Available online at www.sciencedirect.com

SCIENCE DIRECT®

Atmospheric Environment 40 (2006) 4894–4905

ATMOSPHERIC ENVIRONMENT

www.elsevier.com/locate/atmosenv

An operational evaluation of the Eta–CMAQ air quality forecast model

Brian Eder*,[1], Daiwen Kang[2], Rohit Mathur[1], Shaocai Yu[2], Ken Schere[1]

ASMD, ARL, NOAA, Mail Drop E-243-01, Research Triangle Park, NC 27711, USA

Received 10 March 2005; received in revised form 29 July 2005; accepted 22 December 2005

Abstract

The National Oceanic and Atmospheric Administration (NOAA), in partnership with the United States Environmental Protection Agency (EPA), are developing an operational, nationwide Air Quality Forecasting (AQF) system. An experimental phase of this program, which couples NOAA's Eta meteorological model with EPA's Community Multiscale Air Quality (CMAQ) model, began operation in June of 2004 and has been providing forecasts of ozone (O_3) concentrations over the northeastern United States. An important component of this AQF system has been the development and implementation of an evaluation protocol. Accordingly, a suite of statistical metrics that facilitates evaluation of both *discrete-* and *categorical-type* forecasts was developed and applied to the system in order to characterize its performance. The results reveal that the AQF system performed reasonably well in this inaugural season (mean domain wide correlation coefficient = 0.59), despite anomalously cool and wet conditions that were not conducive to the formation of O_3. Due in part to these conditions, the AQF system overpredicted concentrations, resulting in a mean bias of +10.2 ppb (normalized mean bias = +22.8%). In terms of error, the domain-wide root mean square error averaged 15.7 ppb (normalized mean error = 28.1%) for the period. Examination of the discrete and categorical metrics on a daily basis revealed that the AQF system's level of performance was closely related to the synoptic-scale meteorology impacting the domain. The model performed very well during periods when anticyclones, characterized by clear skies, dominated. Conversely, periods characterized by extensive cloud associated with fronts and/or cyclones, resulted in poor model performance. Subsequent analysis revealed that factors associated with CMAQ's cloud cover scheme contributed to this overprediction. Accordingly, changes to the cloud schemes are currently underway that are expected to significantly improve the AQF system's performance in anticipation of its second year of operation.
Published by Elsevier Ltd.

Keywords: Air quality forecasting; Ozone; Model evaluation; Community Multiscale Air Quality (CMAQ) model; Eta model

1. Introduction

Although air quality has improved significantly in the decades following passage of the Clean Air Act (1970), there are still many areas in the United States where the public is exposed to unhealthy levels of air pollutants, most notably ozone (O_3) and fine particulate matter. The cost of poor air quality

*Corresponding author. Tel.: +1 919 541 3994; fax: +1 919 541 1379.
E-mail address: eder@hpcc.epa.gov (B. Eder).

[1] In partnership with the NERL, US Environmental Protection Agency, RTP, NC 27711.

[2] On assignment from Science and Technology Corporation, Hampton, VA 23666.

1352-2310/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.atmosenv.2005.12.062

to the United States from pollution-related illnesses alone has been estimated at 150 billion dollars (http://nws.noaa.gov/ost/air-quality). For many citizens, especially those who suffer from respiratory problems, the availability of air quality forecasts (AQF), analogous to weather forecasts, could make a significant difference in how they plan their daily activities and in turn improve the quality of their lives. It has been estimated that for each 1% reduction in adverse health effects that an AQF could provide, over 1 billion dollars could be saved annually in medical expenses (http://nws.noaa.gov/ost/air-quality).

To help the Nation realize such benefits, the Congress directed the National Oceanic and Atmospheric Administration's (NOAA) to provide National AQF guidance via the *H.R. Energy Policy Act of 2002* (Senate Amendment) S. 517, SA1383, Forecasts and Warnings. Accordingly, NOAA which has, as one of its core missions, weather prediction, or more generally, environmental prediction, has entered into a partnership with the United States Environmental Protection Agency (EPA), which has as one of its core missions the protection of human health and welfare, to develop a real-time nationwide AQF system (Otte et al., 2005). This AQF system is intended to provide local and State agencies with forecast guidance, thereby supplementing, rather than replacing, the numerous and varied techniques that they have employed through the years. A brief history of these techniques, which vary greatly in levels of sophistication, can be found in US EPA (2003).

The initial phase of this AQF system, which couples NOAA's Eta meteorological model with EPA's Models-3 Community Multiscale Air Quality (CMAQ) model (Byun and Ching, 1999), began experimental operation during the summer of 2003. An updated version, incorporating numerous refinements developed by NOAA's ARL, began operation in June of 2004 and provided forecasts of both peak 1- and 8-h ozone concentrations for the northeast quadrant of the United States. Beginning in September of 2004, these forecasts were officially disseminated to the general public via NOAA's website: http://www.nws.noaa.gov/aq/.

The purpose of this paper is to provide the first (of an expected series) of evaluations that characterize the performance the AQF system for the summer using a suite of metrics established in Kang et al. (2003). This evaluation examines the performance of both *discrete forecasts* (observed versus modeled concentrations) for hourly, peak 1- and 8-h O_3 concentrations and *categorical forecasts* (observed versus modeled exceedances/non-exceedances) for both the peak 1- (125 ppb) and 8-h (85 ppb) National Ambient Air Quality Standards as established by the Clean Air Act and its amendments. This evaluation covered a 4-month period (1 June to 30 September, 2004) and used O_3 concentration measurements obtained from EPA's AIRNow network (http://www.airnow.gov/). In addition to the metrics presented in Kang et al. (2005), one new categorical metric, called the Weighted Success Index (WSI), is introduced that provides a more representative measure of model performance.

2. Description of the modeling system

The Eta–CMAQ AQF system is based on the National Centers for Environmental Prediction's (NCEP's) Eta model (Black, 1994; Rogers et al., 1996) and EPA's CMAQ Modeling System (Byun and Ching, 1999). A brief summary of the linkage between the Eta and the CMAQ models, relevant to this study, is presented below. A more in-depth description can be found in Otte et al. (2005).

The Eta model is used to prepare the meteorological fields for input to the CMAQ. The NCEP Product Generator software is used to perform bilinear interpolations and nearest-neighbor mappings of the Eta Post-processor output from Eta forecast domain to the CMAQ forecast domain. The processing of the emission data for various pollutant sources has been adapted from the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system (Houyoux et al., 2000) on the basis of the US EPA national emission inventory. The Carbon Bond chemical mechanism (version 4.2) is used to represent the photochemical simulations.

Detailed information on transport and cloud processes in the CMAQ is described in Byun and Ching (1999). For this application, O_3 concentrations are forecast over the Northeast US using a 12-km horizontal grid spacing on a Lambert Conformal map projection. There are 22 layers in the vertical domain, which are set on a sigma coordinate extending from the surface to approximately 100 hPa. Vertically varying lateral boundary conditions for O_3 are derived from daily forecasts of the Global Forecast System (GFS). The initial condition chemical fields for CMAQ are initiated using the previous forecast cycle. The Eta 12 UTC cycles are used for the forecast cycle (Otte et al., 2005). The primary Eta–CMAQ model forecast for next-day surface-layer O_3 is based on the current day's 12 UTC Eta cycle. The target forecast period is local midnight through local midnight (04 to 03 UTC for the Northeast US). An additional 8 h are required beyond midnight to calculate peak 8-h average O_3 concentrations. As a result, a 48-h Eta–CMAQ forecast is needed (based on the 12 UTC initialization) to obtain the desired 24-h forecast period.

3. O_3 data

Hourly, near real-time, O_3 (ppb) data obtained from EPA's AIRNow program are used in the evaluation (http://www.epa.gov/airnow). Over 600 stations are available (Fig. 1) resulting in nearly 2 million total hourly O_3 observations for the study period. In addition to the hourly data, both the peak 1- and 8-h concentrations are calculated for each station and each day over the 4-evaluation period. The calculation of the 8-h peak is the same as the model forecast using the forward calculation method (i.e. calculation of the last seven 8-h peak concentrations including data from next day). The peak 1- and 8-h concentrations are considered missing if half of the hourly observation data are missing for the day. If two or more monitoring stations are located within the same model grid cell, their average value is used as the representative measurement for that grid cell.

4. Statistics

A description of the various discrete and categorical statistical metrics used in this evaluation is presented below, including a newly designed metric, called the WSI.

4.1. Discrete statistics

For the *discrete forecast* evaluation, basic summary statistics along with two standard and widely used measures of *bias*: the Mean Bias (MB) and the Normalized Mean Bias (NMB); and *error*: the Root Mean Square Error (RMSE) and Normalized Mean Error (NME) were selected and are defined below:

$$\mathrm{MB} = \frac{1}{N}\sum_{1}^{N}(C_\mathrm{m} - C_\mathrm{o}), \quad (1)$$

$$\mathrm{NMB} = \frac{\sum_1^N (C_\mathrm{m} - C_\mathrm{o})}{\sum_1^N C_\mathrm{o}}\,100\%, \quad (2)$$

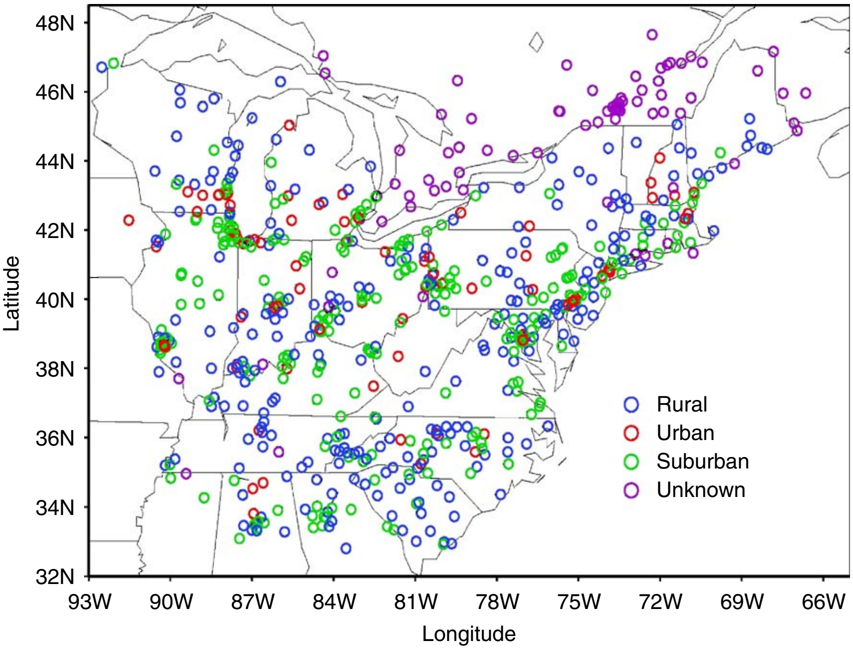


Fig. 1. Modeling domain and AIRNow monitoring locations (with land-use type denoted).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{1}^{N}(C_\mathrm{m} - C_\mathrm{o})^2}, \quad (3)$$

$$\mathrm{NME} = \frac{\sum_1^N |C_\mathrm{m} - C_\mathrm{o}|}{\sum_1^N C_\mathrm{o}} 100\%, \quad (4)$$

where C_m and C_o are modeled and observed concentrations, respectively.

4.2. Categorical statistics

For the *categorical forecast* evaluation, the models' Accuracy (A), Bias (B), Hit Rate (H), False Alarm Ratio (F), and Critical Success Index (CSI) were calculated (Jolliffe and Stephenson, 2003). These metrics were based on the observed exceedances, non-exceedances versus forecast exceedance, non-exceedances for both the 1- and 8-h O_3 standard. Fig. 2 provides a graphical representation of the variables (a, b, c and d) that represent the number of data points within each quadrant used to formulate the categorical metrics. Specifically, a represents all of the forecast 8-h exceedances (>85 ppb) that did not occur, b represents all of the forecast 8-h exceedance that did occur, c all of the forecast 8-h non-exceedances that did not occur and d all of the non-forecast 8-h exceedances that did occur.

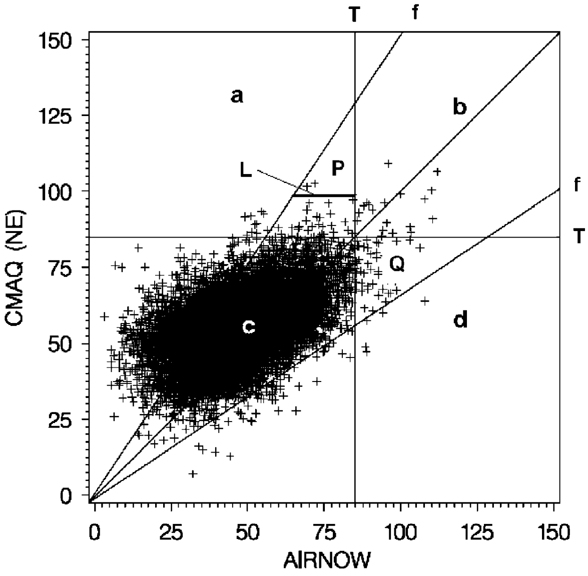


Fig. 2. Example scatter plot for the categorical evaluation: a denotes a forecast 8-h exceedance (>85 ppb) that did not occur (false positive); b a forecast 8-h exceedance that did occur; c a forecast 8-h non-exceedance that did not occur; and d a non forecast 8-h exceedance that did occur (false negative); T denotes threshold values; f denotes factor of 1.5 above (below) the 1:1 line; L denotes distance between factor and threshold; and P and Q denote triangles used in calculation of the Weighted Success Index (WSI).

Accuracy (A) measures the percentage of forecasts that correctly predict an exceedance or non exceedance and is given by

$$A = \left(\frac{b + c}{a + b + c + d}\right) 100\%. \quad (5)$$

As will be discussed in Section 5.2, A is strongly influenced by the number of correctly forecast non-exceedances (c), which is invariably very large, hence care must be taken in its interpretation. The Bias (B) indicates on average, if the forecasts are underpredicted (false negative) or overpredicted (false positives):

$$B = \left(\frac{a + b}{b + d}\right). \quad (6)$$

A value of 1.0 would indicate no bias, values <1.0 indicate underprediction and values >1.0 indicate overprediction. The False Alarm Ratio (FAR) measures the percentage of times an exceedance was forecast and did not occur:

$$\mathrm{FAR} = \left(\frac{a}{a + b}\right) 100\%. \quad (7)$$

Smaller numbers are of course desirable, with a FAR = 0 indicating no false alarms, and a FAR of 50% indicating that half of the forecast exceedances did not actually occur. The CSI indicates how well both forecast exceedances and actual exceedances were predicted:

$$\mathrm{CSI} = \left(\frac{b}{a + b + d}\right) 100\%. \quad (8)$$

Unlike the A, the CSI is not affected by a large number of correctly forecast non-exceedances. A CSI of 50% would indicate that half of the forecasted and actual exceedances were correct. Finally, the Hit Rate (H), which is similar to the CSI, indicates the percentage of forecast exceedances that actually occurred. It is also called Probability Of Detection (POD):

$$H = \left(\frac{b}{b + d}\right) 100\%. \quad (9)$$

4.3. Weighted success index

The categorical statistics discussed above are defined by the numbers of paired data points found in the quadrants defined by the threshold lines as shown in Fig. 2. While informative, these metrics are not infallible in that they do not always represent the model's performance accurately. As an illustration of their limitations, consider $x(O,M)$ representing a paired data point (M is the modeled value, O the observed) that lies within area a (forecast exceedance that did not occur) but also lies within a factor line f inside a triangle designated as P. This individual forecast, though considered a "failure" or false alarm from a categorical standpoint, in actuality, represents a "success" from a discrete standpoint. The same is true for points falling into area d, but within the lower factor line (triangle Q). Accordingly, a new metric is proposed, called the WSI that gives some credit for points located in the triangles P and Q, while penalizing points in area a and d, but outside the triangles. The value of the factors (f) used to determine successful model performance, while arbitrary, were set to 1.5 in this example. The threshold lines (T) mark the exceedance values for both observed and forecast O_3 concentrations (85 ppb for 8-h peak O_3).

If a data point $x(O, M)$ is within triangle P, the length of the line that passes through x and intercepts with both the threshold line (T) and factor line (L) can be computed as

$$L = T - \frac{1}{f}M. \tag{10}$$

L can then be used to define

$$\mathrm{WSM} = 1 - \frac{T-O}{L} = 1 - \frac{T-O}{T-M/f} = \frac{M-fO}{M-fT}. \tag{11}$$

The values of WSM are between 0 and 1 for points within the factor lines. For points outside the factor lines, the values of WSM are negative and their magnitude is dependent on the factor value (but limited to -1 for symmetry and to prevent outliers from dominating the weighting). Similarly for a point in triangle Q, that is observed exceedance but not forecasted, we have

$$\mathrm{WSO} = \frac{O-fM}{O-fT}. \tag{12}$$

WSM and WSO are then used to calculate the WSI as seen below:

$$\mathrm{WSI} = \frac{b + \sum \mathrm{WSM} + \sum \mathrm{WSO}}{a+b+d} 100\%. \tag{13}$$

Values of WSI range from -100% (worst possible forecast) to 100% (perfect forecast).

5. Performance results

The meteorological conditions during the summer of 2004 in the northeast quadrant of the US were unusually cool and wet. More specifically, the vast majority of the states within the domain experienced temperatures either below or much below normal, and precipitation either above or much above normal (http://www.ncdc.noaa.gov/). Because of these anomalous conditions, very few O_3 "episodes" or exceedances occurred during the 4-month period, thereby limiting the efficacy of the evaluation, especially from a categorical standpoint. Because the performance results were very similar for each O_3 value (i.e. hourly, peak 1- and 8-h) and for the sake of brevity, the results presented below focus mainly on the peak 8-h concentrations.

5.1. Discrete evaluation

Examination of Table 1, which provides a summary of discrete statistics for the peak 8-h O_3 forecasts, reveals that O_3 concentrations observed throughout the full domain and 4-month time period were indeed very low, reflecting the non-conducive meteorological conditions discussed

Table 1
Summary of discrete statistics for peak 8-h O_3 forecast

Month	Obs. mean	Mod. mean	r	MB (ppb)	NMB (%)	RMSE (ppb)	NME (%)
June	46.1	53.9	0.51	7.8	16.9	14.1	24.0
July	47.4	57.3	0.55	9.9	20.8	16.2	27.2
August	43.4	55.1	0.62	11.7	27.1	16.5	30.9
September	41.7	52.8	0.65	11.1	26.7	15.6	30.6
All	44.6	54.8	0.59	10.2	22.8	15.7	28.1

above. Due, at least in part, to these non-conducive conditions, the AQF modeling system systematically overpredicted the 8-h O_3 concentrations for this period. The mean modeled value of 54.8 ppb resulted in a "season" long mean bias of 10.2 ppb (NMB = 22.8%). When examined monthly, the bias vary from a low of 7.8 ppb (NMB = 16.9%) during the month of June to a high of 11.7 ppb (NMB = 27.1%) for August. Biases and errors associated with the peak 1-h concentrations (not shown) follow a similar pattern. The errors associated with the forecasts were slightly larger. The RMSE (NMB) averaged 15.7 ppb (28.1%) for the season and ranged from 14.1 ppb (24.0%) in June to 16.5 ppb (30.9%) in August. The correlation coefficient (r) averaged 0.59 for the season and ranged from 0.51 in June to 0.65 in August. These results of the AQF system are comparable to those found in the NOAA sponsored "New England Forecasting Pilot Program", which enlisted three different regional-scale air quality models, serving as prototypes, to forecast O_3 concentrations across the northeastern United States during the summer of 2002 (Kang et al., 2005).

Additional insight into the AQF modeling system's positive bias (overprediction) and error (scatter) can be gained from Fig. 3, which includes a scatter plot of the model forecasts versus AIRNow observations for the peak 8-h concentrations (panel a); and a boxplot of the diurnal variation of the hourly concentration bias (panel b). In the scatter plot the vast majority of O_3 forecasts (ordinate) fall within a factor of 1.5 of the observations (abscissa). It also reveals that much of the overprediction discussed above occurs when the observed O_3 concentrations are relatively small (<50 ppb), which typically coincides with non-conducive (i.e. cloud cover, precipitation and cool temperatures) meteorological conditions. The boxplot (panel b) depicts the distribution of model bias (75th, 50th, 25th percentiles, max., min. and mean) throughout the diurnal cycle. Although the model overpredicts throughout the diurnal period, the positive bias is more prevalent at night, due in large part to the model system's difficulty in simulating the evolution of the nocturnal boundary layer and its impact on surface O_3 concentrations.

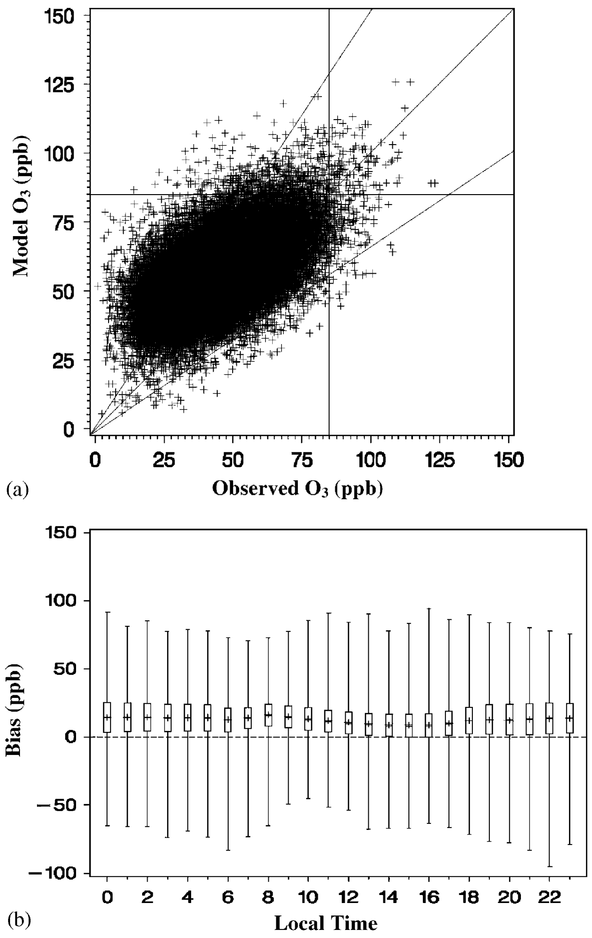


Fig. 3. Overall performance of the forecast system during the 4-month period (June–September, 2004). (a) Scatter plot of daily peak 8-h O_3 concentrations (reference lines are factors of 1.5 above and below the 1:1 line); (b) boxplot for diurnal variation of hourly bias (model–observation) showing 75th, 50th, 25th percentiles, max., min. and mean.

5.1.1. Temporal

In order to investigate the AQF system's performance over time, several of the discrete statistics discussed above were calculated (domain-wide averages) and plotted as a daily time series (Figs. 4 and 5). Fig. 4 displays the observed and forecast daily peak 8-h O_3 concentrations, as well as the corresponding RMSE and MB. Although the forecasts tracked the general temporal pattern well, the overprediction discussed above was prevalent throughout the 4-month period. This resulted in continuously positive MBs (with the exception of 1 day (9/23)) that generally ranged between 5 and 15 ppb. The RMSEs were slightly larger, generally ranging between 10 and 20 ppb. The daily correlation coefficients generally fluctuated between 0.4 and 0.8. There were, however, three notable exceptions occurring on 12 June ($r = -0.18$), 13 July ($r = 0.01$) and 8 September ($r = 0.12$). On each

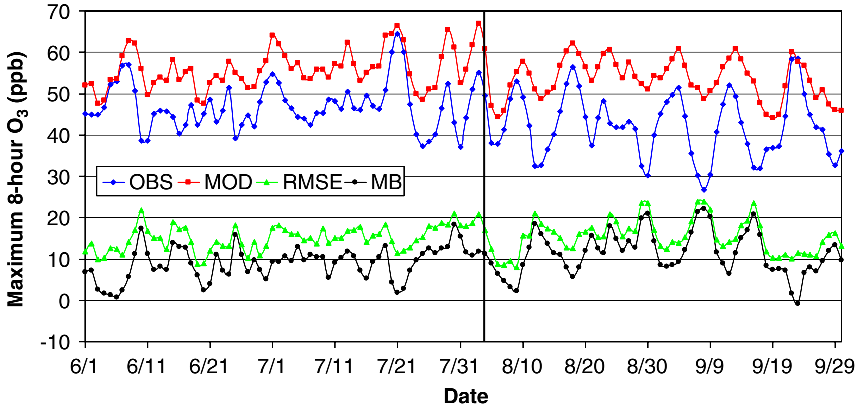


Fig. 4. Daily time series of the modeled and observed peak 8-h O_3 concentrations and the resulting model mean bias and root mean square error (4 August highlighted for case study discussion in Section 5.3).

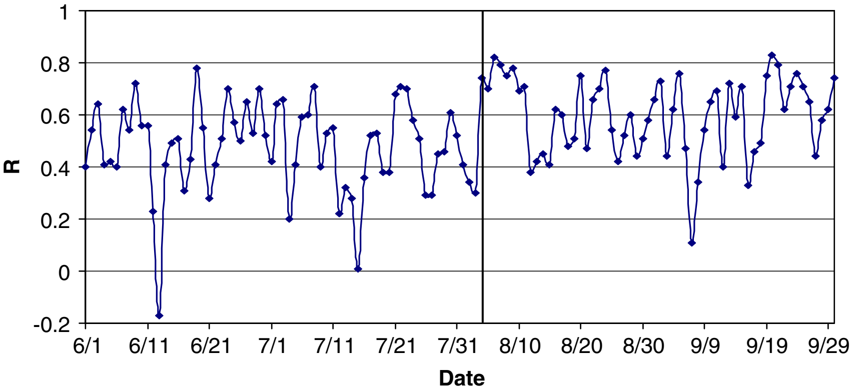


Fig. 5. Daily time series of the correlations between the modeled and observed peak 8-h O_3 concentrations (4 August highlighted for case study discussion in Section 5.3).

of these 3 days, the majority of the domain was covered in extensive convective cloud cover and received heavy precipitation, resulting in low observed O_3 concentrations that the model was unable to replicate.

Closer examination of the time series reveals a systematic pattern of varied modeled performance (i.e. several days of good performance characterized by $r > 0.60$, MB < 10.0 ppb, RMSE < 15.00 ppb, followed by several days of poor performance) that can be traced back to the "synoptic-scale" meteorology impacting the domain during the 4-month period. During days when anticyclones, characterized by clear skies and little or no precipitation (conditions conducive to O_3 formation) dominated the domain, the model performed very well; days characterized by extensive cloud cover and precipitation (conditions not conducive to O_3 formation) associated with fronts and/or cyclones, resulted in poor model performance. A case study is presented in Section 5.3 that will provide a closer examination of these performance characteristics.

5.1.2. *Spatial*

In order to investigate the performance of the AQF system over space, the correlation coefficients and NMB for the peak 8-h forecast were calculated (4 months mean) and plotted across the model domain (Fig. 6). In terms of correlation (top panel), the model performed better ($0.50 \leqslant r \leqslant 0.75$) along a band stretching from the piedmont regions of Alabama, Georgia and the Carolinas, north and east through the northeast corridor to the coast of Maine. Equally good performance is found within the area surrounding the Great Lakes. Three, fairly distinct areas of poorer performance

($0.25 \leqslant r \leqslant 0.50$) are also evident in the figure, including an area surrounding the St. Lawrence River in Quebec (that stretches eastward into northern Maine) and a large area encompassing the Ohio River Valley. The anomalously cool and wet conditions discussed earlier were especially prevalent in these regions and may be responsible for the poorer performance. Poor model performance ($r < 0.50$) was also found along the Appalachian Mountains from Pennsylvania southwest into South Carolina, with several locations in western North Carolina recording correlations less than 0.25. These low values are likely attributable to the elevated and complex terrain found throughout this region and the AQF system's inability to capture such variable terrain when using 12 km horizontal resolution.

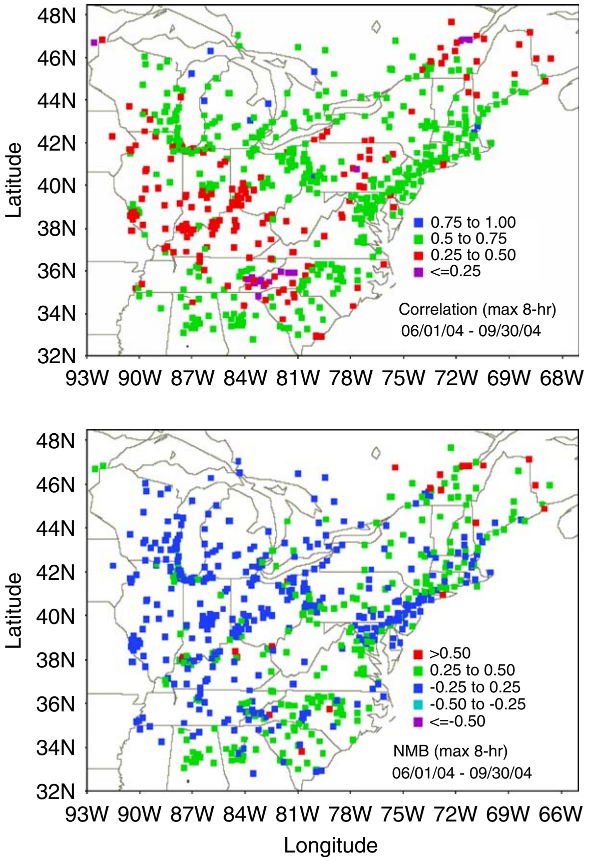


Fig. 6. Spatial distribution of correlation coefficient (top panel), and normalized mean bias (NMB) (bottom panel) associated with the peak 8-h O_3 concentrations.

The AQF system's tendency to generally overpredict O_3 concentrations is also seen in Fig. 6 (bottom panel), which depicts the spatial distribution of NMBs. This figure reveals that roughly half of the domain has NMB exceeding 25%. One area of concentrated positive bias is found stretching from central Pennsylvania northeastward into most of New England and southeastern Quebec. Another area extends from Alabama and Georgia into the Virginias. The reminder of the domain has biases between ±25%.

5.2. Categorical evaluation

Table 2 provides both the monthly and overall categorical statistics associated with the peak 8-h O_3 forecasts, along with the actual exceedance and non-exceedance numbers (a, b, c, d) used in their calculation. Because of the non-conducive conditions discussed earlier, very few exceedances occurred during the 4-month period. More specifically, only 327 out of 74,492 (or less than 0.5%) of the total possible cases experienced exceedances.

As seen in the table, the accuracy (A) exceeds 98% for each individual months as well as the overall forecast. As discussed in Kang et al. (2005), care must be used in interpretation of this metric, however, as it is greatly influenced by the

Table 2
Summary of categorical statistics for peak 8-h O_3 forecast

Month	A (%)	B	H (%)	FAR (%)	CSI (%)	WSI[a] (%)	a	b	c	d
June	99.5	0.79	25.3	67.8	16.5	50.9	40	19	18274	56
July	98.0	1.9	50.5	74.1	20.7	46.7	272	95	18258	93
August	98.7	5.2	39.6	92.4	6.8	15.0	232	19	19000	29
September	99.5	4.50	6.25	98.61	1.15	10.0	71	1	18018	15
All	98.9	2.29	41.0	82.1	14.2	34.3	615	134	73550	193

[a] The factor (f) used in calculation of WSI was 1.5.

Table 3
Summary of discrete statistics for peak 8-h O_3 forecast for 4 August

Date	Mod mean	Obs mean	*r*	MB (ppb)	NMB (%)	RMSE (ppb)	NME (%)
4 August	49.6	60.9	0.74	11.3	22.8	16.9	27.6

overwhelming number of correctly forecast non-exceedances (73,550). The values of bias (*B*) indicate that the system slightly underpredicted exceedances during June (0.79) but greatly overpredicted them during the remaining months, especially August (5.2) and September (4.5). The hit rate (*H*) and CSI average 41.0% and 14.2%, respectively, for the entire period, with each metric exhibiting considerable variation from month to month. It should be noted, however, that during the month of July (when most of the observed exceedances occurred, 188 out of 327 or nearly 60%) the values of *H* and CSI were considerably better, at 50.5% and 20.7%, respectively. The season-long FAR, which is very high (82.1%) and ranges between 67.8% (June) and 98.61% (September) is consistent with the systematic overprediction by the AQF system.

As discussed earlier, the WSI takes into consideration not only the actual numbers of exceedance (like the CSI), but also paired points that are within a designated factor as well. As seen in Table 2, when the factor is set at 1.5, the overall WSI at 34.3% is much larger than the CSI at 14.2% (an increase of 20.1%). On a monthly basis the increase ranges from 8.2% (August) to 34.4% (June). These marked increases indicate that there are many data points slightly outside the desirable *b* quadrant (a forecast 8-h exceedance that did occur) that the strict CSI metric categorizes as failures. When the proximity of these data points is taken into consideration by the weighting associated with the WSI, a more representative measure of the model's performance is obtained.

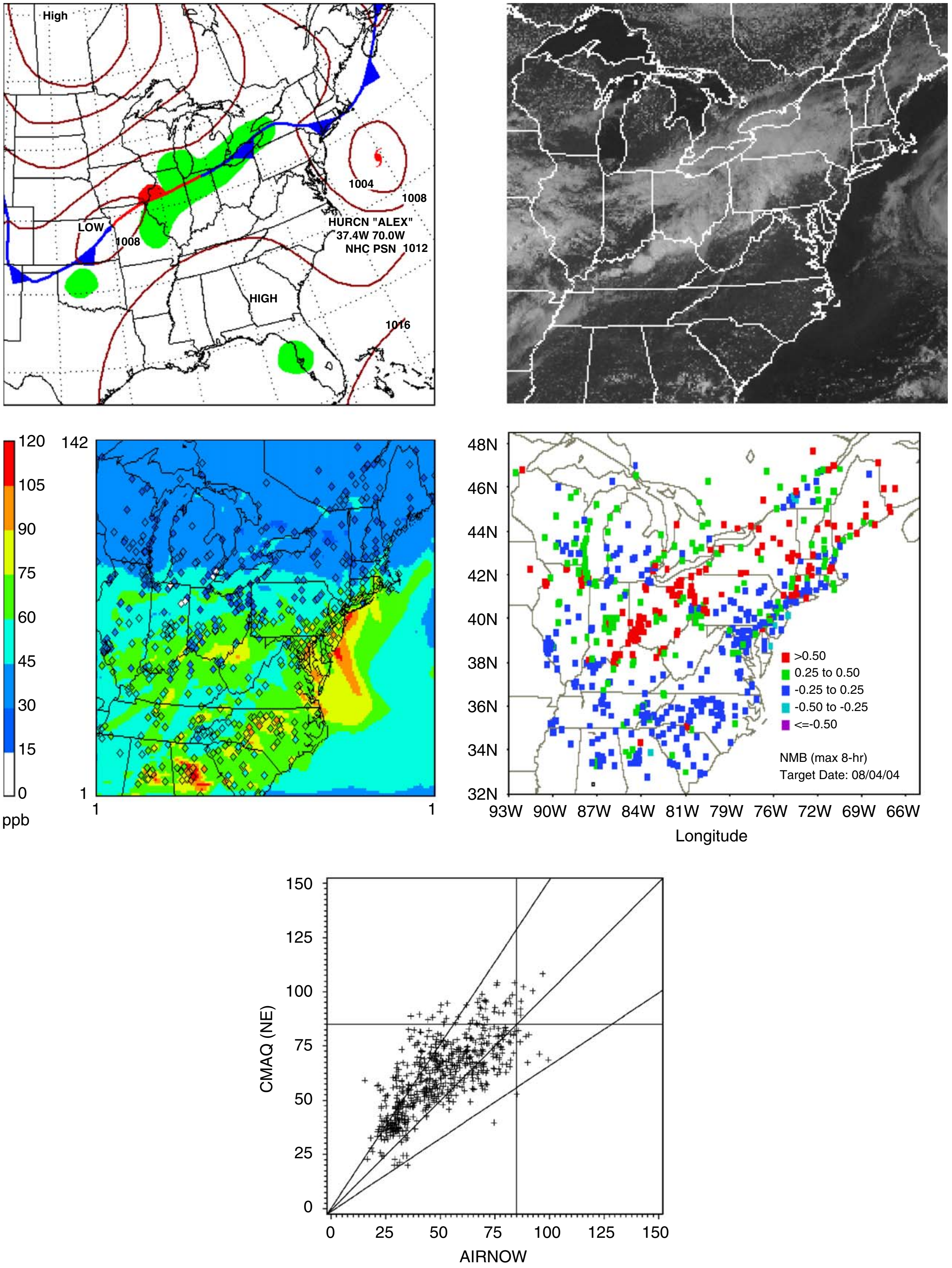
5.3. *Case study—4 August 2004*

The AQF system's forecasts were analyzed, daily throughout its 4-month operation. The "Daily Weather Map" series, obtained from NOAA's National Center for Environmental Prediction (NCEP) (http://www.hpc.ncep.noaa.gov/dailywx-map/), and GOES visible satellite images obtained from NOAA's National Environmental Satellite, Data, and Information Services (NESDIS) (http://www.goes.noaa.gov/) were used to examine the synoptic-scale meteorological conditions impacting the domain. The following discussion examines the performance of the AQF system on a single day (4 August as denoted in Figs. 4 and 5) that typified its performance over the summer (i.e. thorough, daily analysis revealed that the model consistently performed better in areas dominated by clear skies, while performing more poorly in areas dominated by convective cloud cover and precipitation). Discrete metrics associated with this day, which are comparable to the 4-month average, can be found in Table 3. Too few exceedances were recorded to warrant calculation of the categorical metrics.

Examination of the synoptic-scale meteorology for this day revealed an active cold front stretching from the New England coastline southwestward to an area of low pressure over the state of Missouri (Fig. 7). This front was accompanied by convective cloud cover and numerous showers and thunderstorms. Behind this front, an unseasonably strong 1026 millibar (mb) anticyclone was pushing southeastward out of central Canada. This continental Polar (cP) airmass, which was characterized by cool temperatures (<20 °C) and partly cloudy skies (cold advection stratocumulus), was accompanied by low O_3 concentrations. In fact, all observations in this region were less than 45 ppb, with many less than 30 ppb. The pattern of low concentrations was well forecast by the AQF system, although simulated values were biased high in some locations ($25\% \leqslant \mathrm{NMB} \leqslant 50\%$). A warm maritime Tropical (mT) air mass was in place south of the front,

→

Fig. 7. Synoptic-scale meteorological conditions for 12 GMT (top left panel), visible satellite image for 20:15 GMT (top right), AQF system simulation for 20:00 GMT with AQS data overlaid (middle left), daily NMB of the peak 8-h O_3 (middle right); and scatter plot of the peak 8-h O_3 (bottom) for 4 August 2004.


High
LOW
1008
1004
1008
HURCN "ALEX"
37.4W 70.0W
NHC PSN 1012
HIGH
1016
120
105
90
75
60
45
30
15
0
ppb
142
1
1
1
48N
46N
44N
42N
40N
38N
36N
34N
32N
93W 90W 87W 84W 81W 79W 76W 72W 69W 66W
Longitude
🟥 >0.50
🟩 0.25 to 0.50
🟦 -0.25 to 0.25
-0.50 to -0.25
<=-0.50
NMB (max 8-hr)
Target Date: 08/04/04
150
125
100
75
50
25
0
CMAQ (NE)
0 25 50 75 100 125 150
AIRNOW

anchored by a 1016 mb anticyclone. This stagnant mT airmass, which was characterized by mostly clear skies (some fair weather cumulus) and very warm temperatures (>30 °C), resulted in higher concentrations ($45\,\text{ppb} \leqslant O_3 \leqslant 75\,\text{ppb}$). In a few locations, most notably the metropolitan regions of Alabama, Georgia, North Carolina, Maryland and New Jersey had concentrations exceeding 90 ppb. The model replicated this pattern very well resulting in NMB within ±25% in most locations.

The performance along the cold front was, however, considerably worse, as the model greatly overpredicted concentrations resulting in NMBs exceeding 50%. Such overprediction in areas of cloud cover was common throughout the forecast period. Subsequent diagnostic analysis revealed two factors contributing to this overprediction. The first factor involved the use of O_3 profiles derived from the GFS Model in designating the upper boundary conditions. While providing more realistic "near tropopause" O_3 concentrations than those that were used previously (based simply on climatology), these concentrations were found to be too large. This factor was exacerbated by CMAQ's convective cloud scheme. This scheme, which was originally derived from the Regional Acid Deposition Model (RADM) (Chang et al., 1990), was found to be transporting excessive amounts of this "near tropopause" O_3 to the surface, via downdrafts associated with convective clouds. The second factor, which also involved CMAQ's simulation of clouds, revealed that too little attenuation of actinic flux was taking place resulting in too much photolysis and subsequently too much O_3 formation. In combination, these factors resulted in the AQF system's overprediction of O_3 in and around areas of cloud cover.

6. Summary

The purpose of this research has been to provide the first of an expected annual series of operational evaluations of the Eta–CMAQ AQF system using O_3 observations obtained from EPA's AIRNow program and a suite of statistical metrics for both discrete and categorical forecasts. Results from this evaluation revealed that the modeling system performed reasonably well, in this, its first major attempt at forecasting O_3 concentrations over the northeastern United States. The quality of the forecasts was comparable, if not better than similar model forecasts made during the summer of 2002 (Kang, et al., 2005). Examination of the overall, 4-month performance, from a discrete perspective, revealed a systematic tendency to overpredict concentrations resulting in a NMB of 22.8%, a NME = 28.1%, and a correlation of 0.59. The overprediction was also evident from a categorical perspective, as most of these metrics indicated an excessive number of exceedances during the period (e.g. $B = 2.29$, $\text{FAR} = 82.1$). Time series of the metrics associated with the discrete forecasts revealed a systematic pattern of varied modeled performance that could be traced back to the "synoptic-scale" meteorology impacting the domain. During days when high pressure, relative clear skies, and little precipitation occurred within the domain (all conditions conducive to O_3 formation), the model performed well. Conversely, on those days characterized by extensive cloud cover and precipitation (conditions not conducive to O_3 formation) associated with either fronts or areas of low pressure, the model performed poorly.

Subsequent diagnostic analysis revealed two main factors contributing to this overprediction. The first involved the excessive downward transport of O_3 rich air via CMAQ's convective cloud scheme in conjunction with O_3 profiles derived from the GFS Model. The second factor involved too little attenuation of actinic flux by CMAQ's simulated cloud cover, resulting in too much photolysis and subsequently too much O_3. In combination, these factors resulted in the AQF system's systematic overprediction of O_3 in and around areas of cloud cover. Changes to CMAQ's cloud schemes are currently underway that are expected to significantly improve the AQF system's performance in anticipation of its second year of being operational.

Disclaimer

The research presented here was performed under the Memorandum of Understanding between the US Environmental Protection Agency (EPA) and the US Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) and under agreement number DW13921548. This work constitutes a contribution to the NOAA Air Quality Program. Although it has been reviewed by EPA and NOAA and approved for publication, it does not necessarily reflect their policies or views.

References

Black, T., 1994. The new NMC meso-scale Eta Model: description and forecast examples. Weather and Forecasting 9, 265–278.

Byun, D.W., Ching, J.K.S., (Eds.), 1999. Science algorithms of the EPA Models-3 Community Multi-scale Air Quality (CMAQ) modeling system, EPA/600/R-99/030, Office of Research and Development, US Environmental Protection Agency.

Chang, J., et al., 1990. The regional acid deposition model and engineering model. In: National Acid Precipitation Assessment Program, Acidic Deposition, State of Science and Technology, vol. 1, NAPAP SOS/T Rep. 4, National Acid Precipitation Assessment Program, Washington, DC.

Houyoux, M.R., Vukovich, J.M., Coats Jr., C.J., Wheeler, N.M., Kasibhatla, P.S., 2000. Emission inventory development and processing for the seasonal model for regional air quality (SMRAQ) project. Journal of Geophysical Research 105, 9079–9090.

Jolliffe, I.R., Stephenson, D.B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley, West Sussex, England, 240pp.

Kang, D., Eder, B.K., Schere, K.L., 2003. The evaluation of regional-scale air quality models as part of NOAA's air quality forecasting pilot program. In: Proceedings of the 26th NATO/CCMS International Technical Meeting on Air Pollution Modeling and its Application, 26–30 May 2003, Istanbul, Turkey.

Kang, D., Eder, B.K., Stein, A.F., Grell, G.A., Peckham, S.E., McHenry, J., 2005. The New England air quality forecasting pilot program: development of an evaluation protocol and performance benchmark. Journal of the Air & Waste Management Association 55, 1782–1796.

Otte, T.L., et al., 2005. Linking the Eta model with the Community Multi-scale Air Quality (CMAQ) modeling system to build a national air quality forecasting system. Weather and Forecasting 20, 367–384.

Rogers, E., Black, T., Deaven, D., DiMego, G., Zhao, Q., Baldwin, M., Junker, N., Lin, Y., 1996. Changes to the operational "early" Eta Analysis/Forecast System at the National Centers for Environmental Prediction. Weather and Forecasting 11, 391–413.

US Environmental Protection Agency, 2003. Guidelines for developing an air quality (ozone and PM2.5) forecasting program. Office of Air Quality, Planning and Standards, Research Triangle Park, NC, EPA-456/R-03-002, June.