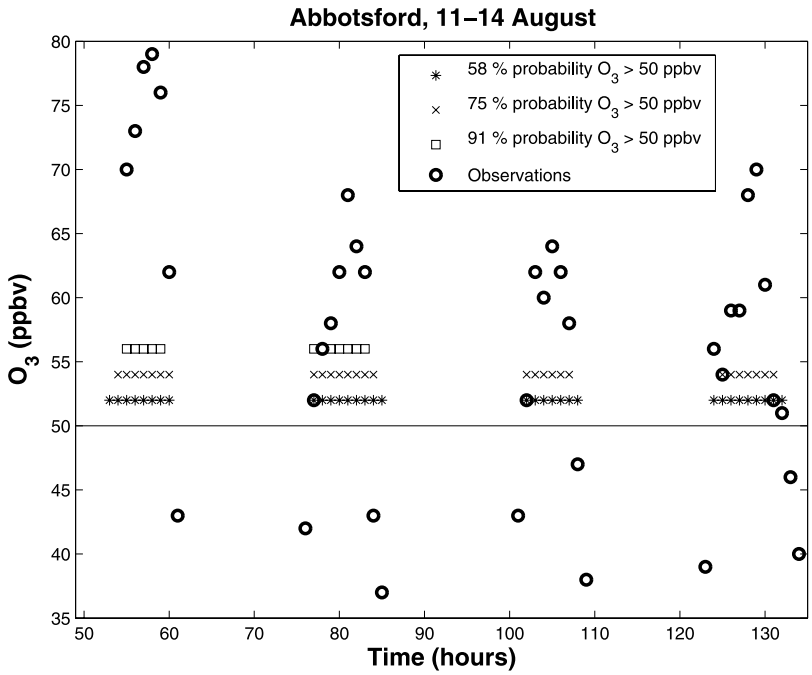

Figure 3. Probabilities of ozone concentration to be above 50 ppbv, as predicted by MET+NO_x at Abbotsford, 11–14 August 2004. Asterisks indicate hours when the forecasted probability is 58% (7 out of 12 ensemble members are predicting the event), crosses when it is 75% (9 out of 12), and squares when it is 91% (11 out of 12). Continuous line represents the ozone 50 ppbv concentration threshold. Circles indicate observations.

ensemble members driven by MC2 at 4 km (MC2-04, seven members). (7) All the ensemble members driven by MM5 at 12 km (MM5-12, seven members). (8) All the ensemble members driven by MM5 at 4 km (MM5-04, seven members). (9) All the control runs (MET, four members). (10) All the ensemble members with 12 km resolution (12 km, 14 members). (11) All the ensemble members with 4 km resolution (04 km, 14 members). (12) All the ensemble members driven by MC2 (MC2-ALL, 14 members). (13) All the ensemble members driven by MM5 (MM5-ALL, 14 members).

[12] MET+NO_x, MET+VOC, and MET+NO_xVOC are ensembles generated with both meteorology and emission perturbations, while MC2-12, MC2-04, MM5-12, and MM5-04 are ensembles where only emission perturbations are considered (i.e., the members in each of them are driven by the same meteorological input field). Ensemble MET, formed by the four control runs, takes into account meteorology perturbations from NWP model differences alone.

[13] Ensembles 12 km and 04 km will help to understand the effects of different horizontal grid spacing for a region such as the LFV having high mountains. Finally, MC2-ALL and MM5-ALL give insights about the different contributions from different NWP models (MC2 and MM5) while including different resolutions.

3. Probabilistic Forecast Verification Statistics

[14] DM06a tested a 12 member ensemble mean for a 5-day episode (August 2004) over the LFV, British Columbia, Canada. *Delle Monache et al.* [2006b] (hereinafter referred to as DM06b), using the same test case as in DM06a, tested the Kalman filter as a post processing approach to remove ozone forecast systematic errors. In the present work, the ensemble system presented in DM06a and DM06b have been expanded to 28 members with the addition of VOC perturbations (as explained in section 2), and probabilistic ozone forecasts have been tested for the same test case.

[15] A probabilistic forecast system (PFS) can be built from a given set of ensemble members by estimating the probability of an event occurrence. This probability can be computed as the 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 ozone concentration above a certain threshold. Figure 3 is an example of the MET+NO_x ensemble forecasted probabilities of ozone to be above 50 ppbv, at Abbotsford, 11–14 August 2004. With higher forecasted probability the predicted time of occurrence of the event (i.e., ozone to be above 50 ppbv) is progressively closer to what is observed.

[16] Probabilistic forecast skill can be evaluated by determining the predictive accuracy of a forecast distribution. With this in mind two important forecast attributes can be computed: resolution and reliability. Both are concerned with the conditional probability $p(o|f)$ of observation (o) given forecast (f). An in-depth discussion of those and other attributes of probabilistic forecasts can be found in *Jolliffe and Stephenson* [2003].

3.1. Reliability

[17] Reliability measures the capability of the PFS to predict unbiased estimates of the observed frequency

Table 2. Out of the 549 Valid Observation Points Available, the Portion of Observations With Ozone Concentration Greater Than the Given Threshold

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

associated with different forecast frequencies. In a perfectly reliable forecast, the forecasted frequency of the event should be equal to the observed frequency of the event for all the cases when that specific event is forecasted. It can be improved with a forecast calibration such as bias correction; e.g., by reassigning the forecast frequency values on the basis of a long series of past forecasts, or by Kalman filtering each individual forecast on the basis of recent past bias values, as shown in DM06b. Reliability is necessary but not sufficient to establish whether a PFS produces valuable forecasts. For instance, a system that always forecasts the climatological frequency of an event is reliable, but may not prove valuable for decision makers.

[18] Reliability can be measured with a Talagrand diagram [*Talagrand and Vautard*, 1997], also known as the 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 the bins all show the same frequency ("ideal bin count"). If each ensemble member represents an equally likely time evolution and spatial distribution of the ozone concentration, then the ensemble exhibits a perfect spread, and the observations are equally likely to fall between any two members.

[19] In this study a new summary index, called a "reliability index" (*RI*), is introduced as the reliability attribute. It is computed as follows:

$$\frac{\textit{mean bin distance from ideal bin count}}{\textit{ideal bin count}} \times 100 = \frac{\frac{1}{N_{bin}} \sum_{i=1}^{N_{bin}} \left| \frac{count_i}{N_{point}} - \frac{1}{N_{bin}} \right|}{\frac{1}{N_{bin}}} \times 100 = \sum_{i=1}^{N_{bin}} \left| \frac{count_i}{N_{point}} - \frac{1}{N_{bin}} \right| \times 100 \quad (1)$$

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, N_{point} is the sum of $count_i$, for $i = 1, \ldots, N_{bin}$ (i.e., the sample size).

[20] Lower *RI* means that the bins are closer to representing requencies associated with a perfectly reliable forecast. This index can be useful since the Talagrand diagram of the 13 PFSs all have similar shapes, as shown in the next section, and can then provide further information to compare their reliability. The *RI* does not provide any information about the Talagrand diagram shape.

[21] When the ensembles are drawn from the same distribution, the *RI* as defined in equation (1) tends to increase with increasing ensemble size following $\sqrt{esize/esizemin}$, where *esize* is the size of the ensemble for which *RI* is computed, and *esizemin* is the size of the smallest ensemble considered. This would prevent its application in cases as here, where ensembles with different sizes are compared with each other. For this reason, equation (1) is normalized as follows:

$$RI = \frac{\left(\sum_{i=1}^{N_{bin}} \left| \frac{count_i}{N_{point}} - \frac{1}{N_{bin}} \right| \times 100 \right)}{\sqrt{\frac{esize}{esizemin}}}. \quad (2)$$

Hereafter, this normalized expression is used because it makes *RI* independent of ensemble size. Again, lower *RI* is better.

[22] The *RI* (%) measures the degree of closeness of a Talagrand diagram to its ideal flat shape, without distinguishing between ensemble bias and underdispersion (i.e., when the ensemble does not have enough spread, defined as the standard deviation of the ensemble members about the ensemble mean, to capture all the observed outcomes). Recently, a similar index (δ) measuring the "deviation of the histogram from flatness" was introduced by *Candille and Talagrand* [2005]. This index also takes into account the distance from the ideal bin height, but does so by considering a sum over the squares of the differences of $count_i$ minus N_{point}/N_{bin} for $i = 1, \ldots, N_{bin}$, 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 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 realizations of the prediction process are not independent". For the analysis here, the *RI* is easier to interpret.

3.2. Resolution

[23] 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 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 2 shows the concentration threshold values used in this study. As the concentration increases, the number of events 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 because of their importance for health-related issues [*Horvath and McKee*, 1994; *Brauer and Brook*, 1995].

[24] Resolution is quantified by the Relative Operating Characteristic (ROC), developed in the field of signal

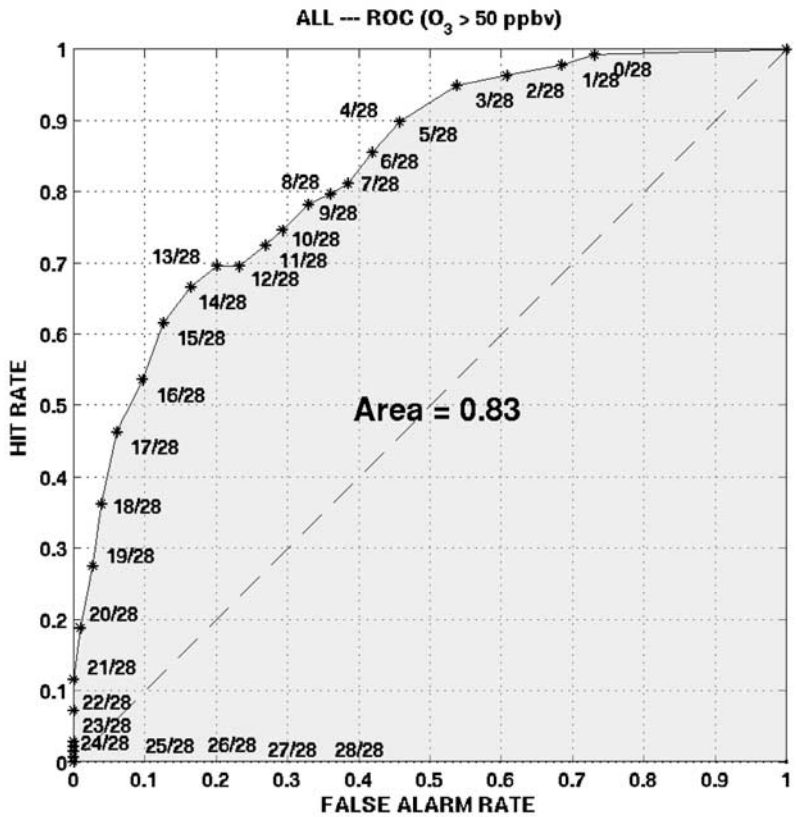

Figure 4. Relative Operating Characteristic (ROC) curve for the "ALL ensemble" (28 members) for observed ozone concentration above 50 ppbv. The better the probabilistic forecast, the closer the ROC curve is to the upper left corner. The shaded portion of the plot represents the ROC area (large areas 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 frequency threshold (the labels adjacent to the asterisks), where the frequency threshold assumes values from 0/28 to 28/28, with increments of 1/28.

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 hit is scored when the ensemble predicts a likelihood of the event is greater than or equal to the given probability threshold. The hit rate is computed as the ratio of the number of correct forecasts of the event to the total number of event occurrences, while the false alarm rate is computed as ratio of the number of noncorrect event forecasts to the total number of event nonoccurrences. Then, hit rates are plotted on the ordinate against the corresponding false alarm rates on the abscissa to generate the ROC curve.

[25] For a PFS with good resolution, the ROC curve is close to the upper left hand corner of the graph (e.g., Figure 4). The area under the ROC curve 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 is the forecast. A value of 0.5 indicates that the forecast system has no skill, relative to a chance forecast from climatology. The ROC curve does not depend on the forecast bias, hence is independent of reliability. It represents the PFS intrinsic value, or the potential value of an unbiased ensemble.

[26] Figure 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. Probability thresholds assume values from 0/28 to 28/28, with increments of 1/28, and are labeled adjacent to the data points on the curve. In this example, a correct forecast of the event occurs if the forecasted frequency is above the given probability threshold when the observed ozone concentration is above 50 ppbv. Similar curves can be produced for other concentration thresholds.

4. Probabilistic Forecast Results

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

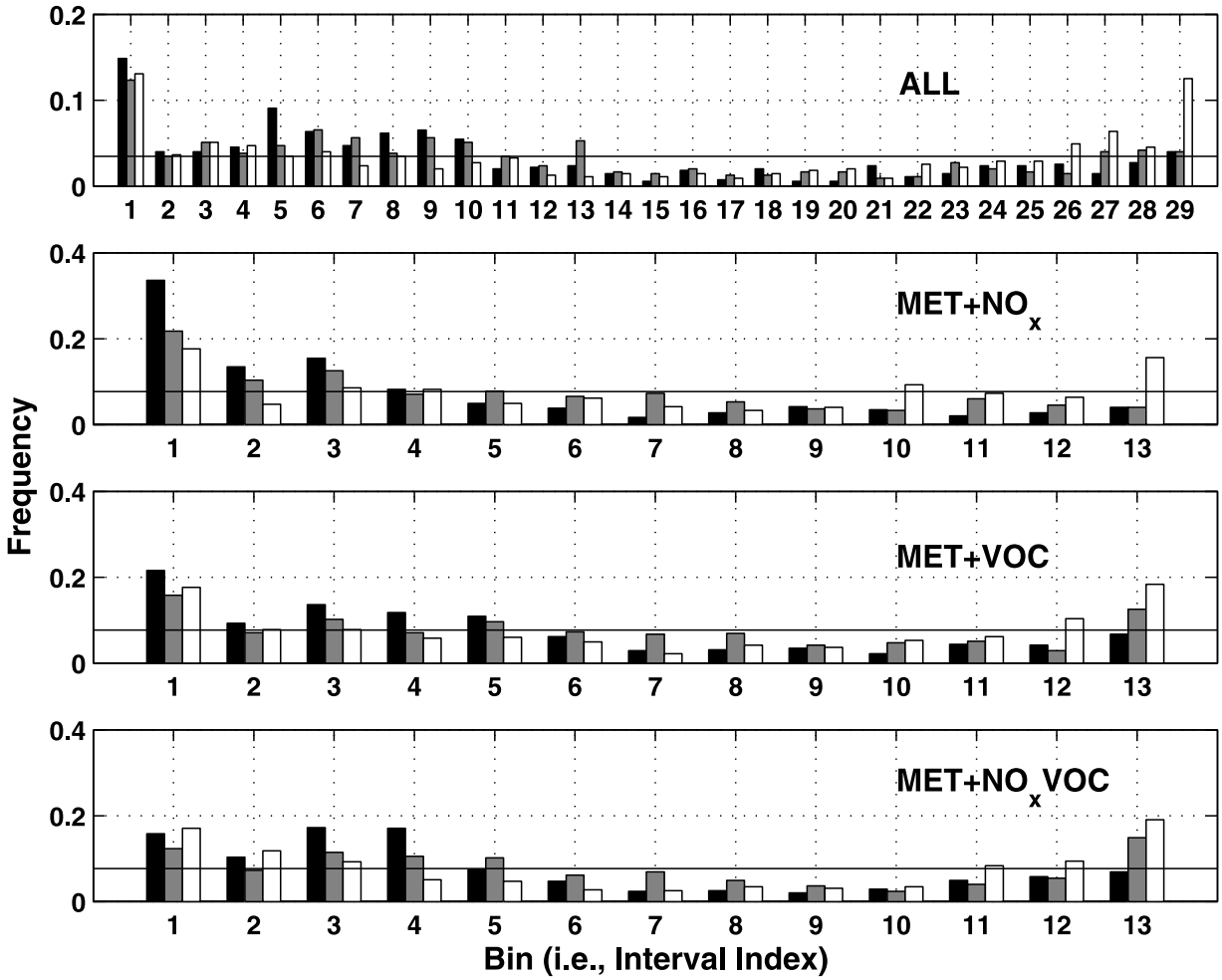

Figure 5. Talagrand diagram (rank histogram) for the ensembles generated by including both meteorology and emission perturbations (from top to bottom): ALL (28 members), MET+NO_x, MET+VOC, and MET+NO_xVOC (all three with 12 members). The number of bins equals the number of ensemble members plus one. Solid black bars are results for the raw forecasts, gray bars are results when the MET biases are removed, and open bars are fully bias-corrected results. Solid horizontal line represents the perfect Talagrand diagram shape (flat).

the same model. A summary of these analyses concludes this section.

4.1. Ensembles With Both Meteorology and Emission Perturbations

[28] The following are the ensembles generated by including both meteorology and emission perturbations: MET+NO_x, MET+VOC, MET+NO_xVOC (12 members each), and ALL (28 members). These ensembles will be referred to collectively as PERT. We expect ALL to outperform the other ensembles because it includes more sources of uncertainty. Including it here provides context for the individual emission perturbations and a measurement of how important each are to bias, ensemble spread, and distinguishing specific events.

[29] Because *RI* does not distinguish between bias and ensemble spread, further refinements can help with interpretation. To better understand the importance of multiple model versus emission perturbations, Talagrand diagrams are plotted for the raw ensemble forecasts, for ensembles with constituent forecasts adjusted by removing the bias from the associated base runs (i.e., those without perturbations to emissions), and for ensembles with constituent forecasts corrected for bias. The sample mean error is computed for each of the 28 possible members of any ensemble, and can be considered the bias for this experiment. This is removed from the forecasts before plotting the bias-free Talagrand diagram (open bars in Figure 5). The resulting *RI* score does not include any bias-induced ensemble spread. This step is desirable because ensemble members with different biases cannot forecast equally likely outcomes.

[30] To help isolate the effects of the emission perturbations, the bias from each of the four MET ensemble members is removed from the associated members that also have emissions perturbations (gray bars in Figure 5). For example, the bias from the 12-km MC2 run with the base CMAQ emissions is removed from all the members of the MC2-12 group. Removing this bias from the emission-perturbed runs shows how the emission perturbations, which are equal and opposite, evolve in the forecast. The resulting Talagrand diagram may contain biases that are not

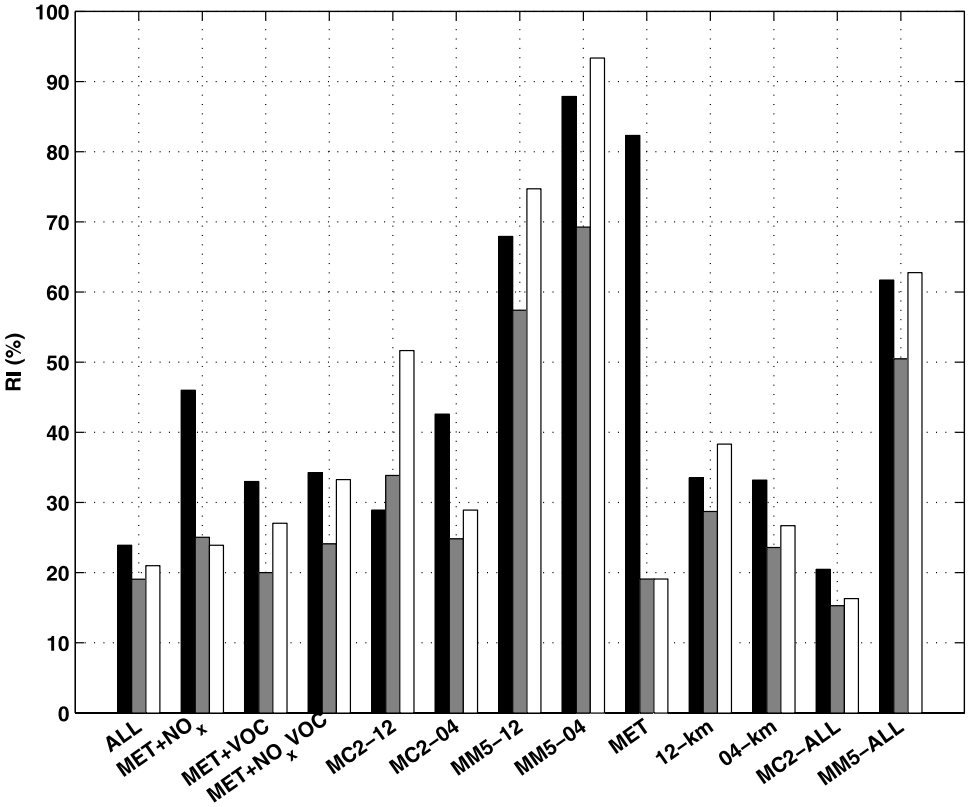

Figure 6. Reliability Index (*RI*) computed as in equation (2). Solid black bars are scores for the raw ensembles, gray bars are scores for the MET-adjusted ensembles, and open bars are scores for the bias-corrected ensembles. The closer *RI* is to 0, the more reliable the probabilistic forecast.

linear functions of the emissions perturbations, which may arise from nonlinear ozone chemistry and its interactions with the driving meteorological model. We will refer to these ensembles as "MET bias adjusted," as opposed to "bias corrected".

[31] Figure 5 shows the Talagrand diagram for the PERT ensembles. The solid horizontal lines indicate the ideal shape (for a perfectly reliable diagram). All the panels show a combination of a "U shape" and an "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. The leftmost bin for the raw ALL, MET+NO_x, and MET+VOC ensembles (black bars) contains an absolute frequency maximum, while the rightmost bin contains a relative frequency maximum. Furthermore, the asymmetric L shape (maximum in the first bin) indicates that the ensemble forecasts are biased toward overprediction of ozone concentrations. Adjusting for the MET bias reduces the overall ensemble bias, but does not remove it, showing that the choice of emissions perturbations leads to a biased ensemble (gray bars). Correcting for the biases of all the constituent members results in more symmetric diagrams (open bars). Reasons for the small remaining asymmetries could include some dependence between ensemble members and sampling error.

[32] Figure 6 shows the *RI* values of all the ensembles and is useful to assess the relative reliability, reliability resulting from emissions perturbations, and reliability due to unbiased ensemble spread. Among the PERT ensembles, ALL shows the least deficiency (in terms of reliability), followed by similar reliability for MET+NO_xVOC and MET+VOC. MET+NO_x shows the greatest positive bias among the four ensembles analyzed in this section, having the highest maximum in the first bin.

[33] Adjusting for the MET biases (gray bars) does reduce the overall *RI* for this group of ensembles, but does not result in an unbiased ensemble, again showing that the emissions perturbations lead to a biased ensemble. The unbiased *RI* (open bars) shows the contributions of the emissions perturbations to the ensemble spread. Perturbed NO_x demonstrates the most reliable spread of the three smaller ensembles when it is unbiased. Rather than promote ensemble spread, combining both NO_x and VOC perturbations leads to less spread than either precursor individually. We hypothesize that this is closely related to the predominant chemical regimes (i.e., NO_x sensitive or VOC sensitive).

[34] The MET+NO_x tendency to overestimate ozone concentrations would appear to suggest 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_xVOC also overestimate the measured ozone concentrations, giving the appearance that the same kind of perturbation shifting toward lower values could improve their forecast skill. However, it is impossible to say whether such a shift is realistic, or that it simply compensates for other errors in this coupled meteorological/AQ ensemble. Furthermore, if the error growth was linear then the MET bias adjusted ensembles would be unbiased,

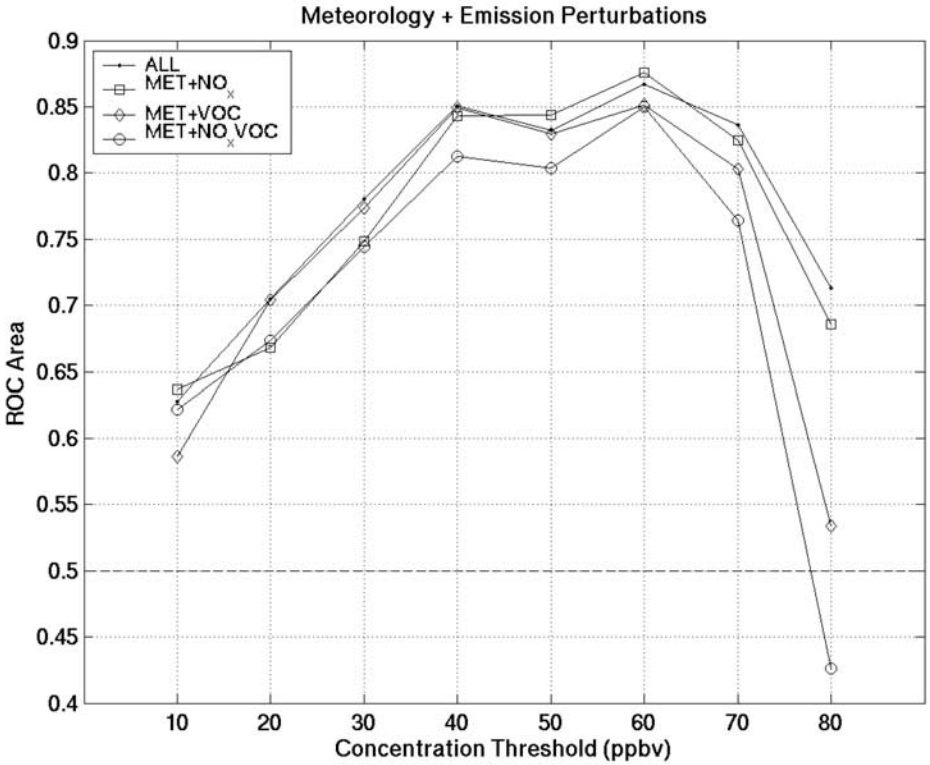

Figure 7. ROC area values for 10 different ozone concentration thresholds (from 10 to 80 ppbv, with increments of 10) and for the ensembles generated by including both meteorology and emission perturbations: ALL (28 members), MET+NO_x, MET+VOC, and MET+NO_xVOC (all three with 12 members). Values are within the interval [0, 1], with the perfect ROC area = 1, and a no-skill ROC area of 0.5 (dashed line).

because the emissions perturbations themselves are equal and opposite. Differences between the gray and open bars in Figure 6 show that some bias effects remain, suggesting that nonlinear effects in the ozone chemistry, and its interaction with the driving meteorological model, play an important role in error growth for this coupled model application.

[35] Figure 7 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 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.

[36] Even though MET+NO_x is the most biased ensemble in this group, it shows probabilistic predictive skill, as indicated by ROC values closest to ALL, and is 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 effective than the VOC (or VOC combined with NO_x) perturbations in spanning the emission uncertainty subspace with the least number of ensemble members. Because of this performance, we expect a bias-corrected MET+NO_x would be the most useful ensemble of this group, excepting ALL.

[37] The NO_x perturbation gives a better prediction of frequency of occurrence 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. In this study, NO_x sensitive means that a fixed percent change in NO_x results in a significantly greater change in ozone concentration relative to the same fixed percent change in VOC (similar but different definitions can be used, as discussed by *Sillman* [1999]). It is beyond the goal of this study to analyze in depth which are 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]) found the LFV to be VOC sensitive for the daily maximum.

[38] Nevertheless, the results of this study suggest a NO_x-sensitive chemistry regime at Hope for this particular

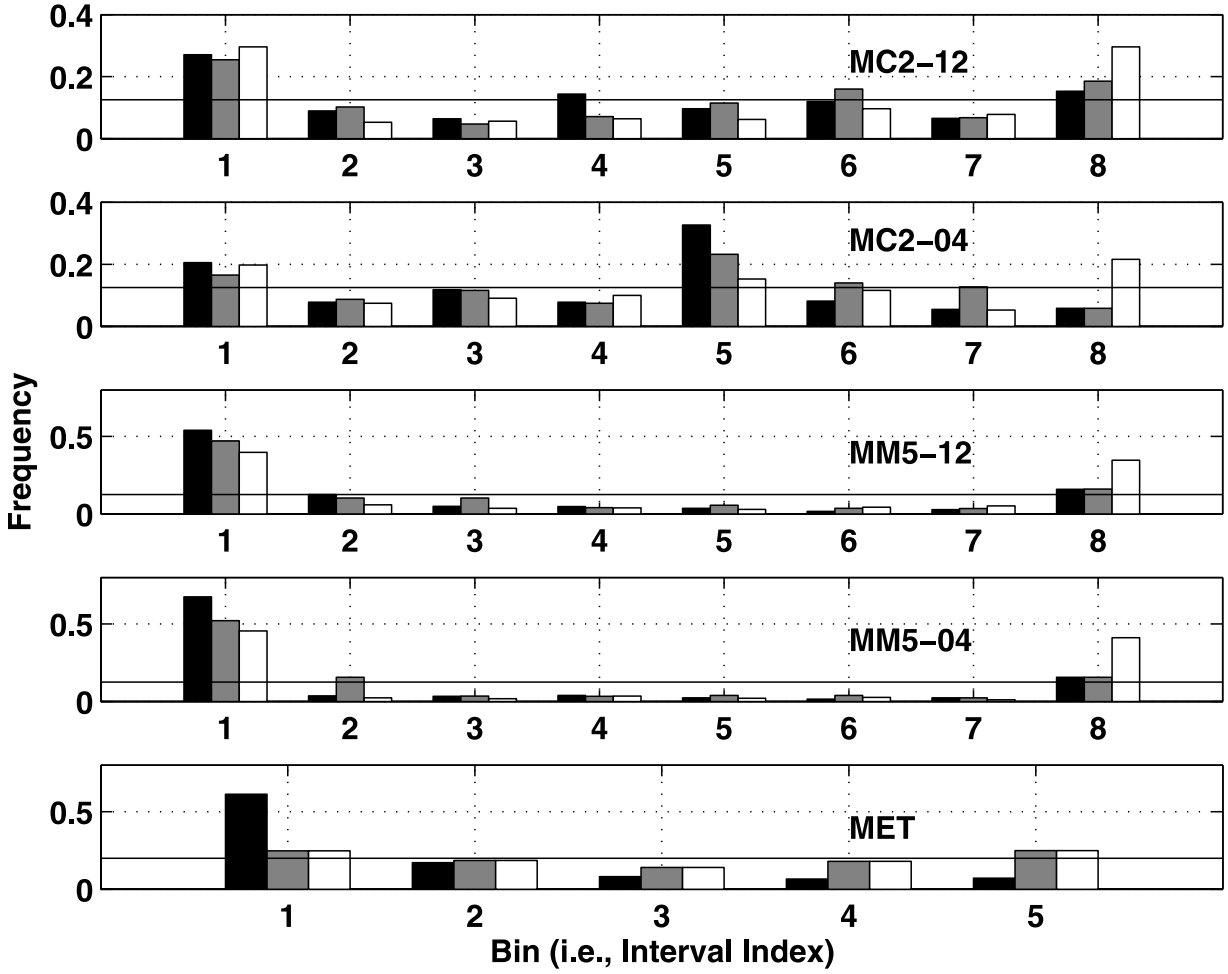

Figure 8. Similar to Figure 5, but for the ensembles generated with only emission perturbations. Namely, the ensembles are formed by forecasts driven with the same meteorological input (MC2-12, MC2-04, MM5-12, and MM5-04, all with seven members), or with only the meteorology perturbation (MET, four members).

11–15 August 2004 event, which can be explained as follows. The aged air mass from the Vancouver 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% values for NO_x are appropriate.

[39] On the basis of reliability and resolution metrics, ensemble ALL is the best forecast in this group, and MET+NO_x shows utility as a small ensemble. ALL demonstrates more reliable spread and less bias, indicated by the flatter Talagrand diagram, and more intrinsic value, indicated by the ROC curve. It is formed by the largest number of members (28) and therefore includes many more degrees of freedom than the others. The extra variability is associated with differences in the meteorological component, and can be expected. Ensemble MET+NO_x, though biased, shows high-ROC scores. Because the bias persists even when the base case mean error is removed, nonlinearity plays a role. A bias correction on each member of MET+NO_x individually improves the reliability without compromising the resolution (not shown). The next two subsections provide additional context for interpreting these emissions perturbations.

4.2. Ensembles With Only Meteorology or Emissions Perturbations

[40] 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 with MET, where only the 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 than for the emission perturbations.

[41] Figure 8 shows the Talagrand diagrams for these PFSs, where the solid lines have the same meaning as in Figure 5. For interpretation we again present Talagrand

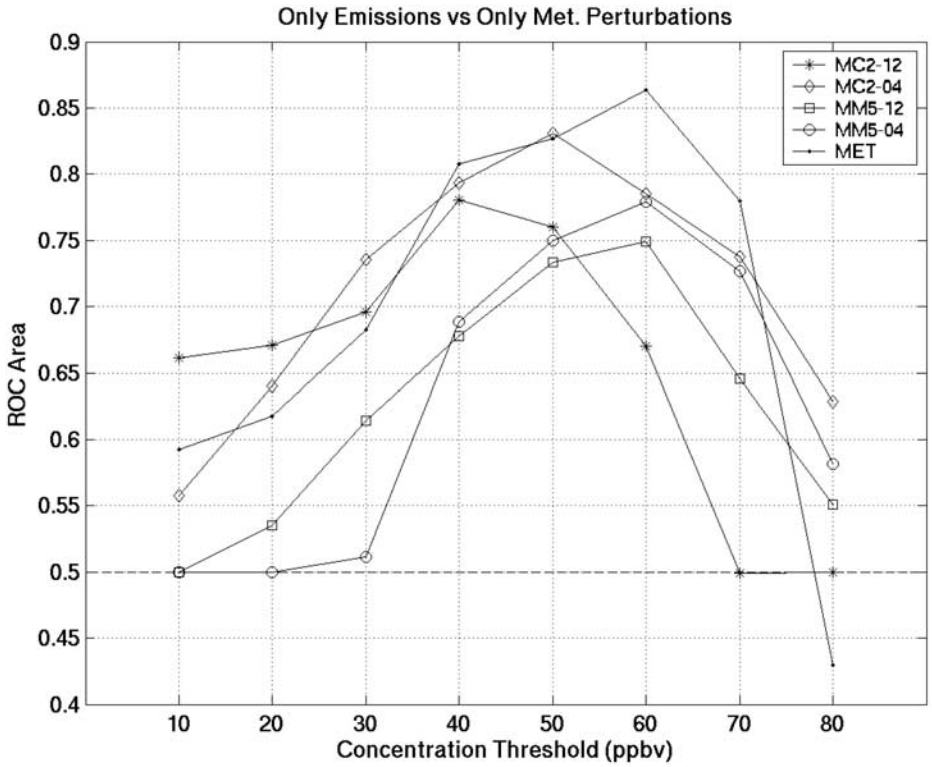

Figure 9. Similar to Figure 7, but with ROC area values for the ensembles generated with only emissions perturbations. Namely, the ensembles are formed by forecasts driven with the same meteorological input (MC2-12, MC2-04, MM5-12, and MM5-04, all with seven members), or with only the meteorology perturbation (MET, four members).

diagrams produced from the raw forecasts, from MET bias adjusted forecasts, and from bias-corrected forecasts. Similar to Figure 5, U-shaped and L-shaped diagrams are observed here. Note the open bars for the MET ensemble show no bias because of this correction, but the U shape indicates a clear underdispersion (i.e., not enough spread). A maximum frequency is observed for MC2-04 in the fifth bin, and to a lesser extent in the fourth bin for MC2-12. As with the PERT group of ensembles, MET bias adjusting the forecasts does not remove all of the bias (except for MET of course). Because each of these ensembles uses the same meteorological input, the remaining systematic errors result from the nonlinear chemistry and its response to the meteorological input.

[42] Overall, the raw MC2-12 has the third best *RI* value (29%), followed by the raw MC2-04 (43%). The two MM5 and the MET PFSs all have very high *RI* values (68% and 88% respectively), resulting in the worst overall performance in this group. The reason is that they are highly positively biased, as shown by the high frequency in the first bin in the Talagrand diagrams.

[43] The bias-corrected *RI* scores show that ensemble MC2-04 demonstrates the widest spread in this group, and that its raw *RI* score primarily results from bias. Its spread is within the range of the PERT ensembles. Conversely, ensembles MC2-12, MM5-12, and MM5-04 suffer from both bias and lack of spread. Bias correcting those ensembles results in higher *RI* scores, and suggests again that systematic behavior of the ozone chemistry is important to the raw ensemble results.

[44] Turning to the resolution (Figure 9), 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. 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.

[45] 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 for 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 topography is much more complex than at the other locations, resulting in an advantage for the finer horizontal resolution runs.

[46] By comparing Figures 7 and 9, the utility of the meteorology and emission perturbations, and their combination, can be inferred. The predictive skill of the PERT

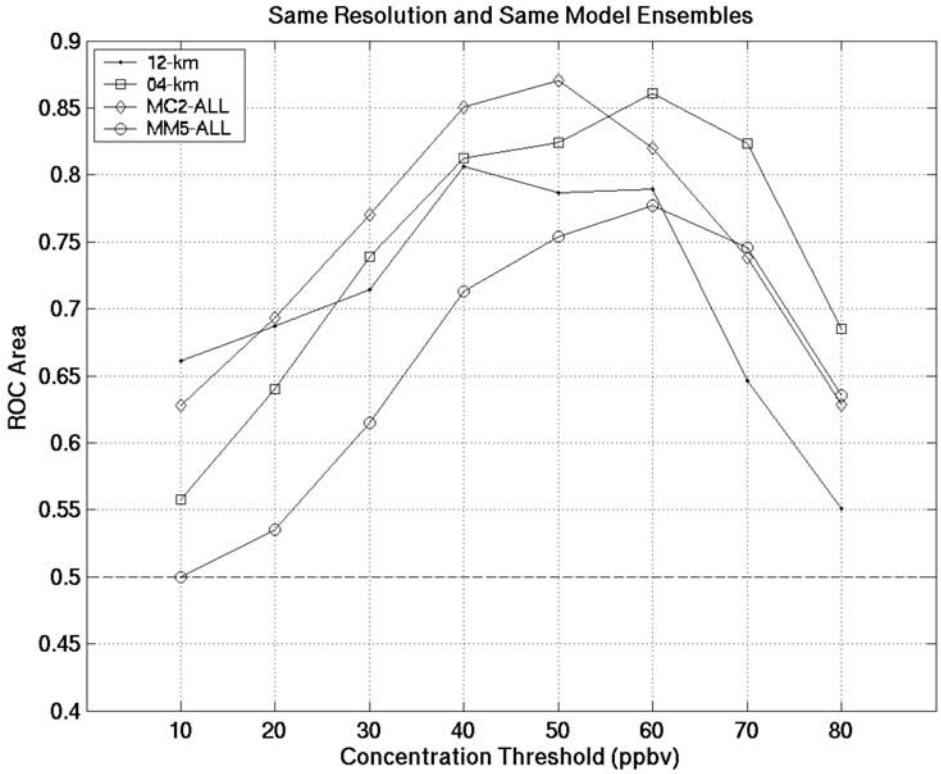

Figure 10. Similar to Figure 7, but for the ensembles formed with the same resolution runs (12 km and 04 km) or driven by the same numerical weather prediction model (MC2-ALL or MM5-ALL).

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.

[47] We can deduce the following from these results: both meteorology and emission perturbations are 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 for capturing ozone concentrations above 80 ppbv. We next examine specific effects of meteorological model differences.

4.3. Ensembles Generated With the Same Model or the Same Resolution

[48] Here the PFS resolution and reliability for 12 km, 04 km, MC2-ALL and MM5-ALL are analyzed (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. We do not present the Talagrand diagrams because their attributes can be deduced directly from Figures 5 and 8. *RI* scores in Figure 6 reinforce the conclusions found above. The MM5-based ensembles suffer from bias and underdispersion, and the higher-resolution ensembles show more reliable spread.

[49] Figure 10 shows the ROC areas 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 becomes more and more complex, giving a clear advantage to the finer resolution runs (as discussed in Section 4.2). Ensembles 04 km and MC2-ALL have high-ROC area values (above 0.8) between 40 and 70 ppbv, while 12 km is above 0.8 only for 40 ppbv. MM5-ALL always has a ROC area below approximately 0.78.

[50] 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 skill when higher ozone concentrations are measured, because it has better ROC area values but is less reliable.

4.4. Summary

[51] Figure 11 shows the ROC areas 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 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_xVOC. 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

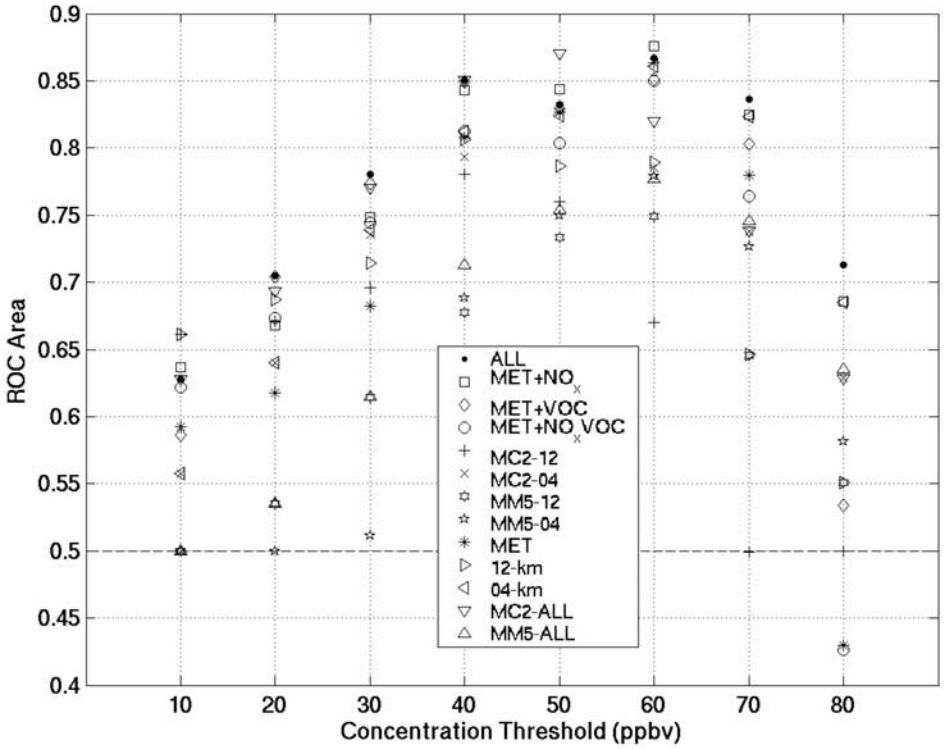

Figure 11. Similar to Figure 7, but for all 13 ensemble groups considered in this study: all the forecasts available (ALL, 28 members), meteorology and NO_x perturbations combined together (MET+NO_x, 12 members), meteorology and VOC perturbations (MET+VOC, 12 members), meteorology and NO_x combined with VOC perturbations (MET+NO_xVOC, 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), 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).

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 appear more important than the meteorology perturbations for capturing the highest ozone concentrations (above 80 ppbv).

[52] Excluding ALL from consideration, 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 for lower ozone concentrations. Instead, by looking at the Talagrand diagram, 04 km (Figure 10) is more reliable than MET+NO_x (Figure 5), which is one of the most positively biased PFSs. However, the MET+NO_x bias could be removed by Kalman filtering its forecasts (as shown in DM06b), resulting in a more reliable prediction.

[53] Revisiting the *RI* scores, the most reliable PFS is MC2-ALL, followed closely by ALL and then MC2-12. The small difference between them is likely within the noise level of this experiment. ALL benefits from the highest number of ensemble members, possibly making the extra computational effort worthwhile. Using ensemble MET as the baseline, ensemble spread is generally improved by the addition of ensemble members when the forecasts are not bias corrected. Conversely, bias-corrected forecasts result in a MET ensemble with spread among the most reliable presented here. Therefore the use of different meteorological models produces some variability that is not attributable to systematic ozone responses to those models.

[54] 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.

5. Conclusions

[55] This study investigates whether ensemble probabilistic ozone forecasts can be made with high verification resolution and reliability. To do this, 28 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-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. 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 air quality (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*, 1999].

[56] The following are the main findings for this one case study:

[57] 1. 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.

[58] 2. 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.

[59] 3. Nonlinear ozone chemistry and its response to different meteorological forcings play an important role that is not captured by varying the meteorology alone.

[60] 4. Correcting the forecasts for mean error significantly improves the reliability of forecasts with good spread characteristics, including the ensemble where meteorology is the only source of uncertainty spanned (MET).

[61] 5. Among the emission perturbations, NO_x perturbations resulted in more skillful probabilistic forecasts for the episode analyzed in this study.

[62] 6. Since NO_x perturbations lead to (positively biased) predictive skill, the ±50% values appear to effectively span the emission uncertainties space for this case.

[63] 7. The finer spatial resolution runs have better predictive skill (but similar reliability) than the coarser runs, particularly in the eastern end of the LFV where the topography progressively becomes more complex.

[64] 8. The MC2 model leads to more ozone variability and better predictive skill than the MM5 in the 5-day period analyzed in this study.

[65] 9. The ALL ensemble (formed by all the 28 ozone forecasts available) is the best probabilistic forecast, when considering both reliability and resolution. Ensembles 04 km and MET+NO_x closely follow.

[66] The results of this study suggest that future work should focus on ozone ensemble forecast systems 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 (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 perturbation for capturing the ozone variability. Conversely, in VOC-sensitive regimes the VOC perturbations could be more effective. In situations where neither of these two regimes is well defined, a combination of NO_x and VOC perturbations may 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].

[67] Here we found some indication that nonlinear ozone chemistry can result in systematic forecast errors, exposing a complex relationship between perturbations to ozone precursors and meteorological drivers. These relationships should be studied further to refine ensemble strategies.

[68] Ideally, each ensemble member should represent 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, 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 or chemistry is their 12-km coarser domain.

[69] 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.

[70] Finally, ensemble weather forecasts often provide information on the uncertainty of the forecasts; if the ensemble members have a large spread, one expects more uncertainty in the forecast. However, similar to *Delle Monache et al.* [2006a, 2006b], no correlation or relationship between ensemble spread and forecast error was found in this study. Much longer experiments, covering many events, would be necessary to evaluate this.

[71] **Acknowledgments.** We thank George Hicks II, 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. We are thankful to Giovanni Leoncini (Colorado State University), Bruce Ainslie, Phil Austin, Allan Bertram, Ian McKendry, and Douw Steyn (University of British Columbia), and Brian Lamb (Washington State University) for carefully reviewing the paper. Also, Bruce Ainslie kindly provided the ozone isopleths in Figure 1. Grant support came from the Canadian Natural Science and Engineering Research Council, the BC Forest Investment Account, the British Columbia 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 providing useful 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, 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*, Rapra Technol., 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, *Rep. EPA/600/R-99/030*, Off. of Res. and Dev., Environ. Protect. Agency, Washington, D. C.

Candille, G., and O. Talagrand (2005), Evaluation of probabilistic prediction system for a scalar variable, *Q. J. R. Meteorol. Soc*, *131*, 2131–2150.

Coats, C. J., Jr. (1996), High-performance algorithms in the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system paper presented at 9th 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.

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 International Technical Meeting on Air Pollution Modeling and its Applications, Eur. Assoc. for the Sci. of Air Pollut., Banff, Alberta, 25–29 Oct.

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

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

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005), Calibrated probabilistic forecasts using ensemble model output statistics and minimum CRPS estimation, *Mon. Weather Rev.*, *133*, 1098–1118.

Greater Vancouver Regional District (2002), 2000 emissions inventory for the Lower Fraser valley airshed, technical report, Greater Vancouver Reg. Dist. 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 TN-398+STR*, Natl. Cent. for Atmos. Res., Boulder, Colo.

Hamill, T., and S. J. Colucci (1997), Verification of Eta–RSM short-range ensemble forecasts, *Mon. Weather Rev.*, *125*, 1312–1327.

Horvath, S. M., and D. J. McKee (1994), Acute and chronic health effects of ozone, in *Tropospheric Ozone, Human Health and Agricultural Aspects*, edited by D. J. McKee, pp. 39–84, CRC Press, Boca Raton, Fla.

Hou, D., E. Kalnay, and K. K. Droegemeier (1998), Objective verification of the SAMEX '98 ensemble forecasts, *Mon. Weather Rev.*, *129*, 73–91.

Jolliffe, I. T., and D. B. Stephenson (2003), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, p. 240, John Wiley, Hoboken, N. J.

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

Mason, I. (1982), A model for assessment of weather forecasts, *Aust. Meteorol. Mag.*, *30*, 291–303.

Mason, S. J., and N. E. Graham (1999), Conditional probabilities, relative operating characteristics, and relative operative levels, *Weather 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, *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.

Pryor, S. C. (1998), A case study of emission changes and ozone responses, *Atmos. Environ.*, *32*, 123–131.

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

Sillman, S. (1999), The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments, *Atmos. Environ.*, *33*, 1821–1845.

Sillman, S., and D. He (2002), Some theoretical results concerning O_3–NO_x-VOC chemistry and NO_x-VOC indicators, *J. Geophys. Res.*, *107*(D22), 4659, doi:10.1029/2001JD001123.

Talagrand, O., and R. Vautard (1997), Evaluation of probabilistic prediction systems, in *Proceedings ECMWF Workshop on Predictability*, pp. 1–25. Eur. Cent. for Medium-Range Weather Forecasts, Reading, U. K.

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

L. Delle Monache, Lawrence Livermore National Laboratory, 7000 East Avenue, L-103 Livermore, CA 94551, USA.

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

J. P. Hacker, Research Applications Laboratory, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000, USA.

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

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