MULTI-MODEL AIR QUALITY FORECASTING OVER NEW YORK STATE FOR SUMMER 2008

Christian Hogrefe[1,2,*], Prakash Doraiswamy[2], Winston Hao[1], Brian Colle[3], Mark Beauharnois[2], Ken Demerjian[2], Jia-Yeong Ku[1], and Gopal Sistla[1]

[1]New York State Department of Environmental Conservation, Albany, NY, [2]Atmospheric Sciences Research Center, State University of New York at Albany, Albany, NY, [3]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY

1. INTRODUCTION

In recent years, the application of comprehensive grid-based one-atmosphere modeling systems has become an integral part of air quality forecasting for both federal and state agencies. For example, since 2005 the New York State Department of Environmental Conservation (NYSDEC) has been utilizing the Community Multiscale Air Quality (CMAQ) model (Byun and Schere, 2006) driven by operational National Center for Environmental Prediction (NCEP) weather forecasts to provide guidance to state air quality forecasters. In an attempt to better quantify uncertainties associated with these ozone and $PM_{2.5}$ forecasts, NYSDEC in collaboration with the University at Albany and Stony Brook University (SUNY-SB) is currently embarking on a research project to apply multiple modeling systems for this task. The concept of ensemble forecasting, a technique that is widely used by the meteorological community to provide for a more robust and accurate weather forecast, is only beginning to be explored in air quality forecasting applications. Building upon various ongoing meteorological and air quality modeling efforts at federal, state, and academic institutions, this study is designed to implement this concept to New York State. As a first step towards this goal, daily air quality simulations for June 4 – August 31, 2008 have been performed with CMAQ driven by four different meteorological forecasts obtained from NCEP and the Stony Brook University Short-Range Ensemble Forecast (SREF) system. Additionally, retrospective simulations were performed in which CMAQ was driven by twelve archived members of the SUNY-SB SREF for June 4 – July 21, 2008. In this study, we present a comparison of the simulations from both the four and twelve member systems against measurements for O_3 and $PM_{2.5}$ over New York State to explore the potential benefits of utilizing a multi-model system to provide air quality forecast guidance.

**Corresponding author:* Christian Hogrefe, BAQAR, New York State Department of Environmental Conservation, 625 Broadway, Albany, NY 12233-3259, Phone (518) 402 8402, Fax (518) 402 9035, Email chogrefe@dec.state.ny.us

2. DATABASE AND METHODS OF ANALYSIS

2.1 Model Setup and Observations

In this study, we analyze CMAQ air quality forecasts for New York State for both a routine four member forecast system and an experimental retrospective twelve member forecast system. In all instances, New York State is covered by horizontal grids with a spacing of 12 km. The four member CMAQ system consists of two CMAQ simulations driven by the NCEP WRF-NMM 12:00 UTC and 00:00 UTC weather forecasts, a third driven by a MM5 member and a fourth driven by a WRF-ARW member of the SUNY-SB 00:00 UTC SREF system (Jones et al., 2007; http://chaos.msrc.sunysb.edu/NEUS/nwp_graphics.html). The WRF-NMM/CMAQ simulations utilize the setup described by Otte et al. (2005), Kang et al. (2005), Yu et al. (2008) and

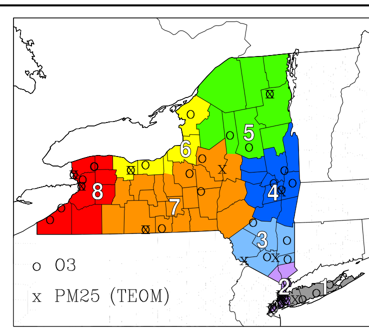

Figure 1. Map showing the eight New York State air quality forecast regions and the locations of the ozone and continuous PM2.5 monitors.

Hogrefe et al. (2007). For the CMAQ simulations driven by the two selected SUNY-SB SREF MM5 and WRF-ARW members, the meteorological fields were processed with the Meteorology-Chemistry Interface Processor (MCIP) and emission inventories were processed with the Sparse Matrix Operator Kernel Emissions Modeling System (SMOKE). It should be noted that emission inputs, domain size, and vertical grid structures vary across the four members of this system. For the retrospective twelve member simulations, CMAQ forecasts were driven by seven daily MM5 and five daily WRF-ARW members from the 00:00 UTC SUNY-SB SREF system. Again, these meteorological fields were processed through MCIP and emissions were processed through SMOKE. For these simulations, horizontal grid structure and emission inputs are constant across all members.

Observations of hourly ozone and total $PM_{2.5}$ concentrations for monitors in New York State were downloaded from the EPA AIRNOW system. Daily maximum 8-hr ozone concentrations and 24-hr average $PM_{2.5}$ concentrations were then determined from the hourly data and used in the subsequent analyses.

2.2 Evaluation Metrics

Routine human-based daily air quality forecasts in New York State are issued by the NYSDEC for the eight forecast regions shown in Figure 1. Consequently, the evaluation of the model-based forecasts follows the same approach in which the ozone or $PM_{2.5}$ values for a given region and a given day are determined by taking the maximum observed and simulated values across all ozone or $PM_{2.5}$ monitors located in that region. This analysis only uses model predictions from grid cells containing the monitors shown in

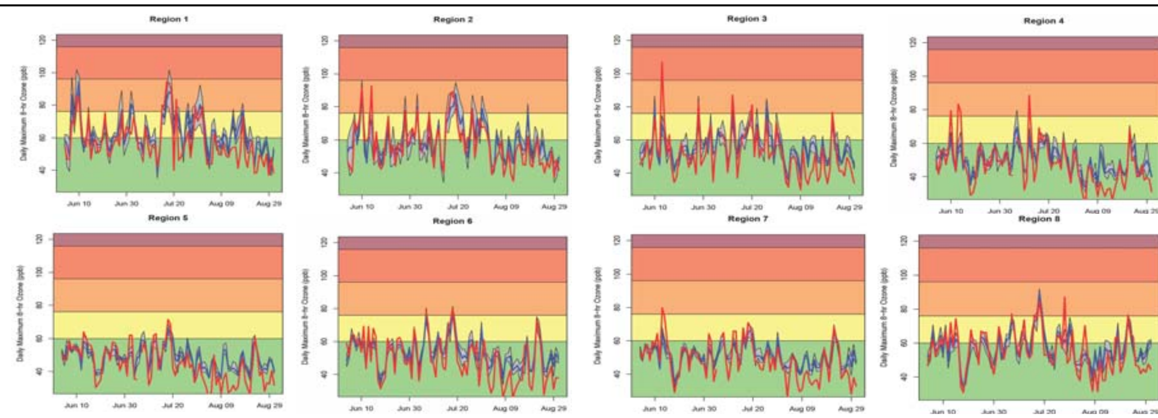

Figure 2a. Time series of observed and simulated daily maximum 8-hr ozone for June 4 – August 31, 2008 for forecast regions 1-4 (top) and regions 5 – 8 (bottom). Further details are provided in the text.

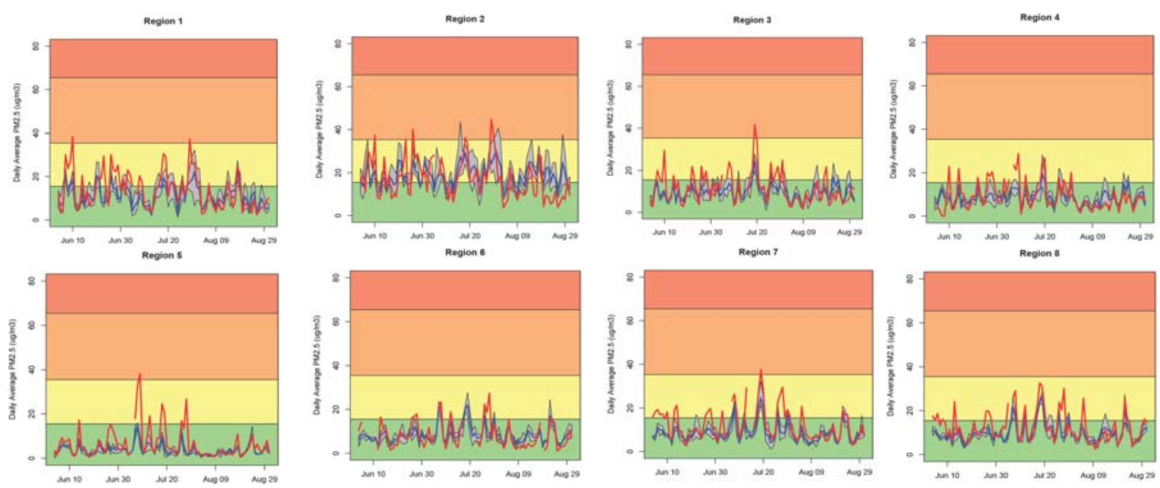

Figure 2b. As Figure 2a but for daily average $PM_{2.5}$

Figure 1. These daily observation/model pairs are then used to compute discrete, categorical, and probabilistic model performance measures for each pollutant and region. The discrete forecast measures included in this analysis are the bias, root mean square error (RMSE), and correlation coefficient as defined by Willmott (1982) and the categorical metrics are the False Alarm Ratio (FAR), Probability of Detection (POD), and Critical Success Index (CSI) as described by Kang et al. (2005). For the categorical metrics, we selected a threshold that corresponds to the transition from the “moderate” to the “unhealthy for sensitivty groups” range of the Air Quality Index (AQI). For ozone, this threshold corresponds to 75 ppb (U.S. EPA, 1999 and 2008), while for $PM_{2.5}$, it corresponds to 35.4 µg/m^3 for forecasts issued in New York State (NYSDEC, 2007). The probabilistic evaluation approach is described in Section 3.2.

3. RESULTS AND DISCUSSION

3.1 Four Member Forecast System

Time series of observed and predicted daily maximum 8-hr ozone and daily average $PM_{2.5}$ concentrations for June 4 – August 31, 2008 are shown in Figures 2a-b for the eight forecast regions in New York. The red lines represent observations, the blue lines represent the average over the four model predictions for each day, and the area shaded in gray represents the range between the minimum and maximum of the four predicted values for each day. While the concentrations in the time series are shown in ppb or µg/m^3, the background in all panels is shaded to correspond to the ranges of the Air Quality Index (AQI). For ozone, the four-model average prediction tracks the observations well for all regions. In terms of regional differences, it can be seen that ozone exceeded a threshold of 75 ppb (AQI 100) on a number of days for regions 1 and 2, i.e. Long Island and New York, while generally lower ozone concentrations with fewer or no exceedances were observed and simulated for the other regions. $PM_{2.5}$ concentrations were generally also higher in region 1 and especially in region 2 compared to other regions, but there were only very few observed or predicted exceedances of the AQI 100 threshold during this time period. For most regions, fluctuations of predicted $PM_{2.5}$ concentrations track well with observations but tend to be too low. Quantitative measures of model performance for ozone and $PM_{2.5}$ for the four individual models as well as the four-model average are shown in Tables 1a-b for regions 1 and 2. For ozone, both the discrete metrics (bias, root mean square error, and correlation coefficient) and categorical metrics for an ozone exceedance threshold of 75 ppb confirm that the modeling systems provided good ozone forecasts for these regions during the summer of 2008. The evaluation metrics for $PM_{2.5}$ in Table 1b show that $PM_{2.5}$ tends to be underestimated by all modeling systems in region 1 but overestimated by two of the four modeling systems and the average model forecast in region 2. Correlation coefficients for $PM_{2.5}$ are lower than for ozone. No categorical metrics are shown for $PM_{2.5}$ because only a few

Table 1a. Discrete and categorical performance statistics for daily maximum 8-hr ozone. M1 denotes the 12:00 UTC NCEP WRF-NMM/CMAQ simulation, M2 the 00:00 UTC NCEP WRF-NMM/CMAQ simulation, M3 the 00:00 UTC SUNY-SB MM5/CMAQ simulation, M4 the 00:00 UTC SUNY-SB WRF-ARW/CMAQ simulation, and Av the forecast calculated by averaging the forecasts from M1 – M4.

	Region 1					Region 2				
	M1	M2	M3	M4	Av	M1	M2	M3	M4	Av
Bias (ppb)	5.4	7.3	3.2	4.1	4.1	1.7	2.5	1.5	3.1	1.6
RMSE (ppb)	11.3	11.5	9.1	9.4	8.9	9.6	8.8	7.9	8.8	7.3
Correlation	0.72	0.77	0.77	0.78	0.81	0.73	0.79	0.83	0.79	0.85
POD (%)	62	77	69	62	46	55	82	55	55	64
FAR (%)	47	44	31	43	40	40	36	33	25	30
CSI (%)	40	48	53	42	35	40	56	43	46	50

Table 1b. As in Table 1a but for daily average $PM_{2.5}$

	Region1					Region2				
	M1	M2	M3	M4	Av	M1	M2	M3	M4	Av
Bias (µg/m^3)	-4.9	-3.4	-0.3	-1.1	-2.4	-1.7	-0.3	4.6	3.1	1.3
RMSE (µg/m^3)	8.1	6.9	6.9	6.7	6.4	8.1	7.5	10.5	9.5	7.9
Correlation	0.7	0.73	0.62	0.64	0.75	0.54	0.61	0.38	0.42	0.56

exceedances of the AQI 100 threshold were observed or simulated during this time period. The ozone performance metrics generally show equal or better model performance than the benchmark values reported by Kang et al. (2005) for a 25-day period during August 2002 and the $PM_{2.5}$ metrics show similar values than those reported by Yu et al. (2008) for a forecasting study for July and August 2004. In addition, the forecasts generated by averaging the forecasts from the four individual modeling systems tends to perform better than the individual model forecasts, suggesting than a multi-model air quality forecasting approach may yield improved forecast guidance.

3.2 Twelve Member Forecast System

To further explore the potential benefits of utilizing an ensemble system to provide air quality forecast guidance, retrospective simulations were performed in which CMAQ was driven by twelve archived members of the SUNY-SB SREF system (Jones et al., 2007) for June 4 – July 21, 2008. To illustrate the variability in meteorological and air quality predictions introduced by using twelve different meteorological forecasts to CMAQ, Figure 3 illustrates spatial fields of the average coefficient of variation for daily maximum temperature, PBL height and 8-hr ozone and daily average wind speed and $PM_{2.5}$. The coefficient of variation for each variable was calculated for each day at each grid cell by dividing the standard deviation of the twelve predicted values by the mean of the twelve predicted values.These daily maps of the coefficient of variation were then averaged over the June 4 – July 21 modeling time period. These figures illustrate that the choice of meteorological ensemble members results in a typical variability of 5%, 15-20%, and 30-40% over land for daily maximum temperature, daily average wind speed, and daily maximum PBL height, respectively. For ozone, the typical variability is on the order of 5-10%, while it is on the order of 20%-25% for $PM_{2.5}$. It is also interesting to note that the coefficient of variation for ozone shows local maxima in urban regions (e.g. New York City and Toronto), suggesting that different meteorological forecast fields lead to a wider divergence of ozone forecasts in areas of high emission densities.

The discrete and categorical evaluation of the ozone and $PM_{2.5}$ predictions for these retrospective twelve member simulations yielded similar results as those described in Section 3.1 for the four member forecast system. In particular, the ozone and $PM_{2.5}$ forecast generated by averaging the forecasts from the twelve individual modeling systems tends to perform better than individual model forecasts. In addition to measuring model performance by these discrete and categorical metrics, however, one can also evaluate the probabilistic aspects of the twelve member ensemble forecasts. As described in Talagrand et al. (1997), DellaMonache et al. (2006) and Jones et al. (2007), this can be accomplished by computing the Brier Skill Score and constructing Talagrand and reliability diagrams among other measures. For illustration, Figures 4a-b present Talagrand diagrams for daily maximum 8-hr ozone and daily average $PM_{2.5}$

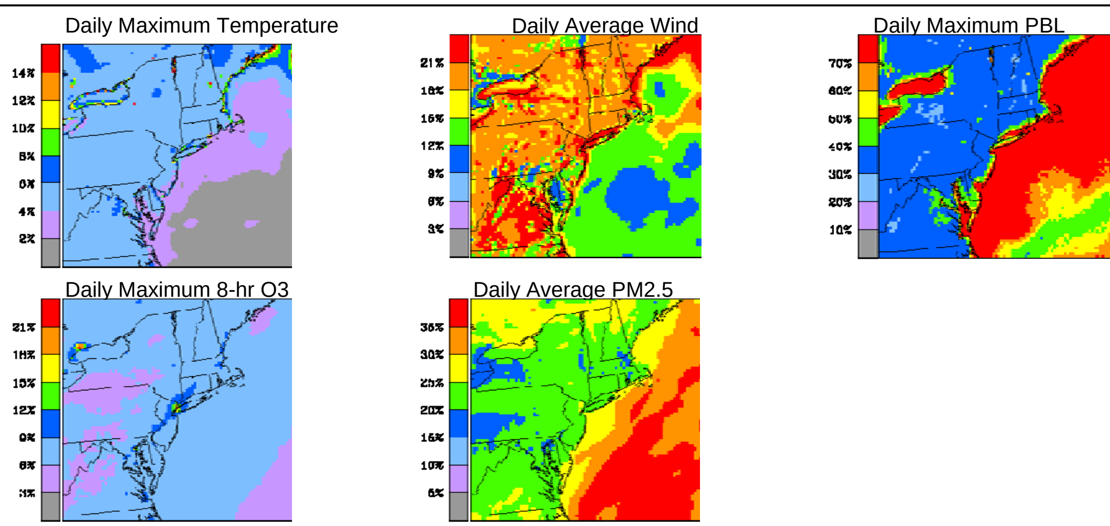

Figure 3. Time-averaged coefficient of variation across the twelve ensemble members for meteorological and air quality parameters as described in the text.

forecasts for three selected regions each. The first step in constructing these Talagrand diagrams is to rank-order the twelve forecasts for a given day. Next, depending on whether the observation for that day is lower than the lowest forecast, falls between the lowest and second-lowest forecast, …, or is higher than the highest forecast, that day is assigned to one of 13 bins. Last, the analysis is repeated for all days and a histogram is created based on the number of days assigned to each of the 13 bins. In an ideal ensemble forecast system, the observations are equally likely to lie between any two ordered adjacent forecast members, including the cases when the observations are outside the ensemble range on either side of the distribution. Therefore, the Talagrand diagram for an ideal ensemble system, i.e. a system that captures the true spread of the underlying observations, would show histogram bars of equal height. A U-shaped (inverted U-shaped) diagram indicates that the spread of the ensemble is too small (large) because the observed event too often falls outside (inside) the range of values sampled by the ensemble. Systematic over (under) prediction biases in the ensemble system lead to a L-shaped (inverted L-shaped) diagram.

To illustrate such cases, Figures 4a-b present Talagrand diagrams for daily maximum 8-hr ozone and daily average $PM_{2.5}$ forecasts for three selected regions each. Since 48 days and 12 ensemble members (13 bins) were used, the expected number of days in each bin of an ideal Talagrand diagram would be between 3 and 4. The three panels in Figure 4a indicate that the ensemble system is close to capturing the true underlying spread of daily maximum 8-hr ozone for regions 1 and 2 and underestimates the spread for region 3. For $PM_{2.5}$, the spread is underestimated for region1, the system is biased high for region 2, and biased low for region 7. Separate computation of reliability diagrams (Jones et al., 2007) for ozone forecasts for regions 1 and 2 for a threshold of 75 ppb indicates good performance of the twelve member system with respect to predicting the probability of exceeding this threshold. In summary, these results show that an air quality ensemble forecast system driven by a meteorological ensemble forecast system could potentially be valuable in providing probabilistic ozone forecasts, especially for regions 1 and 2. However, the system tends to be underdispersed for ozone for other regions and for $PM_{2.5}$ for all regions, indicating that uncertainties from other factors such as emissions or chemistry need to be included for a better treatment of variability. Furthermore, the ensemble simulations need to be carried out for longer time periods and other seasons than the 48 day summertime period described in this study to confirm these initial findings.

4. SUMMARY

Ozone and $PM_{2.5}$ air quality forecasts for eight New York State forecast regions generated by multiple air quality forecast modeling systems were compared against observations for the summer of 2008. Discrete and categorical model

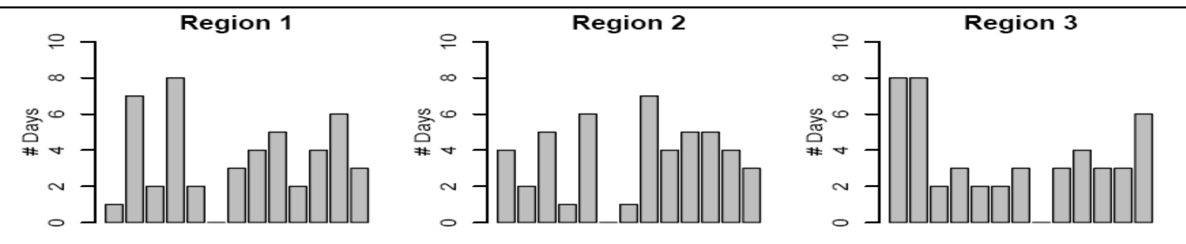

Figure 4a. Talagrand diagrams for daily maximum 8-hr O_3 predictions for forecast regions 1, 2, and 3.

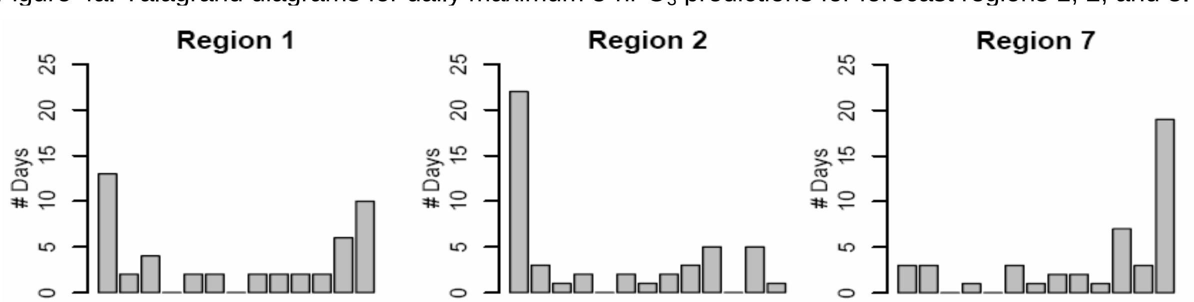

Figure 4b. Talagrand diagrams for daily average $PM_{2.5}$ predictions for forecast regions 1, 2, and 7.

evaluation shows that the four model forecast system operated since June 4, 2008 provided good ozone forecast guidance, especially for Long Island and NYC metro area. On the other hand, forecasts for summertime $PM_{2.5}$ are characterized by a negative bias for all areas except the NYC metro area. The four member mean forecasts often but not always exhibit better performance statistics than individual forecasts. Employing weighting or bias correction approaches prior to averaging may further improve forecast performance. For a 48 day retrospective case study, CMAQ simulations were performed using twelve weather forecasts from the SUNY-SB SREF system. Variations in meteorology introduced by the twelve SREF members cause a typical daily maximum 8-hr ozone variability of 5 – 10% (with higher values in urban areas) and a typical daily average $PM_{2.5}$ variability of 20 – 25% over land areas. An evaluation of the probabilistic aspects of the ensemble forecast reveals that the twelve member system provides a realistic spread of ozone concentrations for regions 1 and 2 but tends to be underdispersed for other regions and for $PM_{2.5}$, indicating that uncertainties from other factors such as emissions or chemistry need to be included for a better treatment of variability. Overall, these results suggest that pursuing an ensemble air quality forecasting approach may yield improved forecast guidance as measured by discrete, categorical, and probabilistic metrics.

5. ACKNOWLEDGMENTS

The work presented in this paper was performed by the New York State Department of Environmental Conservation with partial support from the U.S. EPA under cooperative agreement CR83228001 and the New York State Energy Research and Development Authority (NYSERDA) under agreement #10599. The views expressed in this paper do not necessarily reflect the views or policies of the New York State Department of Environmental Conservation or those of the sponsoring agencies.

6. REFERENCES

Byun, D.W. and K.L. Schere, 2006: Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. Appl. Mech. Rev., 59, 51-77.

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

Hogrefe, C. W. Hao, K. Civerolo, J.-Y. Ku, R.S. Gaza, L. Sedefian, G. Sistla, K. Schere, A. Gilliland, and R. Mathur, 2007: Daily photochemical simulations of ozone and fine particulates over New York State: Findings and challenge, *Journal of Applied Meteorology*, 46, 961–979

Jones, M.S., B.A. Colle, and J.S. Tongue 2007: Evaluation of a Mesoscale Short-Range Ensemble Forecast System over the Northeast United States. Weather and Forecasting, 22, 36–55.

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

NYSDEC, 2007: DEC Using More Stringent Measurement for Air Quality Forecasts. Environment DEC, July 2007, available online at http://www.dec.ny.gov/environmentdec/35375.html

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

U.S. EPA, 1999: Guideline for reporting of daily air quality – air quality index (AQI), EPA-454/R-99-010, United States Environmental Protection Agency, Research Triangle Park, NC 27711, 29 pp.

U.S. EPA, 2008: Revisions to EPA's ozone air quality index fact sheet. Available online at http://www.epa.gov/air/ozonepollution/pdfs/2008_03_aqi_changes.pdf

Talagrand, O., R. Vautard, B. Strauss, 1997. Evaluation of probabilistic prediction systems, in: Proceedings of Workshop on Predictability, ECMWF, Reading, UK, pp. 1-25.

Willmott, C.J., 1982. Some Comments on the Evaluation of Model Performance. *Bulletin of the American Meteorological Society*, **63**, 1309-1313.

Yu S., R. Mathur, K. Schere, D. Kang, J. Pleim, J. Young, D. Tong, G. Pouliot, S. A. McKeen, S. T. Rao (2008), Evaluation of real-time PM2.5 forecasts and process analysis for PM2.5 formation over the eastern United States using the Eta-CMAQ forecast model during the 2004 ICARTT study, J. Geophys. Res., 113, D06204, doi:10.1029/2007JD009226.