JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 112, D10S20, doi:10.1029/2006JD007608, 2007

Evaluation of several $PM_{2.5}$ forecast models using data collected during the ICARTT/NEAQS 2004 field study

S. McKeen,[1,2] S. H. Chung,[1,2] J. Wilczak,[3] G. Grell,[2,4] I. Djalalova,[2,3] S. Peckham,[2,4] W. Gong,[5] V. Bouchet,[6] R. Moffet,[6] Y. Tang,[7] G. R. Carmichael,[7] R. Mathur,[8,9] and S. Yu[9,10]

Received 1 June 2006; revised 11 October 2006; accepted 5 December 2006; published 21 March 2007.

[1] Real-time forecasts of $PM_{2.5}$ aerosol mass from seven air quality forecast models (AQFMs) are statistically evaluated against observations collected in the northeastern United States and southeastern Canada from two surface networks and aircraft data during the summer of 2004 International Consortium for Atmospheric Research on Transport and Transformation (ICARTT)/New England Air Quality Study (NEAQS) field campaign. The AIRNOW surface network is used to evaluate $PM_{2.5}$ aerosol mass, the U.S. EPA STN network is used for $PM_{2.5}$ aerosol composition comparisons, and aerosol size distribution and composition measured from the NOAA P-3 aircraft are also compared. Statistics based on midday 8-hour averages, as well as 24-hour averages are evaluated against the AIRNOW surface network. When the 8-hour average $PM_{2.5}$ statistics are compared against equivalent ozone statistics for each model, the analysis shows that $PM_{2.5}$ forecasts possess nearly equivalent correlation, less bias, and better skill relative to the corresponding ozone forecasts. An analysis of the diurnal variability shows that most models do not reproduce the observed diurnal cycle at urban and suburban monitor locations, particularly during the nighttime to early morning transition. While observations show median rural $PM_{2.5}$ levels similar to urban and suburban values, the models display noticeably smaller rural/urban $PM_{2.5}$ ratios. The ensemble $PM_{2.5}$ forecast, created by combining six separate forecasts with equal weighting, is also evaluated and shown to yield the best possible forecast in terms of the statistical measures considered. The comparisons of $PM_{2.5}$ composition with NOAA P-3 aircraft data reveals two important features: (1) The organic component of $PM_{2.5}$ is significantly underpredicted by all the AQFMs and (2) those models that include aqueous phase oxidation of SO_2 to sulfate in clouds overpredict sulfate levels while those AQFMs that do not include this transformation mechanism underpredict sulfate. Errors in $PM_{2.5}$ ammonium levels tend to correlate directly with errors in sulfate. Comparisons of $PM_{2.5}$ composition with the U.S. EPA STN network for three of the AQFMs show that sulfate biases are consistently lower at the surface than aloft. Recommendations for further research and analysis to help improve $PM_{2.5}$ forecasts are also provided.

Citation: McKeen, S., et al. (2007), Evaluation of several $PM_{2.5}$ forecast models using data collected during the ICARTT/NEAQS 2004 field study, *J. Geophys. Res.*, *112*, D10S20, doi:10.1029/2006JD007608.

1. Introduction

[2] Eulerian model based forecasts of ozone, a common air pollutant, have been publicly available in the United States and Canada for several years, while public forecasts of other criteria pollutants, such as $PM_{2.5}$ aerosol (particulate matter with diameter less than 2.5 μm), are more recent and for the most part in a developmental stage. The need and justification for forecasting ozone also apply to forecasting $PM_{2.5}$ levels; there is a sufficient amount of clinical

[1] Chemical Sciences Division, Environmental Science Research Laboratory, NOAA, Boulder, Colorado, USA.

[2] Also at Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA.

[3] Physical Sciences Division, Environmental Science Research Laboratory, NOAA, Boulder, Colorado, USA.

[4] Global Systems Division, Environmental Science Research Laboratory, NOAA, Boulder, Colorado, USA.

[5] Meteorological Service of Canada, Downsview, Ontario, Canada.

[6] Meteorological Service of Canada, Dorval, Quebec, Canada.

[7] Center for Global and Regional Environmental Research, University of Iowa, Iowa City, Iowa, USA.

[8] Air Resources Laboratory, NOAA, Silver Spring, Maryland, USA.

[9] Also at National Exposure Research Laboratory, Environmental Protection Agency, Research Triangle Park, North Carolina, USA.

[10] Science and Technology Corporation, Hampton, Virginia, USA.

0148-0227/07/2006JD007608$09.00

and epidemiological evidence to relate unhealthy levels of $PM_{2.5}$ to adverse health effects and hospital admittances. $PM_{2.5}$ also contributes adversely to visibility, and has both direct and indirect effects on radiative forcing. Several research centers have been forecasting $PM_{2.5}$ levels on a real-time basis over the past few years, and four research centers in particular (NOAA/ESRL/GSD, EPA/NERL/AMD, Meteorological Services of Canada, and University of Iowa) provided $PM_{2.5}$ forecasts during the summer of 2004 over eastern North America during the International Consortium for Atmospheric Research on Transport and Transformation (ICARTT)/New England Air Quality Study (NEAQS) 2004 field experiment. This field program was unique in that detailed and extensive measurements of aerosol and aerosol composition were made over a large region and from several different mobile platforms, allowing for the evaluation of the various $PM_{2.5}$ forecast models, as well as the assorted processes that are treated with numerical approximations within each model. Reliable, high time resolution aircraft observations of $PM_{2.5}$ speciation are a recent development in measurement capability, and provide a complementary and contrasting view of model performance relative to comparisons restricted to surface observations.

[3] The ICARTT/NEAQS-2004 field program also included a real-time model evaluation project, in which real-time forecasts of nine air quality models were collected at a central facility (NOAA/ESRL/CSD), and the comparisons with the U.S. EPA AIRNow ozone surface network, NOAA *Ronald H. Brown* ship data, and University of New Hampshire AIRMAP air quality network were made available to study participants in near real-time. Five of these models also forecast $PM_{2.5}$, and two additional retrospective $PM_{2.5}$ forecasts were made for the summer of 2004 that were also submitted to the NOAA/ESRL/CSD lab. This collaboration between various research centers allows for a more holistic approach to comparisons with the observations, and has been useful in quickly identifying deficiencies, irregularities and inconsistencies common to all models, as well as relative model performance.

[4] This work documents this multimodel comparison approach to the $PM_{2.5}$ forecasts for the summer of 2004. Section 2 gives a brief review of previous Eulerian aerosol modeling and the models used in this study. Model forecast results are compared against three sets of observations collected during that summer in the following three sections. Section 3 covers comparisons with the U.S. EPA AIRNow $PM_{2.5}$ monitoring network, and section 4 covers comparisons with the NOAA P-3 aircraft during the intensive phase of the field experiment. Comparisons of forecasts from three of the models with data from the U.S. EPA Speciation Trends Network are covered in section 5. Important results of the comparison are summarized in the conclusion along with recommendations for further research to improve forecasts.

2. Model Descriptions

2.1. Review of Existing Eulerian Models That Predict $PM_{2.5}$

[5] Although most current aerosol models tend to treat the same major physical-chemical processes, there are significant differences among models in their characterization of PM chemical composition and size distribution. The major differences arise from treatments of gas phase mechanism, aqueous chemistry, inorganic aerosol thermodynamics, secondary organic aerosol formation, and cloud processing (including wet deposition). Recent and comprehensive reviews of PM models are provided by *Seigneur* [2001] and *Seigneur and Moran* [2004]. A brief discussion of main differences in numerical methodology is given below.

[6] The size distribution of aerosols in tropospheric air quality models can be represented by the moment approach [*Yu et al.*, 2003], the modal approach [*Binkowski and Roselle*, 2003], or the sectional approach [*Zhang et al.*, 2004]. In the simplest moment approach, only PM mass or speciated PM mass is considered; this is the approach of GOCART [*Chin et al.*, 2000] (except for dust particles), and the CHRONOS Canadian forecast model [*Pudykiewicz et al.*, 1997]. In the modal approach the particle size distribution is represented by the sum of several analytical functions. The analytical functions are typically lognormal, each characterized by total number concentration, median diameter, and geometric standard deviation. The modal representation is used in MADE/SORGAM (implemented in WRF/CHEM [*Grell et al.*, 2005]), one version of MOSAIC [*Fast et al.*, 2006], and RPM (implemented in CMAQ [*Binkowski and Roselle*, 2003]). The sectional approach is used in several air quality models, including CAM (implemented in AURAMS [*Gong et al.*, 2003]), CIT [*Meng et al.*, 1998], GATOR [*Jacobson*, 1997], MADM [*Pilinis et al.*, 2000], MADRID [*Zhang et al.*, 2004], MOSAIC, SMOG [*Lu et al.*, 1997], STEM-2K3 [*Tang et al.*, 2004], UAM-AERO [*Lurmann et al.*, 1997], and UAM-AIM [*Sun and Wexler*, 1998]. In these models the particle size distribution is approximated by a discrete number of size sections in which the properties of all particles are assumed to be uniform. Single-moment algorithms are most commonly implemented in PM models that use sectional representation, with the single moment being aerosol volume (or mass). In order to limit numerical diffusion and to improve prediction of particle number, which is important for determining aerosol indirect effect, two-moment algorithms have also been implemented. Among urban and regional applications, only GATOR, MADRID, and MOSAIC use two-moment sectional algorithms. Two-moment schemes have mostly been applied in global models (TOMAS [*Adams and Seinfeld*, 2002], GLOMAP [*Spracklen et al.*, 2005], and GATOR). As a general rule, the modal approach offers the advantage of being computationally efficient, whereas the sectional representation provides more accuracy at the expense of computational cost. Comparison of the modal and sectional approaches in particle size representation is given by *Zhang et al.* [1999].

[7] PM models also differ in their treatment of gas-particle equilibrium and mass transfer between the gas and the particulate phases. The three major methods are (1) the dynamic approach, (2) the equilibrium approach, and (3) the hybrid approach. In theory, the dynamic approach provides the most accurate representation of gas-particle partition, but it is computationally expensive. The equilibrium approach, on the other hand, is computationally efficient but can be inaccurate under certain ambient con-

ditions. The hybrid approaches are attempts to provide the best compromise between accuracy and computational speed. The dynamic approach is implemented in CIT, GATOR, MADM, MOSAIC, STEM-2K3, and UAM-AIM. In the full dynamic approach, the mass transfer between the gas and the particle phases is simulated explicitly; the gas and particle phase concentrations of each species may or may not be in equilibrium. In the equilibrium approach, the gas and the particulate phases are assumed to be in chemical equilibrium. The equilibrium assumption requires that all particles have the same chemical composition for all species involved in gas-particle equilibrium. The equilibrium approach is implemented in RPM and CMAQ [*Binkowski and Roselle*, 2003], WRF/CHEM [*Ackermann et al.*, 1998], CHRONOS and AURAMS (both use the ISORROPIA mechanism of *Nenes et al.* [1998]). In one variation of the hybrid approach, the gas phase is assumed to be in chemical equilibrium with the whole particulate phase, but the distribution of condensable/volatile species among the particles of different sizes is determined by diffusion-limited assumptions. Unlike the full equilibrium approach, the individual particles are not in equilibrium with the gas phase. This type of hybrid approached is implemented in CIT, UAM-AERO, and AURAMS. A different variation of the hybrid approach, proposed by *Capaldo et al.* [2000], is to assume full equilibrium for particles with diameter less than a threshold value (around 1 μm) and to use the dynamic approach for the larger particles. This version of the hybrid approach is implemented in MADRID. *Zhang et al.* [2000] and *Koo et al.* [2003] examine the accuracy of the three approaches to modeling gas/particle mass transfer. A review of inorganic aerosol equilibrium modules used in PM models is provided by *Zhang et al.* [2000].

[8] A major difference between PM models is that not all models include aqueous phase oxidation of SO_2 by H_2O_2 and O_3 as a source of aerosol sulfate. Global modeling studies indicate that aqueous phase oxidation is a major sink pathway of SO_2 and contributes approximately 80% of the global production rate of aerosol sulfate [e.g., *Koch et al.*, 1999; *Barth et al.*, 2000]. Among the models studied in the work, AURAMS, CMAQ/ETA, and STEM-2K3 include aqueous phase oxidation. The CMAQ/ETA and STEM-2K3 models assume equilibrium partitioning of SO_2, H_2O_2 and O_3 into cloud liquid water using Henry's Law, which is applicable under most atmospheric conditions but can also lead to overprediction of aqueous phase reactions under certain conditions due to mass transport limitation [*Schwartz*, 1988]. The AURAMS model uses a kinetic mass transfer approach modified from the equilibrium approach of *Gong* [2002].

[9] Most PM models require meteorological inputs (e.g., wind speed and direction, turbulence, radiation, clouds, and precipitation) from a 3D host meteorological model or from observations. A limitation of these models is that transport and transformations of chemical and PM components are decoupled from meteorological and radiation calculations. Among urban and regional models, the exceptions are GATOR, SMOG, and WRF/CHEM. These models include detailed online treatments of meteorology, tracer transport, chemistry, and PM processes that treats shorter timescale interactions with meteorology relative to off-line formulations. In addition, these models incorporate online calculation of PM optical properties, which allows for studying the coupling of PM radiative forcing and meteorology. WRF/CHEM has the ability to study the role of PM in redistributing and absorbing solar radiation using the MOSAIC PM module [*Fast et al.*, 2006]. However, for this study MADE/SORGAM is used as the PM module in the WRF/CHEM runs, and aerosol/radiation interactions are not explicitly treated.

2.2. Air Quality Forecast Models (AQFMs) Used in the Evaluation

[10] Seven models are incorporated within the following $PM_{2.5}$ model-measurement comparisons: WRF/CHEM-1 (27 km res.), WRF/CHEM-2 (27 km res.), WRF/CHEM-2 (12 km res.), AURAMS (42 km res.), CHRONOS (21 km res.), STEM-2K3 (12 km res.) and CMAQ/ETA (12 km res.). A fairly detailed description for most of these models, relating in particular to model framework, initial and boundary conditions, emissions, and gas phase oxidation mechanisms, is given by *McKeen et al.* [2005]. Additional information related to the treatment of $PM_{2.5}$ is provided in references associated with each model. For the NOAA/ESRL/GSD WRF/CHEM models the treatment of $PM_{2.5}$ is covered by *Grell et al.* [2005]. The treatment of aerosols in AURAMS is given by *Gong et al.* [2003], and for CHRONOS in the work by *Pudykiewicz et al.* [1997]. A description of the STEM-2K3 implementation of aerosols is given by *Tang et al.* [2004]. The developmental CMAQ/ETA model uses the same aerosol formalism as CMAQ described by *Binkowski and Roselle* [2003] and updates as described by *Yu et al.* [2007]. The aerosol size distribution is modeled as a superposition of three lognormal models corresponding to the ultrafine (diameter (D_p) < 0.1 μm), fine (0.1 μm < D_p < 2.5 μm), and coarse (D_p > 2.5 μm) particle sizes. Model results for $PM_{2.5}$ concentrations are obtained by summing species concentrations over the first two modes. The two WRF/CHEM-2 models are identical in terms of formulation and physics options except for two important features: the horizontal grid resolution (27 km versus 12 km) and the parameterization of planetary boundary layer (PBL) transport.

[11] The forecast domains of the seven models within this study are shown in Figure 1. The region of model overlap is defined by the STEM-2K3 domain boundaries, but analysis here is restricted to a slightly smaller area that excludes the northwest corner of the STEM-2K3 domain and is the same area defined for the O_3 comparison and ensemble study of *McKeen et al.* [2005]. All of the models except STEM-2K3 use fixed boundary conditions of $PM_{2.5}$ mass and composition, and thus ignore PM2.5 sources outside the model domain such as Asian pollution or dust events. The 12-km resolution STEM-2K3 model is nested within a 60-km resolution that covers much of North America. The time-varying lateral and top boundary conditions of the 60-km resolution model are driven by results from the MOZART-2 global chemical transport model [*Horowitz et al.*, 2003], and therefore include global sources to background PM.

[12] None of the models include sources of $PM_{2.5}$ from wildfires. Smoke and enhanced CO from forest fires originating in Alaska and western Canada were observed over the continental United States during the summer of 2004 [*Pfister et al.*, 2005]. Smoke and emissions from these fires were also detected on several occasions during the ICARTT

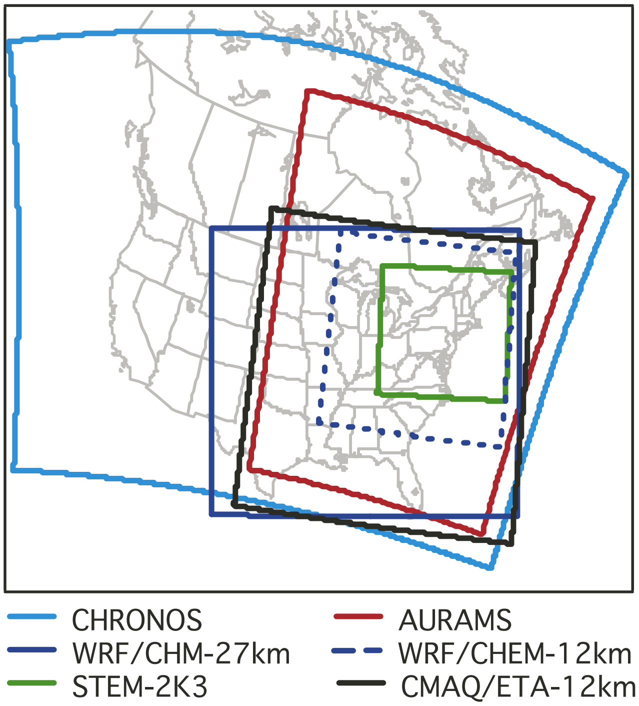

Figure 1. Location of forecast model domain boundaries. The STEM-2K3 model boundaries determine the domain used in the analysis.

field study by the NOAA WP-3 aircraft [*de Gouw et al.*, 2005], mostly in the 2 to 5 km altitude range. As explained in section 4, flight sections that intercepted smoke plumes are eliminated in the model comparisons by using observations of acetonitrile, a chemical marker of wildfire emissions, to window out those air masses modified by wildfires. The impact these wildfire sources have on the $PM_{2.5}$ comparisons with surface monitors (sections 3 and 5) are not as easy to dismiss or assess. *Warneke et al.* [2006] show evidence that the emissions from the Alaskan/Canadian wildfires impacted surface and low altitude aircraft measurements over New England on one particular day, 11 July. CO data from the AIRMAP network in New Hampshire [*DeBell et al.*, 2004] has been used in the past to show the influence of Canadian fires on the northeast United States during the summer of 2002. Analysis of the AIRMAP data for the summer of 2004 (http://soot.sr.unh.edu/airmap/archive/) shows none of the extreme CO events seen in 2002. The one clear case of enhanced fire CO occurs on 11 July (at Mount Washington), the same day that *Warneke et al.* [2006] show the influence of the fires in near-surface air. The dates for the analysis used in this study are from 14 July to 17 August, thus avoiding the largest impact of the wildfires on the surface comparisons. Minor contributions from wildfires on $PM_{2.5}$ levels at the surface sites cannot be entirely discounted, though their impact is minimized by the time interval used in the comparisons.

3. Evaluation of Several $PM_{2.5}$ Forecasts Using Summer of 2004 AIRNow $PM_{2.5}$ Data

[13] Latitudes and longitudes of 128 AIRNOW $PM_{2.5}$ monitoring stations falling within the domain of analysis are mapped into the model grid coordinates of each model. For all models observed $PM_{2.5}$ values are compared against model grid values that contain the monitor. In other words no spatial interpolation is performed on the model results, but depending on model resolution, $PM_{2.5}$ from several stations could be evaluated against results from only one model grid.

[14] The AIRNOW observations are reported as hourly averages centered on the half hour, and some temporal averaging of model results is necessary to allow for consistent comparisons. The CMAQ/ETA model provides results already averaged over these hourly periods. The WRF, STEM-2K3, CHRONOS, and AURAMS model results come as snapshots at the top of each hour. For these models the hourly average centered at the half hour is taken as the average of the two adjacent hourly snapshots. With data for all models available on a uniform time base an equal-weighted ensemble forecast is also generated by calculating the arithmetic mean of 6 individual models. The WRF-1 model is not included in the ensemble because of its older emission inventory and other deficiencies within its $PM_{2.5}$ formulation, as discussed further below. Ensembles based on the geometric and arithmetic means of the 6 models are also calculated, and results for these ensembles are also shown for comparison.

3.1. Observations and Details of Analysis of AIRNOW

[15] $PM_{2.5}$ data for the months of July and August of 2004 were provided by Sonoma Technologies Corporation through the EPA AIRNow program. All measurements are made using tapered element oscillating microbalance (TEOM) instruments, averaged over hourly intervals from the top of one hour to the next. It should be recognized that TEOM measurements are somewhat uncertain, and believed to be lower limits because of volatilization of soluble organic carbon species in the drying stages of the measurement [*Eatough et al.*, 2003; *Grover et al.*, 2005]. No attempt is made here to account for this uncertainty, and measurements are used "as is." The location of the 128 stations within the domain of model overlap is shown in Figure 2 along with some information from AIRNow on surrounding population.

[16] The 35 day period between 0000 Z 14 July 2004 and 0000 Z 17 August 2004 (0000 Z refers to Zulu, or Universal

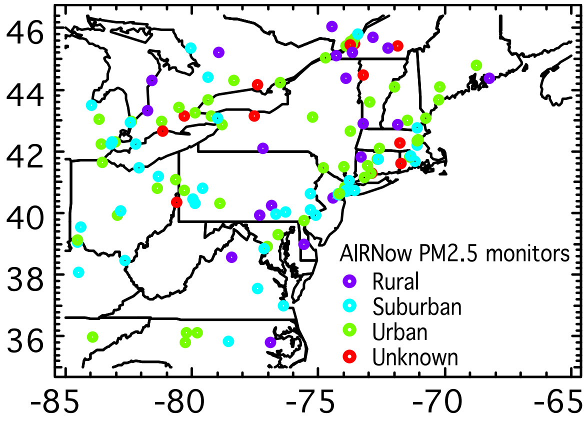

Figure 2. Location of AIRNow sites providing real-time $PM_{2.5}$ data to Sonoma Tech.

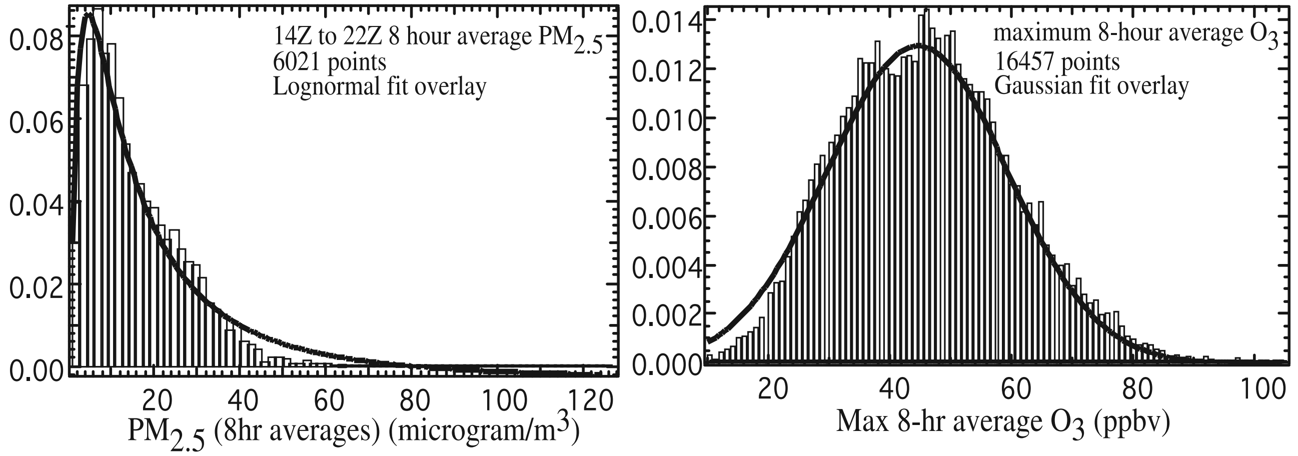

Figure 3. Sorted frequency histograms of $PM_{2.5}$ and O_3 between 6 July 2004 and 30 August 2004 for all monitors in the domain of Figure 2 (128 $PM_{2.5}$ monitors, and 358 O_3 monitors).

Time Coordinated) is the sampling period used in this analysis, corresponding to data availability of CMAQ/ETA. The statistical evaluation is for results from the 0000 Z forecasts, except for the CMAQ/ETA model, which only provided 1200 Z forecasts. One day (1 August) was missing from the CHRONOS 0000 Z forecast, leaving 34 days with coincident, and 6-member ensemble results. Two sets of analysis are presented here; one set giving spatial information on common statistical measures (r-coefficients, bias and RMSE), and another set that looks at average diurnal cycles for each model compared to observations.

[17] For the spatial analysis, daily values of 8-hour afternoon average $PM_{2.5}$ (1400 Z to 2200 Z), and daily average (midnight to midnight LT) are calculated from the hourly $PM_{2.5}$ observations and compared with similar averages from each model. There are three reasons for comparing afternoon averages. First, the observed afternoon values are representative of a larger footprint during this time of day, due to efficient boundary layer mixing, and therefore provide the best conditions for comparison to models having 12 to 42 km horizontal resolutions. Second, as will be discussed below with regard to diurnal variations, some models show distinct positive $PM_{2.5}$ biases during morning rush hour peaks in urban and suburban regions that tend to skew statistics that include the 0600 to 1000 LT data. Third, and importantly, it is of interest to compare statistical measures for $PM_{2.5}$ with those of O_3 within a similar diurnal context. The maximum 8-hour average O_3 is the quantity of interest in terms of regulatory and health advisory issues, and this has been the quantity used in previous statistical evaluations [*McKeen et al.*, 2005, and references therein]. The timing of the maximum 8-hour average O_3 values tend to center between 1400 and 1500 local time (LT) for the companion O_3 statistics presented here, and thus the choice of the 1400 Z to 2200 Z averaging window for $PM_{2.5}$. If a monitor has less than 6 hours of data available between 1400 Z and 2200 Z, the data for that day are discarded. Statistics for the 24 hour average are also shown here, since the 24 hour average is used for regulatory and compliance purposes by the U.S. EPA. Statistical results are calculated only for monitors and days when forecasts from all six models as well as observations are available, and are restricted to 118 sites that have 20 or more days of observations that fit the above criteria.

3.2. Spatial Analysis of 8-Hour $PM_{2.5}$ Averages

[18] Frequency distributions of $PM_{2.5}$ and O_3 are shown in Figure 3 for the observed 8-hour $PM_{2.5}$ averages and maximum 8-hour averages of O_3 between 6 July 2004 and 30 August 2004, and for monitors available within the domain shown in Figure 2. The frequency distribution for O_3 is nearly Gaussian and symmetric about a central value. The frequency distribution of $PM_{2.5}$ is decidedly non-Gaussian, and a chi-square analysis shows that its frequency histogram approximates a lognormal distribution more accurately than a standard Gaussian. For this reason statistical comparisons are performed on the log of $PM_{2.5}$ concentrations, rather than the concentration levels themselves, as is typically done for O_3.

[19] The following statistical measures are therefore used in the spatial analysis and are modifications to the standard bulk statistical measures used by *McKeen et al.* [2005] to account for log-scaling of $PM_{2.5}$:
the r-correlation coefficient:

$$\mathrm{r(i)} = \frac{\sum\limits_{days}\left(\Phi^{modl}(i,day)-\Phi^{modl}(i,avg)\right)\left(\Phi^{obs}(i,day)-\Phi^{obs}(i,avg)\right)}{\sqrt{\sum\limits_{days}\left(\Phi^{modl}(i,day)-\Phi^{modl}(i,avg)\right)^2\sum\limits_{days}\left(\Phi^{obs}(i,day)-\Phi^{obs}(i,avg)\right)^2}}, \tag{1}$$

the model/observed ratio;

$$\mathrm{Md/Ob\ \ Ratio(i)} = \exp^{\left\{\left(\frac{1}{\mathrm{N}_{days}}\right)\sum\limits_{days}\left[\Phi^{model}(i,day)-\Phi^{obs}(i,day)\right]\right\}}, \tag{2}$$

Table 1. $PM_{2.5}$ Statistics for 8-Hour Averages (1400 Z to 2200 Z) and Maximum 8-Hour Average O_3 From the AIRNow Surface Networks for the 118 $PM_{2.5}$ Monitors and 342 O_3 Monitors for the 14 July 2004 Through 17 August 2004 Time Period[a]

	$PM_{2.5}$, Log-Transformed, Statistics				O_3 Standard Statistics			
Institute, Model, Horizontal Resolution	r Coefficient	Modl/Obs Ratio	RMSE (Factor)	Skill, %	r	Bias, ppbv	RMSE, ppbv	Skill, %
NOAA/ESRL, WRF/CHEM-1, 27 km	0.42	1.17	2.19	33	0.67	14.3	20.9	24
NOAA/ESRL, WRF/CHEM-2, 27 km	0.64	0.81	1.97	64	0.73	3.4	11.6	61
NOAA/ESRL, WRF/CHEM-2, 12 km	0.54	0.64	2.38	40	0.67	11.9	16.6	31
MSC Canada, CHRONOS, 21 km	0.65	0.77	2.14	50	0.68	17.0	23.2	16
MSC Canada, AURAMS, 42 km	0.46	0.85	2.16	59	0.54	5.9	16.2	27
University of Iowa, STEM, 12 km	0.63	1.12	1.97	70	0.60	26.4	31.	2
CMAQ/ETA, 12 km	0.65	0.76	2.03	60	0.63	13.4	17.9	24
Six-model arithmetic mean ensemble	0.73	0.89	1.78	76	0.76	10.2	15.0	47
Six-model geometric mean ensemble	0.74	0.79	1.83	73				
Persistence (previous day observations)	0.38	1.0	2.13	50%	0.48	0.0	13.7	50

[a]All quantities except skill are median values for the number of monitors in each network.

and the ratio-equivalent root mean square error;

$$\text{ratio RMSE(i)} = \exp\left\{\sqrt{\left(\frac{1}{N_{days}}\right)\sum_{days}\left(\Phi^{model}(i,day)-\Phi^{obs}(i,day)\right)^2}\right\}, \quad (3)$$

where Φ is the natural logarithm of the $PM_{2.5}$ concentration, i refers to $PM_{2.5}$ monitor i (i = 1 to 118), N_{days} refers to number of observing days at each site, "obs" refers to observed, and "model" or modl refers to model value. It is important to note that the Md/Ob ratio and the ratio RMSE values are multiplicative equivalents to mean bias and RMSE, and a value of one for these quantities means no bias and perfect agreement, respectively. The ratio RMSE value is a measure of the distance from the one-to-one line on a log-log plot of model versus observations, or a multiplication/division factor, rather than an addition/subtraction constant associated with the standard RMSE.

[20] Table 1 gives median values of the log-scaled $PM_{2.5}$ statistical parameters for the various models and the two ensembles for all monitors shown in Figure 2. Also listed is a skill factor (in percent). This is defined as the percentage of monitor comparisons for a given model that has lower ratio RMSEs than for persistence (or simply the previous day's 8-hour average). A skill factor greater than 50% means more than half of the model results have lower mean errors compared to persistence, and therefore the model possesses some skill. Also included in Table 1 are equivalent statistics for O_3 [*McKeen et al.*, 2005] for the 342 AIRNow O_3 monitors that are within the domain limits of Figure 2, and for the same time period as the $PM_{2.5}$ comparisons. Because of the log scaling for the $PM_{2.5}$ statistics, mean biases and RMSE cannot be compared between O_3 and $PM_{2.5}$, but the r-coefficients and skill factors can be. The r-coefficients for $PM_{2.5}$ are significantly reduced compared to the O_3 r-coefficients for the three WRF/CHEM models and AURAMS, slightly reduced for CHRONOS, but slightly better $PM_{2.5}$ r-coefficients for the STEM and CMAQ/ETA models. The skill factors are always higher for $PM_{2.5}$ compared to O_3. As discussed by *McKeen et al.* [2005], the high O_3 bias in all models directly affect RMSE, keeping O_3 skill scores unacceptably low, but useful forecasts are still possible with simple bias corrections. The $PM_{2.5}$ forecasts, on the other hand, show low bias (within 25% of unity) for all models but the WRF/CHEM-2 12 km case, resulting in skill scores for all but 2 of the WRF/CHEM models that match or beat persistence. Also, similar to the results given by *McKeen et al.* [2005], the ensemble $PM_{2.5}$ forecasts show improved r-coefficients compared to any single model.

[21] The results of Table 1 show that 4 models (CHRONOS, CMAQ/ETA, STEM-2K3, and WRF/CHEM-2) are quite similar in terms of the three statistical measures, while the AURAMS, WRF/CHEM-1, and WRF/CHEM-2 (12 km) models show reduced performance. This is shown graphically with the summary statistics for r-correlation coefficients in Figure 4. Figure 4 also shows the obvious improvement of both ensemble calculations compared to the other individual models. The geometric mean ensemble appears to perform slightly better than the arithmetic mean ensemble, especially for lower values of the r-coefficient

[22] Figure 5 shows the spatial pattern of the statistical parameters in Table 1 for the arithmetic mean ensemble forecast. The ensemble statistics represent a collective understanding (or misunderstanding) of the suite of models, though patterns shown in Figure 5 are generally representative of individual models. Figure 5 shows a fairly heterogeneous distribution of r-coefficients, but overall low bias except in isolated urban source regions (New York City, Boston, Toronto, Detroit). The ratio RMSE is highest for

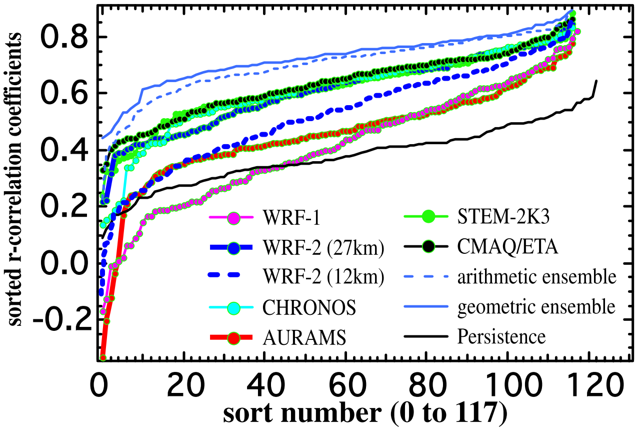

Figure 4. Sorted $PM_{2.5}$ r-correlation coefficients for the models, ensembles and persistence for the monitors in Figure 2 and data summarized in Table 1.

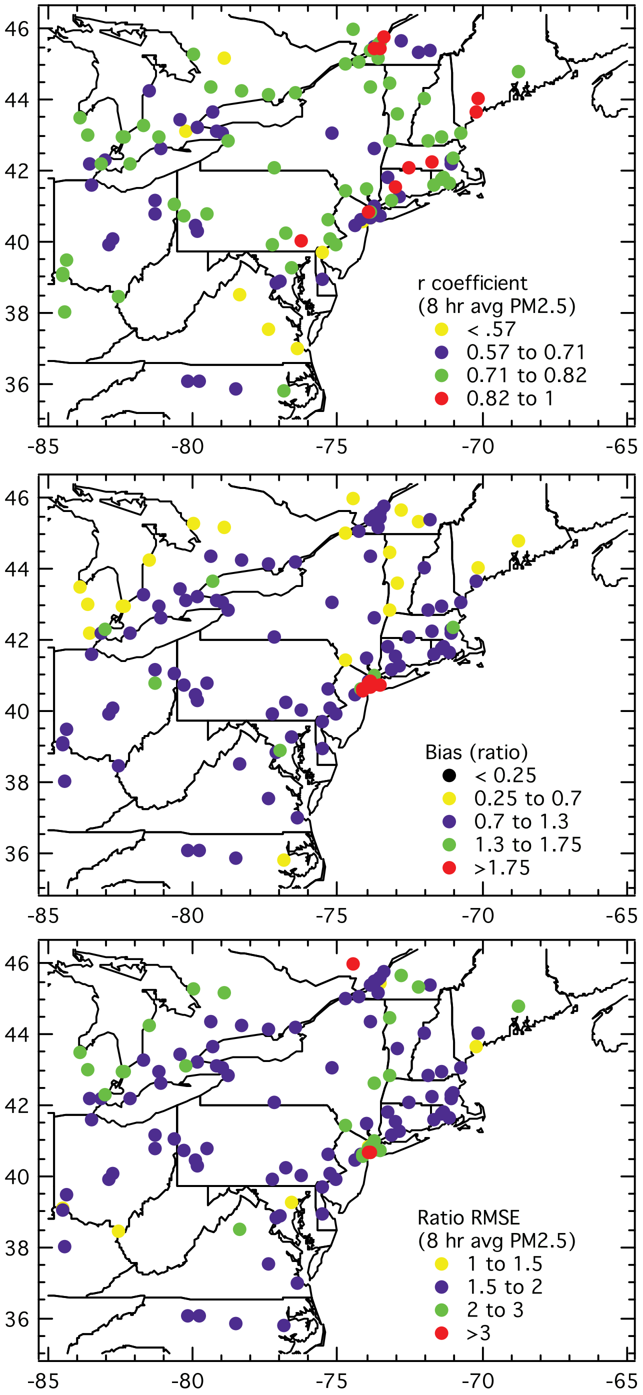

Figure 5. Spatial patterns of r-coefficients, Model/Obs. ratio, and the ratio RMSE for $PM_{2.5}$ from the arithmetic mean ensemble forecast.

locations with either very low or very high bias. Figure 5 suggests that as a whole, the $PM_{2.5}$ models are underpredicting $PM_{2.5}$ over the larger, regional scale, and significantly overpredicting $PM_{2.5}$ in only a few isolated urban regions. As mentioned previously, the measurements are from TEOM and are considered to be lower limits. The underprediction by models is likely more severe than this analysis suggests. As discussed later in section 4.2, one possible cause of the low bias is underestimation of particulate organic carbon (POC) by the models.

[23] The reason for differences in performance between the individual models needs further explanation that only the forecast centers can provide. There are valid explanations for the relative performance of the WRF/CHEM models. First, the WRF-Chem/1 emissions are based on the EPA NET-96 emissions inventory for the United States, and a 1985 base year inventory for Canada. WRF/CHEM-2 emissions are from the EPA NEI-99 inventory for the United States and a 2000 base year inventory for Canada. The older inventory used in WRF/CHEM-1 has urban $PM_{2.5}$ emissions more than a factor of 2 higher than in the NEI-99 inventory for the United States, and differences for Canadian cities are more on the order of a factor of 4. Additionally, the deposition velocity for $PM_{2.5}$ in WRF/CHEM-1 was specified incorrectly, giving too high of deposition velocities for this model. The WRF/CHEM-2 (12 km) model differs from the WRF/CHEM-2 (27 km) in horizontal resolution, the other difference between these models is with the parameterization of PBL transport. As discussed further below with regard to comparisons with aircraft data, the 27 km versions of WRF/CHEM use the original default for PBL transport, the Mellor-Yamada-Janjic (MYJ) scheme with order 2.5 closure [*Janjic*, 2002]. This parameterization was found to predict PBL heights that were too low, PBL temperatures too low, and high PBL water vapor and cloud biases when compared to NOAA WP-3 aircraft observations collected during the ICARTT/NEAQS-2K4 field experiment. For this reason the PBL scheme of *Hong and Pan* [1996] also known as the YSU (Yung Sun University) scheme, was used within the WRF/CHEM-2 (12 km) model. Using this scheme results in more consistent PBL heights, but with significant positive temperature and low relative humidity biases near the surface. All WRF/CHEM models use the same land surface model (LSM) of *Smirnova et al.* [2000], though the numerical coupling between the LSM and the two PBL parameterizations may require further inspection and refinement and may influence the PBL meteorological biases. Using the YSU scheme with deeper boundary layers makes overall $PM_{2.5}$ bias lower than the MYJ scheme, and it also appears to degrade the r-correlation comparisons.

[24] Because regulatory and compliance issues associated with $PM_{2.5}$ are based on 24-hour averages, it is useful to compare forecast statistics based on this quantity. Table 2 shows the equivalent statistics based on 24-hour averages (midnight to midnight LT) that are given for the 8-hour averages in Table 1. In general the correlation coefficients based on the 24-hour averages are not significantly different than those for the 8-hour averages, except for the persistence forecast which is much higher. Median model/observed ratios are noticeably higher for the 24-hour average comparisons for the WRF/CHEM-2 (12 km), CHRONOS and CMAQ/ETA models. As shown in the next section these three models overpredict the observed early nighttime and postsunrise peaks in $PM_{2.5}$, while the other models do not. RMSE factors within Table 2 are somewhat lower than those for the 8-hour average statistics in Table 1 for all models and ensembles, but the RMSE factor for the persistence forecast is significantly lower. The net result is that the skill (fraction of monitors with RMSE less than that of persistence) is reduced significantly for all models and ensembles. The persistence forecast is so successful for the 24-average

Table 2. $PM_{2.5}$ Statistics for 24-Hour Averages (Midnight to Midnight EST) From the AIRNow Surface Networks for the 118 $PM_{2.5}$ Monitors for the 14 July 2004 Through 17 August 2004 Time Period[a]

	$PM_{2.5}$, Log-Transformed, Statistics			
Institute, Model, Horizontal Resolution	r Coefficient	Modl/Obs Ratio	RMSE (Factor)	Skill, %
NOAA/ESRL, WRF/CHEM-1, 27 km	0.51	1.11	1.96	21
NOAA/ESRL, WRF/CHEM-2, 27 km	0.63	0.81	1.86	33
NOAA/ESRL, WRF/CHEM-2, 12 km	0.49	1.08	2.03	22
MSC Canada, CHRONOS, 21 km	0.69	0.94	1.96	28
MSC Canada, AURAMS, 42 km	0.51	0.93	1.92	36
U of Iowa, STEM, 12 km	0.69	1.15	1.77	51
CMAQ/ETA, 12 km	0.60	0.94	1.79	41
Six-model arithmetic mean ensemble	0.72	1.03	1.68	66
Six-model geometric mean ensemble	0.74	0.94	1.69	64
Persistence (previous day observations)	0.49	1.0	1.77	50

[a]All quantities except skill are median values for the number of monitors in each network.

$PM_{2.5}$ that only the model ensembles and one model forecast (STEM-2K3) possess forecast skill, using the definition of skill applied here.

3.3. Analysis of Diurnal Variations at the AIRNow $PM_{2.5}$ Monitors

[25] In this section the data from the 128 AIRNOW monitors are segregated into urban (56), suburban (42), and rural (20) stations (see Figure 2), and the average diurnal cycles for each of these three classes of stations are analyzed. Two sets of analysis are presented. One is based on a ‘‘diurnal factor’’ that is calculated according to a 24-hour running average centered on any given. The second is based on time the diurnal cycle of the geometric means of the hour-specific $PM_{2.5}$ concentrations themselves. The diurnal factor is defined as follows, for the model or observed value at hour t, and monitor i:

$$\text{Diurnal Factor}(t, i) = \frac{PM_{2.5}(t, i) \bullet 24(\text{obs./dy})}{\sum\limits_{t-12\text{hour}}^{t+12\text{hour}} PM_{2.5}(t', i)}. \tag{4}$$

Examples of the median and mean diurnal factors from the observations are shown in Figure 6.

[26] A large fraction of monitors is in the urban and suburban categories within Figure 2, and the diurnal factors for the suburban stations are nearly identical to the urban diurnal factors (and are not shown). The diurnal factors for all stations combined are very similar to the urban factor. The morning peak (1300 Z, or 0900 to 1000 LT), followed by a sharp decrease due to boundary layer growth and dilution, stands out in the urban category, and is much less apparent in the rural category. The peak in urban $PM_{2.5}$ at 0200 Z (2200 to 2300 LT) can be attributed to evening emissions within a stable surface layer, followed by $PM_{2.5}$ deposition once these emissions subside. The peak to valley differences of ~16% are much smaller than diurnal factors calculated for O_3 (peak to valleys from 1.6 to 0.4), and standard deviations of the $PM_{2.5}$ factors are very large (0.38 averaged over all hours). The median and mean ratio RMSE for persistence of the 8-hour average $PM_{2.5}$ values are both 2.1. This factor of 2 variability in the larger-scale day-to-day forcing clearly dominates the diurnal variability displayed by the averages in Figure 6. The opposite is true for O_3. The O_3 RMSE from persistence is ~9.5 ppbv, while average peak to valley differences are 30 ppbv.

[27] Figure 7 shows the model median diurnal factors overlaid on the diurnal factors from the observations in Figure 6. It should be kept in mind that these diurnal factors only provide a relative, normalized view of the impact that several processes are having on the diurnal cycle. One also needs to look at the absolute average diurnal variation, shown in Figure 8, to use the diurnal cycle information diagnostically. The unrealistically high deposition velocity for WRF/CHEM-1 shows up clearly in Figure 7, along with the overestimated urban and suburban emissions during the morning rush hour. The strong and persistent draw down of WRF/CHEM-2 (12 km) from early to midmorning reflects a strong nighttime inversion and trapped emissions followed by rapid PBL growth and subsequent dilution associated with the YSU boundary layer parameterization. The timing of the morning peak is quite different between models. For the urban and suburban averages the WRF/CHEM 27 km and Canadian models tend to peak an hour earlier than the observations, while the CMAQ/ETA and WRF/CHEM 12 km model peak 2 hours early, and the STEM-2K3 shows no morning buildup. All models except WRF/CHEM-1 tend to show the evening buildup, but CHRONOS (suburban only), WRF/CHEM-2 (12 km) and particularly CMAQ/ETA do not capture the decrease from 0100 to 0600 LT. The diurnal factors for the rural sites are not as clear cut, most likely because of the added importance of timing and meteorology that define source-receptor relationships at the limited number of rural sites (20).

[28] Figure 8 shows the diurnal geometric averages of the 0000 Z forecast models for all times available in the forecast. CMAQ/ETA only has the 1200 Z forecast available for this comparison. Figure 8 generally reflects the information in Figure 7, but there are a couple of additional interesting aspects within Figure 8. The observed median levels for rural sites are only 2 to 3 $\mu g/m^3$ less than the observed levels at urban/suburban sites, but all models show markedly lower rural $PM_{2.5}$ levels compared to those at urban/suburban sites. For the urban and suburban comparisons all models show a decrease in the second-day averages compared to the first, indicating trends in the biases. Previous NEAQS-2K2 analyses of MM5, WRF, and CMAQ/ETA meteorology also show trends in temperature and wind biases as well, but it is unclear how meteorological bias trends would affect $PM_{2.5}$ bias trends, and why all models would have the same trend. Finally, STEM-2K3 and the WRF/CHEM-2 models use the same anthropogenic emissions inventory for $PM_{2.5}$, SO_2, NOx, and VOC.

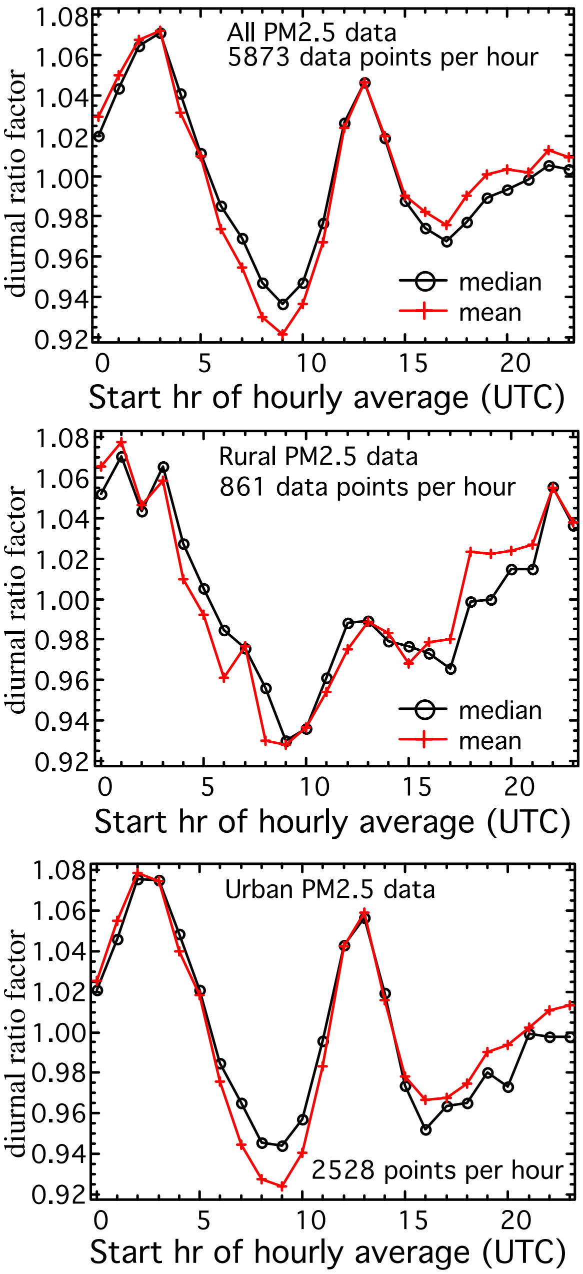

Figure 6. AIRNow observations only. Median and mean diurnal $PM_{2.5}$ factors (top) for all $PM_{2.5}$ monitors, (middle) for only rural monitors, and (bottom) only urban monitors shown in Figure 2.

However, there are significant difference in the diurnal profiles and averages, particularly for the rural sites. This suggests that model processes other than emissions are significant contributors to absolute $PM_{2.5}$ levels.

4. $PM_{2.5}$ Forecast Model Evaluation Using NOAA P-3 Aircraft Data

[29] During the summer of 2004 the ICARTT/NEAQS-2K4 intensive field study was operational for roughly the 5 July to 15 August time period. Mobile platforms participating in the field study include thirteen aircraft and the NOAA *Ronald H. Brown* research vessel (http://www.al.noaa.gov/csd/2004/rhbplatform.shtml). Two aircraft (the NASA DC-8 and the NOAA WP-3) along with the RV *Ronald H. Brown* included aerosol composition and aerosol size measurements within their payloads. The NOAA WP-3 aircraft in particular spent a significant fraction of its allotted flight hours immediately upwind, within and downwind of the New England region. The WP-3 aerosol

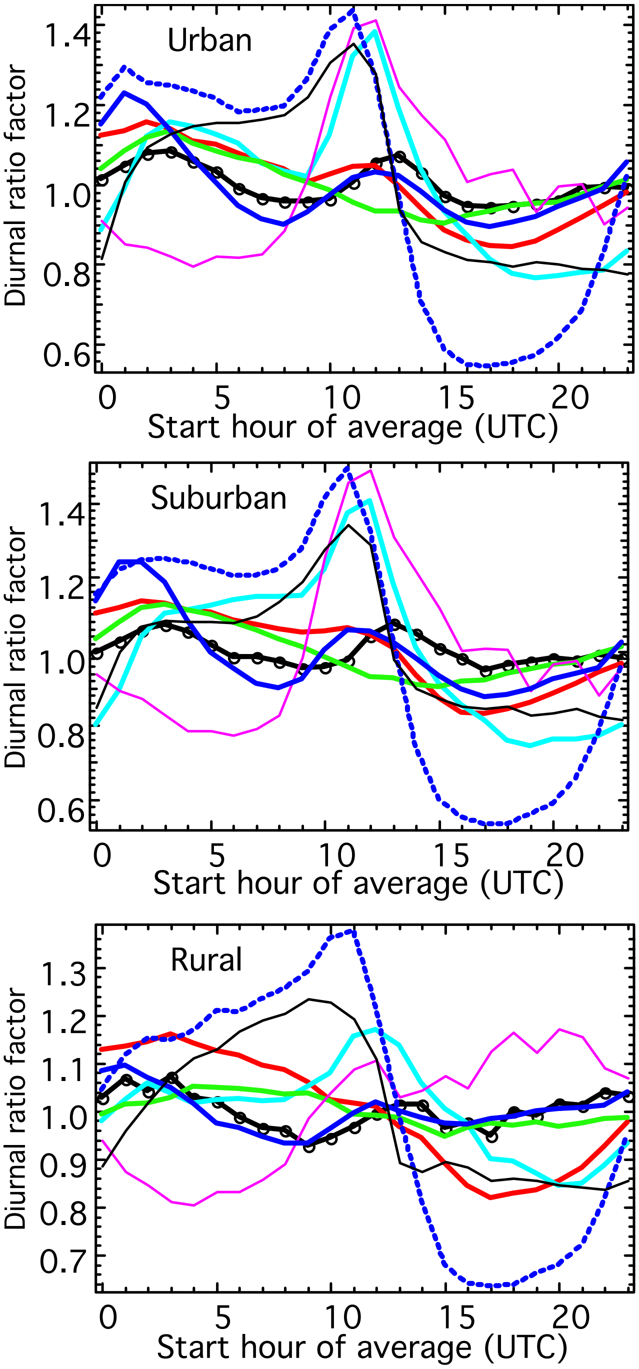

Figure 7. Median diurnal factors for the seven $PM_{2.5}$ forecast models compared to observations for the three different monitor categories. The color and line type assignments are the same as in Figure 4. Observed factors are given by the black line with open circles.

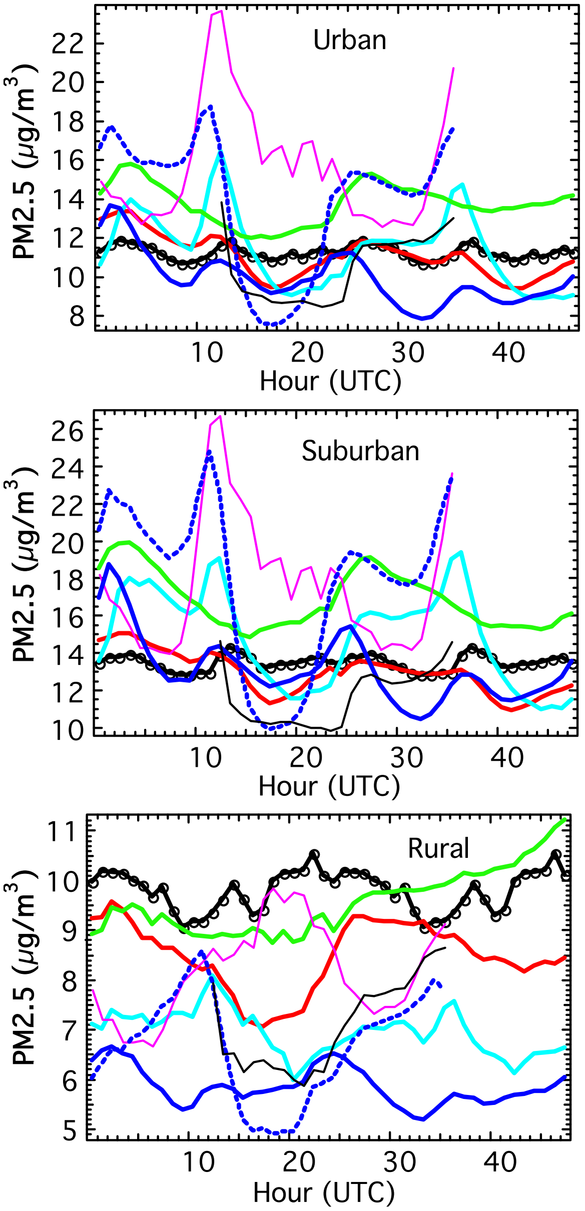

Figure 8. Diurnal averages (geometric means) for the seven $PM_{2.5}$ forecast models compared to observations for the three different monitor categories. The color and line type assignments are the same as in Figure 4. Observed factors are given by the black line with open circles.

composition data therefore provide valuable observations with which to compare forecast model $PM_{2.5}$ composition, particularly since information on the vertical distribution of aerosol species is very limited. Further details related to the design and coordination of the experiment, and the role of the NOAA WP-3 aircraft within the suite of mobile platforms can be found at http://www.al.noaa.gov/2004.

[30] In late March of 2005 most of the data sets collected by the NOAA W-P3 were finalized and made available to experiment participants. A public Web site (http://esrl.noaa.gov/csd/ICARTT/modeleval/) was constructed that overlays results from 9 AQ forecast models with results from the NOAA WP-3 aircraft and RV *Ronald H. Brown*. Thousands of plots are available at this site showing day-by-day comparisons for the *Ronald H. Brown* and every vertical profile and horizontal transect of the 17 NOAA WP-3 flights for several chemical, meteorological, aerosol and radiation variables (O_3, CO, NO, NO_2, NOy, HNO_3, PAN, NOx, NO_3, N_2O_5, SO_2, NH_3, $PM_{2.5}$ composition (SO_4^{2-}, NH_4^+, organic carbon (POC), NO_3^-, elemental carbon (EC)), total sulfur, isoprene, CH_3CHO, C_2H_4, C_3H_6, toluene, xylenes, CH_3COOH, temperature, virtual potential temperature, H_2O, relative humidity, wind speed, wind direction, solar radiation, J^{NO2}, sea surface temperature). Along with the detailed comparisons, a set of summary statistics for each of the nine models, and for each of the variables, is also provided for the NOAA WP-3 aircraft data. The following analysis is based upon the summary WP-3 statistics provided in the evaluation Web site, and the reader should consult this site for statistical comparisons of the individual models for the 23 variables that are not covered here.

4.1. Observations and Details of Analysis

[31] This section focuses on the comparison of $PM_{2.5}$ composition between six $PM_{2.5}$ forecast models and data collected on board the NOAA WP-3 aircraft. Size (and volume) distribution measurements of aerosol were made aboard the NOAA P-3 at 1 s time resolution by laser optical particle counters, similar to the measurements made during the ITCT 2002 field project [*Brock et al.*, 2004]. The aerosol size cutoff is ~1.0 μm diameter for this technique, a fixed refractive index for all particles is assumed, and a density of 1.6 g/m^3 is assumed in order to convert size distribution data to mass mixing ratio. It is therefore important to keep in mind that the $PM_{2.5}$ comparisons shown here have large inherent uncertainties (estimated to be 30–40%). The 90-s samples from the PILS (particle-into-liquid sampling) measurements [*Orsini et al.*, 2003; *Weber et al.*, 2001] form the observational basis of the aerosol composition comparisons, and six $PM_{2.5}$ forecasts are analyzed in terms of bias. All of the models evaluated in section 3, except the WRF/CHEM-1 model, are also evaluated in this section. Data from 15 flights between 15 July and 15 August 2004 and for flight tracks within the area of model overlap used by *McKeen et al.* [2005] are analyzed here in order to provide compatible statistics with that study and the preceding evaluation of the AIRNow $PM_{2.5}$ surface network.

[32] Numerically, comparisons are done by flying the aircraft through each model domain using the three-dimensional model field specific to each flight, and for the nearest hour of model output. If the aircraft flies through a model grid cell, the observational average is calculated for the time spent in that grid, and the model value at the nearest hourly time slice for that grid is also recorded, regardless of the time spent in the grid cell. If the sample time overlaps two model grids then both model grid values are compared against the observed average over the sample time. Similar to the surface comparisons described previously, there is no interpolation of model or observed data either in space or time in the comparisons. Further refinements to the comparisons should include a more rigorous way of limiting comparisons to well sampled grid cells, or weighting of averages according to time spent in grid. Here we rely on comparisons of median values or median errors which

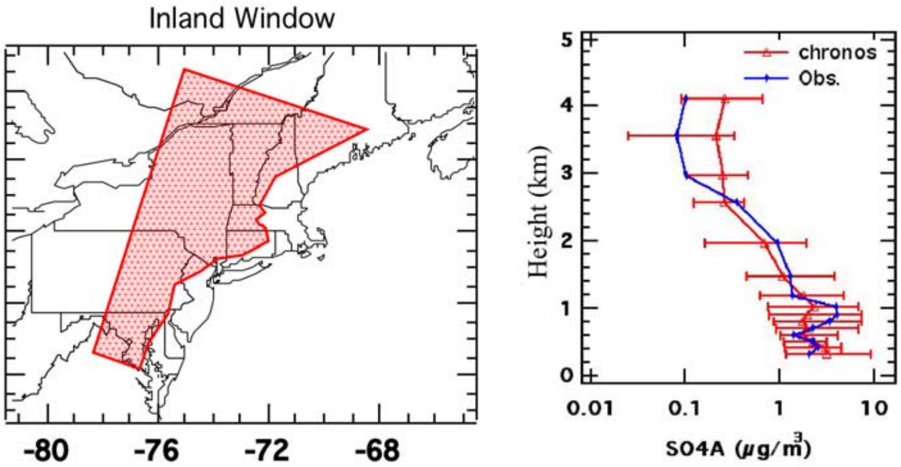

Figure 9. (left) Spatial extent of the inland-daytime window. (right) Median vertical profile of the observations (blue) for the inland-daytime (1100 to 1800 LT) window and the median model profile for CHRONOS (red) with the central two thirds of the model data given by the red bars.

should be relatively insensitive to these sampling and averaging issues.

[33] Median vertical distribution of $PM_{2.5}$ components for the 6 $PM_{2.5}$ forecast models and observations, as well as the median model over observed ratios, are given in the model evaluation web site (http://esrl.noaa.gov/csd/ICARTT/modeleval/) for three data windows; inland-daytime, coastal-daytime, and for all data. Figure 9 shows the limits of the inland-daytime window, and an example (CHRONOS) of $PM_{2.5}$ SO_4 median profiles taken from the Web site. This window is designed to compare profiles and model statistics for inland conditions unaffected by the coast. Since many flights were nighttime flights, and several flights focused on downwind pollution from Boston and New York, this window only comprises ~11% or 13 hours of NOAA WP-3 flight time from the 15 flights. Also excluded are measurements influenced by Alaskan and Canadian forest fires as determined by acetonitrile measurements from the PTRMS instrument [*de Gouw et al.*, 2006], mostly above 2 km altitude.

4.2. Results of NOAA WP-3 Comparisons

[34] Figure 10 shows the vertical distribution of median model over observed ratios for the inland-daytime window and for the six $PM_{2.5}$ forecast models. All models tend to overestimate SO_4 above 2 km altitude, and below this height three of the models overestimate SO_4 by a factor of 2 or more. Only the WRF/CHEM-2 12 km model underestimates SO_4 below 2 km significantly, though the coastal-daytime window (not shown) does not show the low bias or error gradient indicated in Figure 10 for this model.

[35] In order to relate upper air model comparisons with the statistics derived from the surface networks a more useful window of data is made by limiting comparisons to daytime (1100 to 1800 LT), over land and at lower altitudes (between 410 and 670 m above ground). The spatial distribution of this data set is shown in Figure 11. It comprises 4.5 hours or about 4% of available flight data. Data sampling is heavily weighted to the flights of 25 July 2004 (46%), 20 July (23%), 15 July (17%), and 22 July (9%).

[36] Figure 12 summarizes several comparisons related to the partitioning of sulfur and ammonia between the gas and particulate phase that are posted on the evaluation web page. The comparisons for total sulfur (SO_2 + SO_4, in ppbv) show that all models tend to overpredict total sulfur by a minimum of 40% and up to a factor 3. One possible explanation is that emissions inventories of total sulfur are simply too high. However, total sulfur for any given model is a balance between emissions, vertical transport, rainout and deposition. Notice that sulfur emissions within WRF/CHEM and STEM-2K3 are identical, and the emissions within AURAMS and CHRONOS are identical, but large differences in the median ratios are shown for models with similar emissions.

[37] For CHRONOS and the WRF/CHEM models the overestimates of total sulfur are associated with overestimates of SO_2. For AURAMS, CMAQ/ETA and STEM-2K3 the total sulfur estimate has a significant contribution from overestimated $PM_{2.5}$ SO_4. These three models include cloud oxidation of SO_2 into $PM_{2.5}$ SO_4 within their chemical mechanisms, while the other three models do not. Figure 12 therefore strongly suggests the 3 models with overestimated SO_4 have too much SO_2 cloud oxidation occurring in their simulations either from too much cloud, or too high an oxidation rate. Figure 13 illustrates this point more dramatically by comparing SO_2/(total sulfur) ratios for two representative models. The CMAQ/ETA model clearly underestimates this ratio, which is probably because of too fast or effective cloud oxidation. The WRF-Chem model underestimates SO_2 conversion to sulfate, but further model evaluation is needed to determine if this is because gas phase oxidation rates are too slow in this model (a defect in the photochemical formulation) or naturally occurring cloud oxidation is not included in the WRF/CHEM formulation, or both. Likewise, there is also the possibility that those models which include aqueous phase oxidation are overpredicting gas phase oxidation, or that the WP-3 aircraft undersampled cloud-processed air by attempting to avoid clouds. Further analysis focusing on the relative role of gas

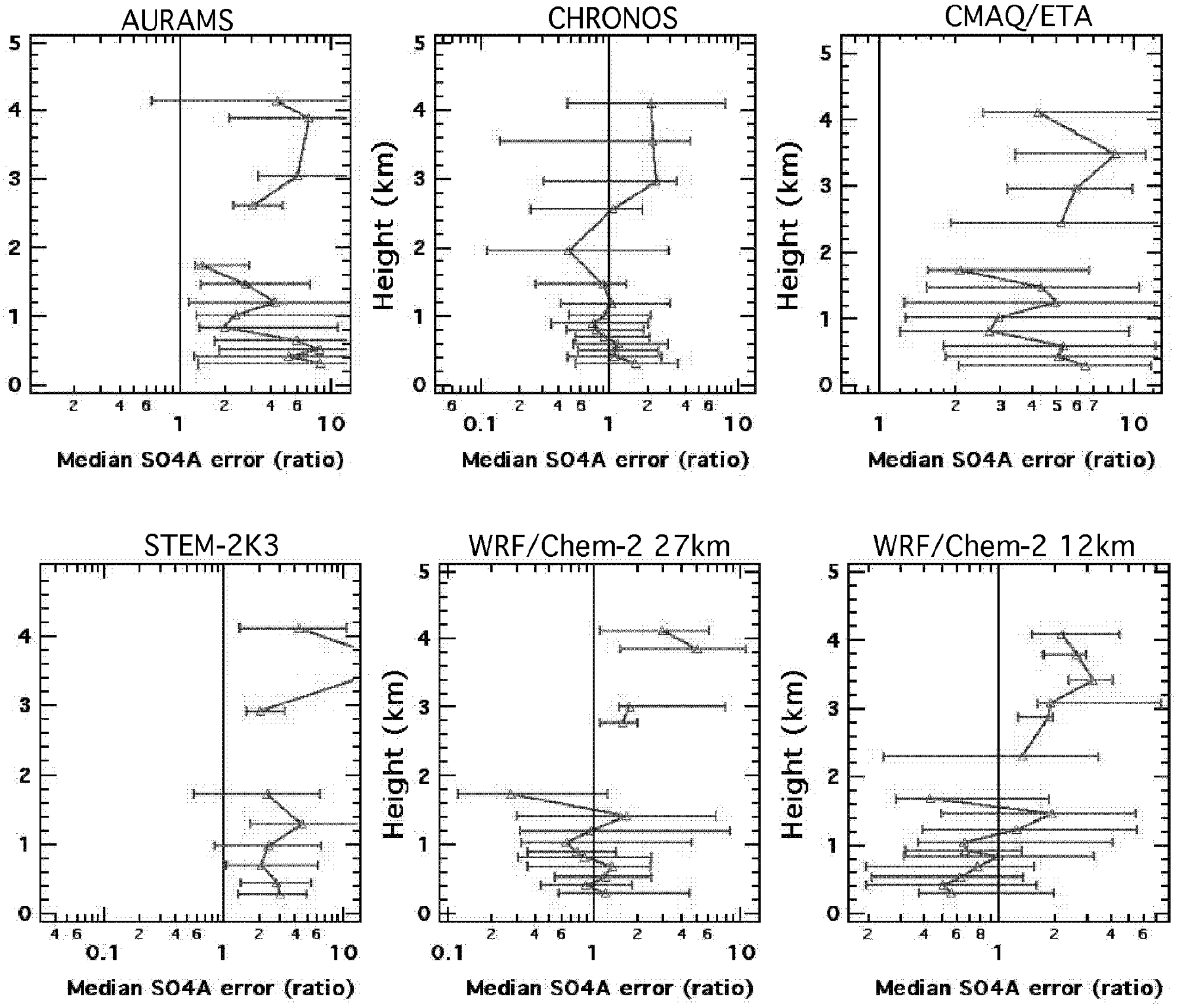

Figure 10. Vertical profiles of median model/observed ratios of $PM_{2.5}$ sulfate for the inland-daytime data window of the six forecast models. The central two thirds of the mode/observed ratios (within a sorted histogram) is given by the vertical bars. Only model vertical levels with more than 20 comparison points are shown.

phase versus aqueous phase sulfate formation is needed to unequivocally assign specific defects in these models.

[38] Figure 12 also shows the effect that model predicted sulfate levels have on ammonia partitioning between the gas and particulate phase. The CMAQ/ETA model has no gas phase ammonia as all ammonia partitioned to the aerosol phase due to the high sulfate coupled with the equilibrium assumptions within the aerosol formulation. It is important to note that the observations are almost always in an ammonia rich regime relative to ammonium sulfate ((NH_3 + NH_4)/SO_4 $\gg$ 2 on a molar basis) with a median ratio of 3.8. The high sulfate in CMAQ/ETA puts that model into an ammonia poor regime (median (NH_3 + NH_4)/SO_4 molar ratio of $\sim$ 1.). The WRF/CHEM-27 km model shows agreement for median NH_3 and NH_4 conditions, but large variation in model errors leads to very little correlation with the observations of total NH_3. Similar to the observations the WRF/CHEM-2 27 km and 12 km models are both in an ammonia rich regime with median (NH_3 + NH_4)/SO_4 molar ratios of 4.1 and 5.1, respectively. However, the ammonia partitioning in the WRF/CHEM-12 km model is clearly overpredicting the gas phase, and the only significant difference between the two WRF/CHEM models (other than horizontal resolution) is with the parameterization of boundary layer mixing. Temperature and water vapor biases are quite different between the two models with the 27 km WRF/CHEM-2 being too cold and humid on the median, which increases partitioning of NH_3 into the aerosol, and the 12 km model being too warm and dry, especially for daytime conditions of Figure 9 below 1 km (see the Web page). Under summertime conditions ammonia partitioning within WRF/CHEM is expected to have a strong dependence upon ambient humidity [e.g., *Yu et al.*, 2005]. Further analysis shows that for this data window the 27 km WRF/CHEM predicts 99.5% of the aerosol to be ammonia rich (NH_4/SO_4 molar ratios greater than or equal to 2), but the

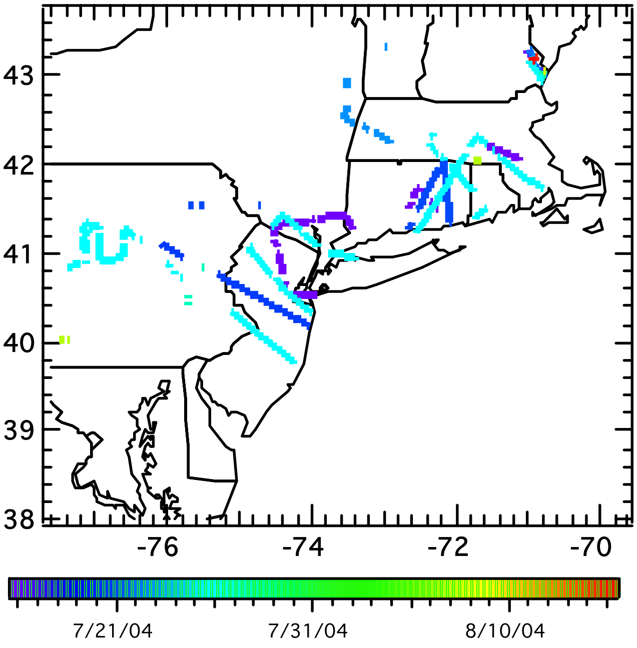

Figure 11. Location and time of NOAA W-P3 flights between 1100 to 1800 LT, between 410 and 670 m radar altitude, over land, and within the domain of model overlap.

12 km model predicts ~40% of the aerosol to be ammonia rich. For the inorganic equilibrium solution method the drier conditions of the 12 km WRF/CHEM limit the $PM_{2.5}$ NH_4 to sulfate ratio to less than 2, while the cold and moist conditions of the 27 km WRF/CHEM allow incorporation of NH_4 and NO_3 so that can exceed the NH_4 to SO_4 ratio can exceed 2.

[39] Comparisons of particulate nitrate and organic carbon (POC) with the PILS measurements are shown in Figure 14, along with the comparisons to total $PM_{1.0}$ determined from the laser optical counters. As discussed previously the $PM_{2.5}$ comparisons are somewhat uncertain because of the 1.0 μm cutoff of the sampling technique and the assumption of particle density used to convert aerosol volume to mass. Only two models (CMAQ/ETA and STEM) show correspondence between the $PM_{2.5}$ biases shown in Figure 14 and the biases from the AIRNow surface networks within Tables 1 and 2. AURAMS shows much higher biases aloft compared to the surface, while CHRONOS and the WRF/CHEM models show much lower biases aloft. Though the contribution of NO_3 to $PM_{2.5}$ mass is small and more than 80% the NO_3 measurements were below detection limit (0.1 $\mu g/m^3$), all models underpredict this component when measurements were above this limit. This is despite the fact that HNO_3 comparisons show that AURAMS, CMAQ/ETA and WRF/CHEM-27 km overestimate gas phase HNO_3 in the median. The numerical solution of $PM_{2.5}$ NO_3 is determined by equilibrium and ion balance conditions within the inorganic aerosol formulation with ammonium nitrate as the particulate phase intermediate. Although nitrate formation channels not included in the models are possible (i.e., secondary organic nitrate formation from VOC photochemistry), further analysis of model and observed $PM_{2.5}$ nitrate availability under conditions of strong acid displacement and ammonium limitations are necessary to verify or eliminate this possibility.

[40] The organic carbon comparisons in Figure 14 show significant underpredictions of this component for all models, but for AURAMS and the WRF/CHEM models in particular. The organic carbon errors in Figure 14 are upper limits, since the PILS instrument detects only soluble organic carbon. The WRF/CHEM models are expected to underpredict POC since the RADM2 photochemical mechanism does not include monoterpene photochemistry, which is a known biogenic pathway for POC formation. The RADM2 mechanism is likewise expected to underpredict secondary organic aerosol (SOA) formation from anthropogenic VOC oxidation relative to more up-to-date mechanisms (RACM and SAPRC-99) because of its limited treatment of anthropogenic VOC oxidation. The CMAQ/ETA and STEM-2K3 models do include biogenic SOA formation mechanisms but still underpredict POC by 40 to 50% for median conditions. As described by *de Gouw et al.* [2005], POC levels observed aboard the NOAA RV *Ronald H. Brown* during the 2002 NEAQS field study cannot be explained by currently accepted SOA mechanisms for both anthropogenic and biogenic conditions. Since POC slightly outweighs sulfate as the dominant aerosol component for the data window considered here, understanding the POC underpredictions in Figure 14 could help explain a large portion of the low $PM_{2.5}$ biases evident in the rural $PM_{2.5}$ comparisons in Figure 8, and also in Table 1.

5. $PM_{2.5}$ Forecast Model Evaluation Using the Surface Trends Network Data

[41] This section compares observations collected from the EPA Speciation Trends Network (STN) with model results from only three models, the two WRF/CHEM version 2 models (27 km and 12 km horizontal resolution), and the CMAQ/ETA model. As in previous comparisons the time period of comparison is between 14 July 2004 and 17 August 2004. The STN observations are 24-hour filter samples that are collected at over 200 sites throughout the United States every 3rd or 6th day through the U.S. EPA Air Quality System (AQS) (http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm). For the STN data used here field blank measurements that are sampler specific for the entire network were averaged to arrive at an average blank number, which was subtracted from the measured OC on a daily basis (V. Rao, U.S. EPA, personal communication, 2006). Figure 15 shows the location of the various STN monitor sites for the domain of model overlap applied in this study. Many of the 76 STN monitors are colocated with the $PM_{2.5}$ monitors shown in Figure 2.

[42] Comparison results for the three models are shown for five aerosol components in Figure 16. As in previous comparisons results are shown in terms of median model to observed ratios. When the results of Figure 16 are compared with the results of Figures 12 and 14, there are distinct qualitative differences between the upper air statistics and the STN surface statistics for the WRF/CHEM models, but no large qualitative difference for the CMAQ/ETA model. For CMAQ/ETA the ammonium overestimate and organic carbon underestimate are nearly identical between the two comparisons, while median biases are significantly reduced

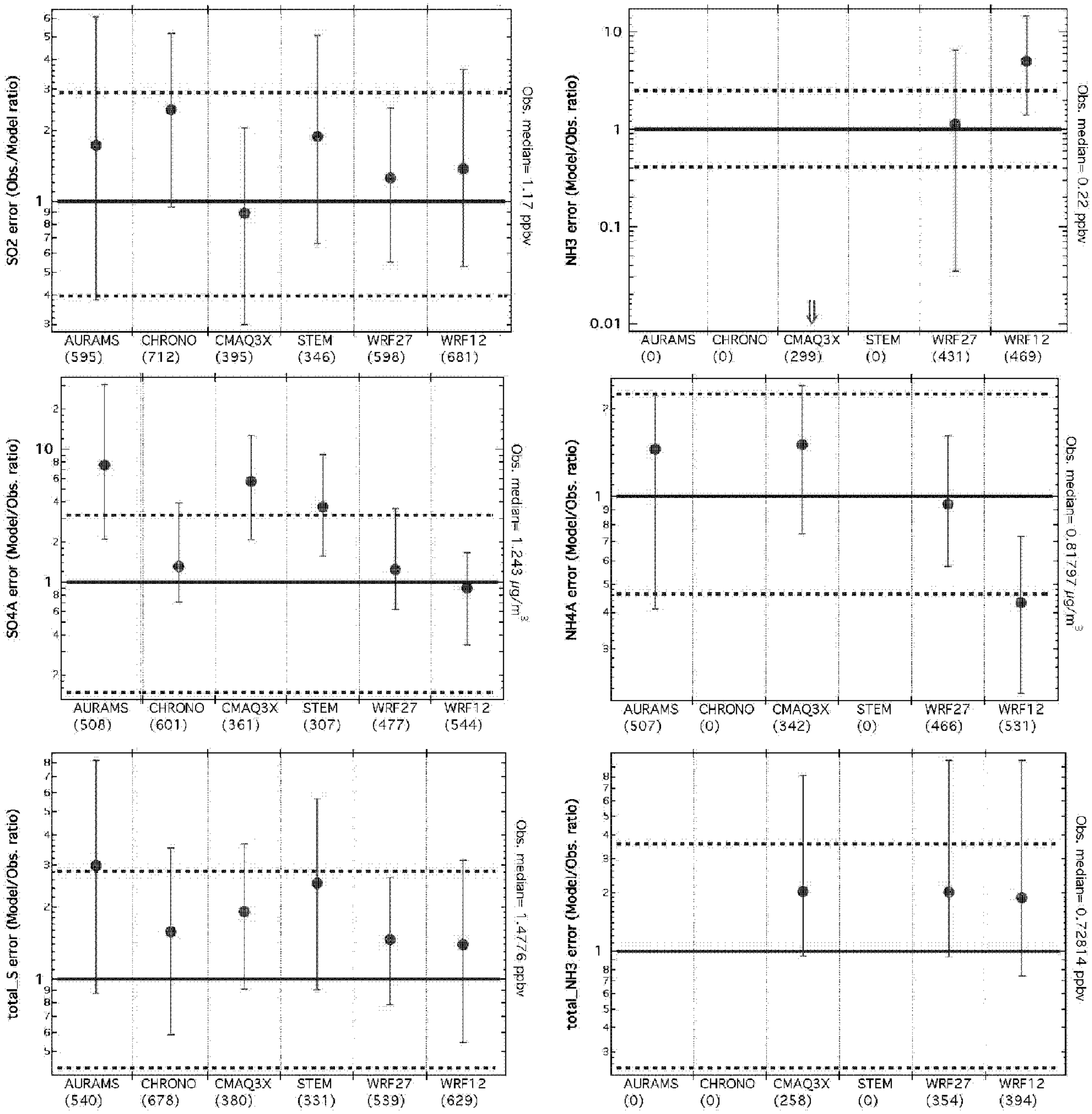

Figure 12. Median model to observed ratios (dots) for (top left) SO_2, (middle left) $PM_{2.5}$ SO_4, and (bottom left) SO_2 + $PM_{2.5}$ SO_4, (top right) NH_3, (middle right) $PM_{2.5}$ NH_4, and (bottom right) NH_3 + $PM_{2.5}$ NH_4. The central two thirds of the sorted model error distribution are the vertical limits. Dotted lines give the central two thirds of the sorted observation distribution relative to the observed median listed on the right of each panel. The comparison window is the 1500 Z to 2200 Z, 410 to 670, inland window shown in Figure 11. The number of comparison points in each comparison sample are listed below each model acronym.

for sulfate and nitrate. For the WRF/CHEM models organic carbon biases are similar to the upper air biases, but sulfate is biased very low and nitrate is biased very high for the STN comparisons, contrary to the upper air statistics in Figures 12 and 14. While the negligible ammonium bias is the same for the WRF/CHEM-27 km model, the WRF/CHEM-12 km model is biased high for the surface comparisons and biased low for the upper air comparisons.

[43] The differences between upper air and surface comparisons for the WRF/CHEM models can be explained in terms of the 24-hour averages used in the surface comparisons as opposed to only daytime comparisons for the upper air data. As discussed in section 3, the WRF/CHEM models tend to have shallow and nearly nonexistent mixing at night. Since SO_4 originates primarily from elevated point sources of SO_2, the nighttime decoupling of the surface and upper

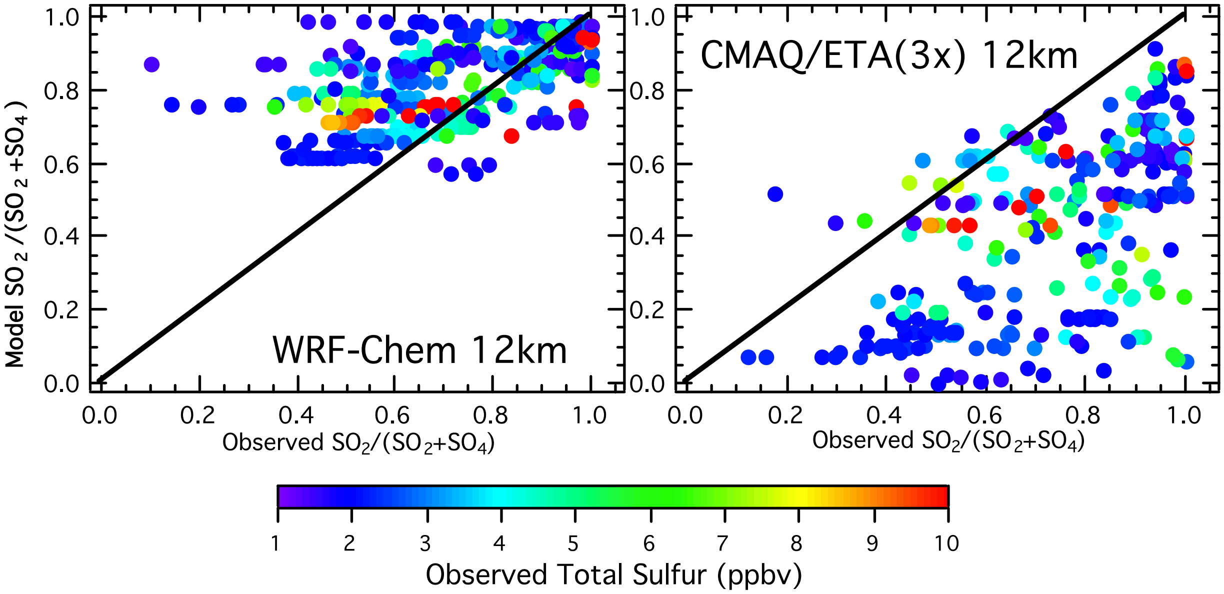

Figure 13. SO_2/(total sulfur) ratios for the WRF/CHEM-12 km and CMAQ/ETA models. The comparison data are for the same set used to derive Figure 12.

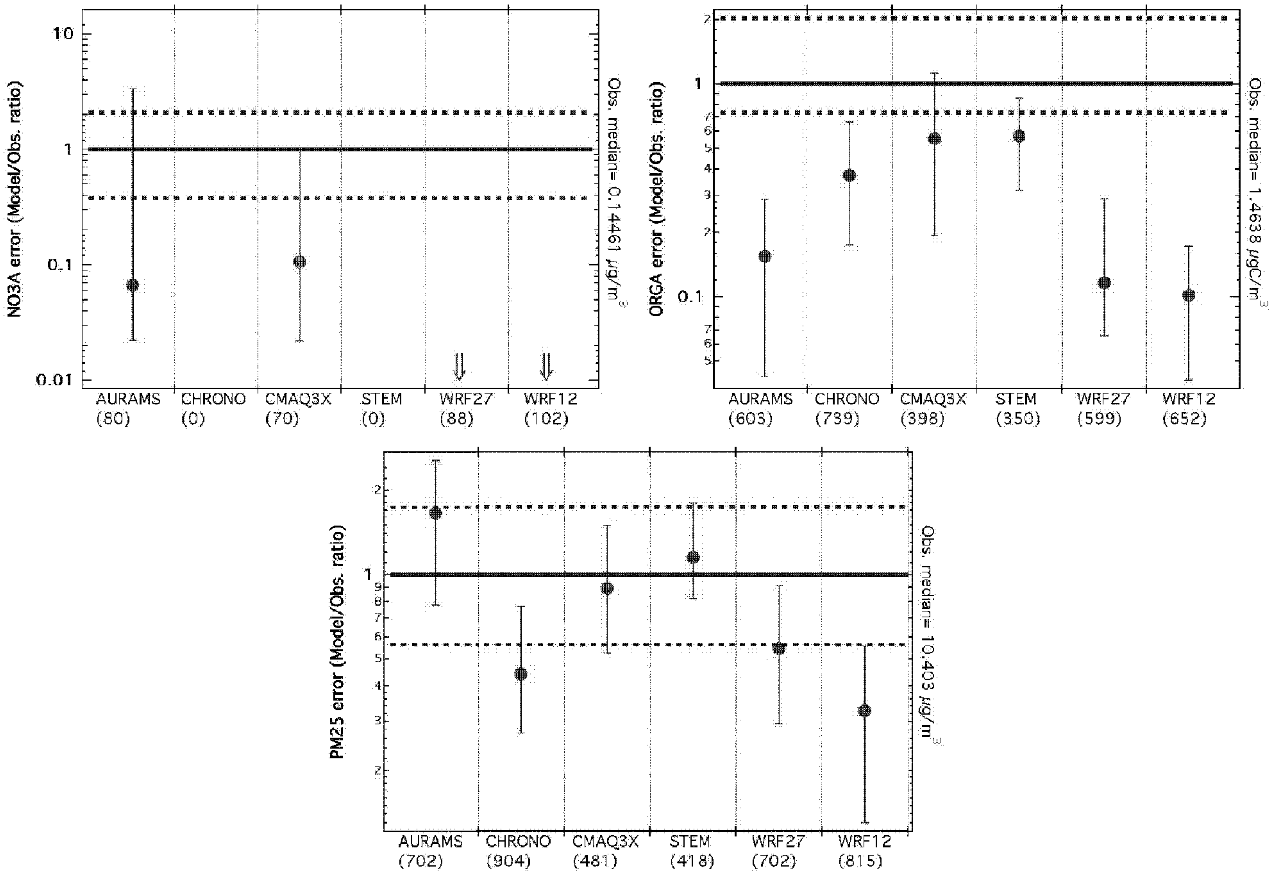

Figure 14. Median model to observed ratios (red lines) for (top left) aerosol NO_3, (top right) aerosol organic carbon, and (bottom) $PM_{2.5}$ mass derived from size distribution measurements. The comparison window, the meaning of lines, limits and values are the same as in Figure 12.

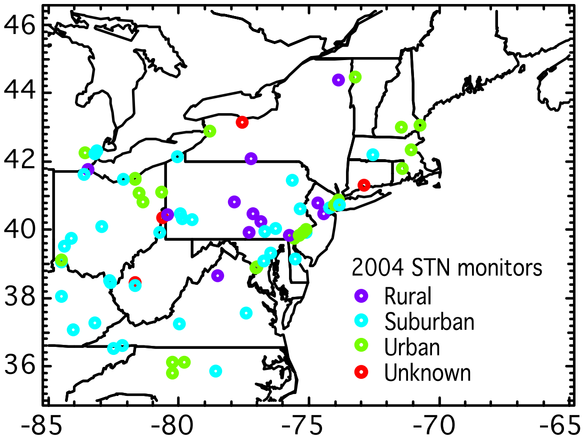

Figure 15. Location and population description of the 76 EPA STN surface sites used in the model/observation comparisons.

layers in the WRF/CHEM models deplete surface layer SO_4, and also traps surface emissions of EC, ammonia and NOx. The low sulfate, high humidity, ammonia, and gas phase NO_3 all contribute to ionic-equilibrium calculations favoring nitrate formation in the nighttime surface layers of the WRF/CHEM models. Ion balance diagrams show the WRF/CHEM to be completely neutralized with nitrate as the dominant anion. CMAQ/ETA shows much smaller nighttime ammonia levels, acidic $PM_{2.5}$ with sulfate as the dominant anion, and very little ammonium buffering available for inclusion of nitrate.

[44] Median values of six model $PM_{2.5}$ constituents are compared to observed medians in Table 3 for two separate categories of monitors in Figure 15. Medians for suburban and urban monitors are very similar so results for these two categories are combined, and compared to median rural conditions in Table 3. The relative medians between models and observations are generally representative of the median ratios in Figure 16, and these medians can be considered rough estimates of the median aerosol composition for the observations and models. Also listed for the models are median primary $PM_{2.5}$ levels, which represent inert, unspeciated primary $PM_{2.5}$ included in the emission inventories of the models. It is important to note that organic carbon in Table 3 is expressed in terms of carbon only, and that associated hydrogen and oxygen makes its relative contribution to $PM_{2.5}$ mass larger by anywhere from 20 to 70% for the models, and an unknown fraction for the observations.

[45] Table 3 again illustrates the low organic carbon fraction within the models compared to observations. However, a fraction of primary unspeciated $PM_{2.5}$ emissions are expected to be in the form of organic carbon, which complicates these direct comparisons. The CMAQ/ETA model also shows too high a sulfate fraction, but otherwise appears to match observed composition levels and fractions quite well, especially compared to the WRF/CHEM models. Both of the WRF/CHEM models underpredict sulfate, have nearly half the mass contributions from primary unspeciated $PM_{2.5}$, and severely overpredict nitrate levels. As discussed above, the combination of lack of vertical mixing, the shallowness of the lowest model level (15 m), and evening emissions all contribute to unrealistic surface concentrations that affect 24-hour averages, which also contributes to the high primary

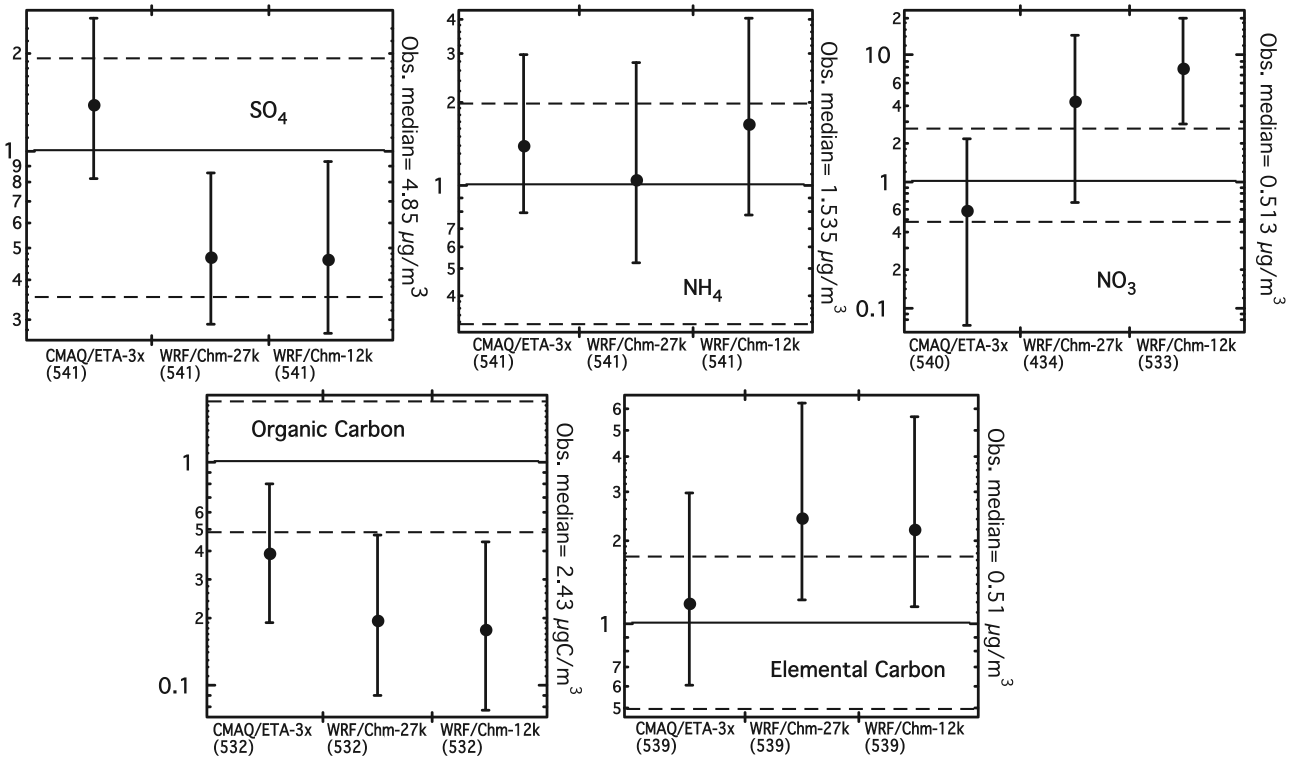

Figure 16. Median model to observed ratios of $PM_{2.5}$ constituents for all data from the STN sites shown in Figure 15. The meaning of lines, limits, and numerical values are the same as in Figure 12.

Table 3. Median Values for 24-Hour Averages From the STN Surface Network for the 60 Suburban and Urban Monitors, and the 11 Rural Monitors in Figure 15[a]

	Urban and Suburban				Rural			
$PM_{2.5}$ Component	ETA/CMAQ	WRF/CHEM 27 km	WRF/CHEM 12 km	Obs.	ETA/CMAQ	WRF/CHEM 27 km	WRF/CHEM 12 km	Obs.
Sulfate	6.7	2.2	2.1	4.8	7.8	2.1	1.9	5.4
Ammonium	2.0	1.7	2.5	1.6	2.0	1.4	2.2	1.5
Nitrate	0.4	3.3	4.7	0.5	0.3	2.1	3.8	0.4
Organic carbon	1.0	0.5	0.4	2.6	0.6	0.3	0.2	1.5
Elemental carbon	0.6	1.3	1.2	0.6	0.3	0.6	0.4	0.3
Primary $PM_{2.5}$	2.7	8.2	7.4		1.7	5.5	4.3	

[a]All units are in μg/m^3 except for organic (and elemental) carbon, which is in μg-carbon/m^3.

$PM_{2.5}$ and elemental carbon medians shown in Table 3. Looking only at the observations, the organic to elemental carbon ratios are nearly identical for the rural and the urban/suburban categories. Since natural sources of elemental carbon are thought to be negligible, the close association of aerosol organic carbon with elemental carbon suggests an anthropogenic origin of aerosol organic carbon for this particular set of eleven rural STN sites. However, Figure 15 shows that a majority of these "rural" sites are located close to urban or suburban settings. Biogenic contributions to organic carbon for truly rural conditions may therefore be underrepresented by the median values shown in Table 3.

6. Conclusions and Recommendations for Improved $PM_{2.5}$ Forecasts

6.1. Total $PM_{2.5}$ Mass

[46] When afternoon, 8-hour average $PM_{2.5}$ forecasts are log-transformed (justified by the lognormal $PM_{2.5}$ probability distribution functions), the $PM_{2.5}$ comparisons show reasonable biases (on the order of 20% or less for 6 out of 7 models), and correlations with observations that are comparable or exceed similar correlations in ozone. Five of the seven models exhibit low $PM_{2.5}$ bias, while only the earliest version of WRF/CHEM (using older emissions) and the STEM-2K3 model show high bias. As discussed in section 2.2, one caveat to the analysis is that effects of Alaskan/Canadian wildfires may have on $PM_{2.5}$ levels and model comparisons in the northeast United States. While the choice of dates used this study should minimize their effect, small but significant contributions during the last half of June cannot be completely discounted. If skill is defined in terms of fraction of comparisons having lower RMSE than persistence forecasts, all $PM_{2.5}$ forecasts show more skill than the O_3 forecasts, largely because of the relatively higher biases for the case of O_3 forecasts. From the standpoint of summertime afternoon average $PM_{2.5}$ concentrations, this would indicate that current $PM_{2.5}$ models are just as useful and accurate as current O_3 forecast models. Statistics based on 24-hour average $PM_{2.5}$ levels generally reflect those of the 8-hour averages. However, median model/observed ratios increase significantly for those models that overpredict the $PM_{2.5}$ peak (0600 to 0800 LT) associated with emissions in the stable early morning boundary layer. The skill of the 24-hour average forecasts is significantly less for all models and ensembles compared to the 8-hour forecasts. This is due to the definition of skill (RMSE relative to previous day's persistence forecast), and the fact that persistence forecasts for 24-hour averages have much lower RMSE (and higher correlation) than for the afternoon 8-hour persistence forecasts.

[47] There are simple techniques that can be used to improve the $PM_{2.5}$ forecasts. Similar to results found for O_3, ensembles of the $PM_{2.5}$ forecasts show significant statistical improvement over any individual forecast. Though not presented here, analysis of bias corrected $PM_{2.5}$ forecasts have also been performed where corrections have been applied analogous to previous O_3 bias removal studies [*McKeen et al.*, 2005; *Wilczak et al.*, 2006]. For all models and ensembles bias correction offers additional improvement in RMSE and forecast skill.

[48] Analysis of the diurnal cycles from the AIRNow $PM_{2.5}$ monitors and comparison with model median diurnal cycles illustrates some inconsistencies with certain processes within the models and the observations. For example there is very little diurnal variation in the median observed diurnal cycles, but significant diurnal variability exhibited by the CMAQ/ETA, WRF/CHEM-1, WRF/CHEM-2 (12 km), and CHRONOS models. The variability within WRF/CHEM-1 can be attributed to older emission inventories overpredicting urban $PM_{2.5}$ emissions. The WRF/CHEM-2 (12 km) variability appears to be due to the utilization of the YSU PBL parameterization within the WRF formalism. Further investigation is needed to understand the reasons for the high diurnal variability in the CMAQ/ETA model results, which could be due to the adopted PBL parameterization. The diurnal cycles of both CMAQ/ETA and WRF/CHEM-2 (12 km) show a reduced role for aerosol loss, presumably through surface deposition, during the late night and early morning hours. The timing of midmorning drawdown associated with PBL growth is also quite different between all models and observations for the urban and suburban monitors, and is typically too early by 1 or 2 hours.

[49] The TEOM instruments used in the analysis are uncertain and probably represent lower limits to actual mass loadings. Reducing measurement uncertainties of $PM_{2.5}$ mass is a prerequisite to any effective model evaluation studies, as well as compliance, trends, or management issues. It would be very useful to have $PM_{2.5}$ estimates from the NOAA mobile platforms to contrast and compare with the aerosol speciation data. Although size distribution measurements are available from these platforms, there is no direct information on particle densities to relate particle mass and size distributions unambiguously.

6.2. Aerosol Composition

[50] From the NOAA WP-3 comparisons two inconsistencies between models and observations are immediately

apparent: (1) Organic carbon is underpredicted by all models and (2) sulfate is overpredicted by those models that include sulfate formation from SO_2 cloud oxidation within their formulations. Particulate organic carbon (POC) and sulfate comprised the two largest components of dry $PM_{2.5}$ mass for almost all conditions encountered by the NOAA WP-3 during the field program, illustrating the importance for $PM_{2.5}$ forecast models to correctly characterize their relative contributions. One complication in comparing model $PM_{2.5}$ composition with observations is the specification of a primary, unspeciated $PM_{2.5}$ component within the emissions inventory. This is usually a significant portion of the model $PM_{2.5}$ mass, and model POC shortfalls could easily be explained by less than 50% of this primary emission component.

[51] The models with smallest median underprediction of POC ($\sim$45% for both CMAQ/ETA and STEM-2K3) include more detailed mechanisms of secondary organic aerosol (SOA) formation than the WRF/CHEM models. The WRF/CHEM model with RADM2 photochemistry does not include emissions or oxidation of biogenic monoterpenes, which are known to be a significant source of SOA. The RADM2 mechanism is also somewhat dated in terms of the lumping scheme used for anthropogenic VOC oxidation, and more efficient SOA formation from anthropogenic VOCs can be expected from more up-to-date mechanisms (e.g., RACM, CBM-Z, SAPRC). Improvements in WRF/CHEM's ability to reproduce POC levels should include an upgrade to the current RADM2 mechanism.

[52] The fact that all models underpredict POC by at least a factor of 2 demonstrates a need for further model investigation, further updates and awareness of recent experimental studies related to SOA formation, and further analysis of the observations to indicate if the missing model POC is of primary or secondary origin, or is from anthropogenic or biogenic sources. The STN data shows a correlation between POC and elemental carbon (EC) when contrasting urban/suburban and rural sites, suggesting a significant anthropogenic component of POC at the 11 "rural" STN sites considered here. Previous analysis of secondary versus primary has shown a large fraction of POC is secondary during the summer in the northeast [e.g., *Yu et al.*, 2004; *de Gouw et al.*, 2005]. Elemental carbon is a useful marker for anthropogenic $PM_{2.5}$ sources, and is usually highly correlated with CO measurements. We therefore recommend analysis of EC and CO measurements during intensive field missions, along with other markers, in order to provide insight into biogenic versus anthropogenic, and primary versus secondary sources of POC for conditions downwind of select urban and rural locations.

[53] All models tend to overpredict total sulfur (sulfate plus SO_2) as well as total ammonia (aerosol NH_4 plus gas phase NH_3) within the lowest 1 km and for the data windows considered here. One explanation is that emissions of sulfur and ammonia are simply overestimated. However, transport mechanisms, rainout and deposition processes in the models also influence these combined species. The results shown here (Figure 12) suggest that model differences affecting sulfur loss are significant, since two or more models having the same sulfur emissions can have a factor of two difference in the median model error of total sulfur. Whether these model differences are related to model parameterizations of transport, deposition and rainout, or due to meteorological biases (e.g., differences in total rain fall) is an open question.

[54] Those models that include SO_2 cloud oxidation as a source of sulfate clearly overpredict $PM_{2.5}$ sulfate (Figures 12 and 13) with CMAQ/ETA and AURAMS having more than a factor of 5 median overprediction of aerosol sulfate compared to aircraft measurements. Comparing the SO_2 to total sulfate ratios with aircraft data shows that SO_2 conversion is occurring too rapidly or efficiently in the CMAQ/ETA model. The surface STN data comparisons show that sulfate is overpredicted by $\sim$40% in the median at the surface. One can infer that the source of this overprediction at the surface is due to the rapid SO_2 conversion in the upper levels seen in the WP-3 comparisons. The WRF/CHEM models, which only include gas phase conversion of SO_2 to sulfate, show too slow or inefficient SO_2 conversion in the upper levels. In order to estimate the degree to which cloud versus gas phase SO_2 oxidation actually occurred, it is recommended that addition evidence, possibly through hydrocarbon relationships from WP-3 observations and model results, be analyzed to look for consistency between inferred gas phase VOC oxidation rates and the SO_2 to total sulfur ratios. Additional analysis of the relative importance of gas phase versus aqueous phase sulfate production in those models that include aqueous phase production is also needed to unequivocally assign specific defects their formulations.

[55] The high sulfate levels at upper levels in CMAQ/ETA are reflected in the NH_3 and NH_4 errors (Figure 12) with all available NH_3 incorporated into the particulate phase. The aircraft observations used in the daytime, inland 410 to 670 m window are ammonia rich relative to ammonium sulfate (median $(NH_3 + NH_4)/SO_4$ molar ratios $\approx$ 3.8), but the CMAQ/ETA results are ammonia poor (median $(NH_3 + NH_4)/SO_4$ molar ratio $\approx$ 1.) because of the high sulfate loading. The 40% median overprediction of NH_4 for CMAQ/ETA in the STN comparisons is roughly consistent with the 40% overprediction of sulfate and the degree of neutralization by NH_4 seen in the observations. The NH_3/NH_4 partitioning between gas and particulate phase for the WRF/CHEM (27 km) model seems very consistent with the observations for the daytime 410 to 670 m window, but the 12 km WRF/CHEM model shows nearly all partitioning in the gas phase. The reason for this appears to be due to a sharp sensitivity to ambient humidity in the equilibrium solution of NH_3/NH_4 partitioning. Particulate NH_4 is limited to ammonium sulfate ratios for the drier conditions of the 12 km WRF/CHEM model, while the moister conditions of the 27 km WRF/CHEM allows NH_3 incorporation into the particulate phase above the fully neutralized limit. This sensitivity of particulate NH_4 to different thermodynamic states, induced by the different PBL mixing schemes, could be an artifact of the equilibrium calculations for ammonia, nitrate and sulfate, and deserves further investigation.

[56] Observations of particulate NO_3 by the WP-3 aircraft show low concentrations and small contributions to total particulate mass, and were often times below the detection limit of the PILS instrument. For times when observed nitrate was appreciable all models tend to underpredict the observations. *Yu et al.* [2005] show that summertime nitrate

levels at the surface are highly dependent on total sulfate and total ammonia for conditions near the surface, and the overpredicted sulfate would certainly account for the low bias of NO_3 in CMAQ/ETA for both the aircraft and surface STN monitor comparisons. The total absence of aerosol NO_3 in the comparison of the WRF models (Figure 12), despite having gas phase ammonia and HNO_3 available, is not consistent with the limited observations. This could be due to equilibrium assumptions within the inorganic aerosol portion of the WRF/CHEM formalism, and further analysis is needed to confirm this and identify the underlying reasons.

6.3. PBL Mixing Parameterizations

[57] It is clear from raw model output, and from model comparisons of primary emitted species (CO, NOy, VOC) with observations taken aboard the *Ronald H. Brown* (see the model evaluation Web page) that the WRF/CHEM models have very little turbulent exchange below 200 m during stable nighttime conditions, allowing pollutants to build up to unreasonably high levels in the lowest model level. Exacerbating the situation, the lowest model level of WRF/CHEM is 15 m thick, which is half to a third of the thickness of the other models considered here. As an online model system, WRF/CHEM is forced to use the same vertical mixing parameterization for the aerosol and chemical species that is used for sensible heat and water vapor transport in the thermodynamic portion of the weather/chemistry forecasts. It therefore appears that PBL parameterizations applicable for meteorological forecasting within WRF are not entirely consistent with PBL mixing of chemical and aerosol components. The limitations and artifacts of the PBL parameterizations within WRF/CHEM during the night are detrimental to the 24-hour average surface comparisons with the STN data by reducing the transport to the surface of sulfate formed aloft, and by trapping surface emissions of elemental carbon, as well as NOx that contributes to high nitrate levels in the lowest model level. Until this fundamental inconsistency in the nocturnal PBL mixing within WRF/CHEM is identified and resolved, model comparisons that rely on nighttime surface model values are compromised.

[58] **Acknowledgments.** This research is partially funded by Early Start Funding from the NOAA Office of Atmospheric Research Weather and Air Quality Program and the NOAA/NWS Air Quality Forecast Program, and would not be possible without the participation of the AIRNow program and participating stakeholders. Computational and logistic assistance from the following individuals and organizations is also gratefully appreciated: Eirh-Yu Hsie (NOAA/ESRL/CSD), Jenny Fox (NOAA/ESRL/CSD), Amenda Stanley (NOAA/ESRL/GSD), Sophie Cousineau (MSC), L.-P. Crevier (MSC), Stephane Gaudreault (MSC), Mike Moran (AQRB/MSC), Paul Makar (AQRB/MSC), Balbir Pabla (AQRB/MSC), and the NOAA/ESRL High Performance Computing Facility. Special thanks to Paula Davidson (NOAA/NWS/NCEP) and S. T. Rao (EPA/NERL) for suggestions and comments and to Tim Dye and Alan Chan of Sonoma Tech for making the AIRNow $PM_{2.5}$ data available. The NOAA WP-3 comparisons presented here are only possible through the measurements made by John Holloway (SO_2, NOAA/ESRL/CSD), John Nowak (NH_3, NOAA/ESRL/CSD), Chuck Brock (aerosol size, NOAA/ESRL/CSD), and Rodney Weber (PILS aerosol composition, Georgia Tech/EAS). The comprehensive review of Prakash Bhave (EPA/NERL/AMD/AMDB) contributed significantly and is greatly appreciated. One author's (R.M.) research presented here was performed under the Memorandum of Understanding between the U.S. Environmental Protection Agency (EPA) and the U.S. Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) and under agreement 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

Ackermann, I. J., H. Hass, M. Memmesheimer, A. Ebel, F. S. Binkowski, and U. Shankar (1998), Modal aerosol dynamics model for Europe: Development and first applications, *Atmos. Environ.*, *32*(17), 2981–2999.

Adams, P. J., and J. H. Seinfeld (2002), Predicting global aerosol size distributions in general circulation models, *J. Geophys. Res.*, *107*(D19), 4370, doi:10.1029/2001JD001010.

Barth, M. C., P. J. Rasch, J. T. Kiehl, C. M. Benkovitz, and S. E. Schwartz (2000), Sulfur chemistry in the National Center for Atmospheric Research Community Climate Model: Description, evaluation, features, and sensitivity to aqueous chemistry, *J. Geophys. Res.*, *105*(D1), 1387–1415.

Binkowski, F. S., and S. J. Roselle (2003), Models-3 Community Multiscale Air Quality (CMAQ) model aerosol component: 1. Model description, *J. Geophys. Res.*, *108*(D6), 4183, doi:10.1029/2001JD001409.

Brock, C. A., et al. (2004), Particle characteristics following cloud-modified transport from Asia to North America, *J. Geophys. Res.*, *109*, D23S26, doi:10.1029/2003JD004198.

Capaldo, K. P., C. Pilinis, and S. N. Pandis (2000), A computationally efficient hybrid approach for dynamic gas/aerosol transfer in air quality models, *Atmos. Environ.*, *34*, 3617–3627.

Chin, M., R. B. Rood, S. J. Lin, J. F. Muller, and A. M. Thompson (2000), Atmospheric sulfur cycle simulated in the global model GOCART: Model description and global properties, *J. Geophys. Res.*, *105*, 24,671–24,687.

DeBell, L. J., R. W. Talbot, J. E. Dibb, J. W. Munger, E. V. Fischer, and S. E. Frolking (2004), A major regional air pollution event in the northeastern United States caused by extensive forest fires in Quebec, Canada, *J. Geophys. Res.*, *109*, D19305, doi:10.1029/2004JD004840.

de Gouw, J. A., et al. (2005), Budget of organic carbon in a polluted atmosphere: results from the New England Air Quality Study in 2002, *J. Geophys. Res.*, *110*, D16305. doi:10.1029/2004JD005623.

de Gouw, J. A., C. Warneke, A. Stohl, A. G. Wollny, C. A. Brock, O. R. Cooper, J. S. Holloway, M. Trainer, and F. C. Fehsenfeld (2006), VOC compounds composition of merged and aged forest fire plumes from Alaska and western Canada, *J. Geophys. Res.*, *111*, D10303, doi:10.1029/2005JD006175.

Eatough, D. J., R. W. Long, W. K. Modey, and N. L. Eatough (2003), Semivolatile secondary organic aerosol in ruban atmospheres: Meeting a measurement challenge, *Atmos. Environ.*, *37*, 1277–1292.

Fast, J. D., W. I. Gustafson Jr., R. C. Easter, R. A. Zaveri, J. C. Barnard, E. G. Chapman, G. A. Grell, and S. E. Peckham (2006), Evolution of ozone, particulates, and aerosol direct radiative forcing in the vicinity of Houston using a fully coupled meteorology-chemistry-aerosol model, *J. Geophys. Res.*, *111*, D21305, doi:10.1029/2005JD006721.

Gong, S. L., et al. (2003), Canadian Aerosol Module: A size-segregated simulation of atmospheric aerosol processes for climate and air quality models: 1. Module development, *J. Geophys. Res.*, *108*(D1), 4007, doi:10.1029/2001JD002002.

Gong, W. (2002), Modelling cloud chemistry in a regional aerosol model: Bulk vs. size resolved representation, in *Air Pollution Modeling and Its Applications*, vol. 15, edited by C. Borrego and G. Schayes, pp. 285–293, Springer, New York.

Grell, G. A., S. E. Peckham, S. A. McKeen, R. Schmitz, and G. Frost (2005), Fully coupled "online" chemistry within the WRF model, *Atmos. Environ.*, *39*, 6957–6975.

Grover, B. D., M. Kleinman, N. L. Eatough, D. J. Eatough, P. K. Hopke, R. W. Long, W. E. Wilson, M. B. Meyer, and J. L. Ambs (2005), Measurement of total PM2.5 mass (nonvolatile plus semivolatile) with the Filter Dynamic Measurement System tapered element oscillating microbalance monitor, *J. Geophys. Res.*, *110*, D07S03, doi:10.1029/2004JD004995.

Hong, S.-Y., and H.-L. Pan (1996), Nonlocal boundary layer vertical diffusion in a medium-range forecast model, *Mon. Weather Rev.*, *124*, 2322–2339.

Horowitz, L. W., et al. (2003), A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version 2, *J. Geophys. Res.*, *108*(D24), 4784, doi:10.1029/2002JD002853.

Jacobson, M. Z. (1997), Development and application of a new air pollution modeling system. II Aerosol module structure and design, *Atmos. Environ.*, *31*, 131–144.

Janjic, Z. I. (2002), Nonsingular implementation of the Mellor-Yamada level 2.5 scheme in the NCEP Meso model, *NCEP Off. Note 437*, 61 pp., Natl. Cent. for Environ. Predict., Camp Springs, Md.

Koch, D., D. Jacob, I. Tegen, D. Rind, and M. Chin (1999), Tropospheric sulfur simulation and sulfate direct radiative forcing in the Goddard

Institute for Space Studies general circulation model, *J. Geophys. Res.*, *104*(D19), 23,799–23,822.

Koo, B., T. M. Gaydos, and S. N. Pandis (2003), Evaluation of the equilibrium, dynamic, and hybrid aerosol modeling approaches, *Aerosol Sci. Technol.*, *37*, 53–64.

Lu, R., R. P. Turco, and M. Z. Jacobson (1997), An integrated air pollution modeling system for urban and regional scales: 1. Structure and performance, *J. Geophys. Res.*, *102*, 6063–6079.

Lurmann, F. W., A. S. Wexler, S. N. Pandis, S. Musarra, N. Kumar, and J. H. Seinfeld (1997), Modeling urban and regional aerosols—II. Application to California's South Coast Air Basin, *Atmos. Environ.*, *31*, 2695–2715.

McKeen, S., 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.

Meng, Z. Y., D. Dabdub, and J. H. Seinfeld (1998), Size-resolved and chemically resolved model of atmospheric aerosol dynamics, *J. Geophys. Res.*, *103*, 3419–3435.

Nenes, A., C. Pilinis, and S. N. Pandis (1998), ISORROPIA: A new thermodynamic equilibrium model for multiphase multicomponent marine aerosols, *Aquat. Geochem.*, *4*, 123–152.

Orsini, D., Y. Ma, A. Sullivan, B. Sierau, K. Baumann, and R. J. Weber (2003), Refinements in the particle-into-liquid sampler (PILS) for ground and airborne measurements of water soluble aerosol chemistry, *Atmos. Environ.*, *37*, 1243–1259.

Pfister, G., P. G. Hess, L. K. Emmons, J.-F. Lamarque, C. Wiedinmyer, D. P. Edwards, G. Pétron, J. C. Gille, and G. W. Sachse (2005), Quantifying CO emissions from the 2004 Alaskan wildfires using MOPITT CO data, *Geophys. Res. Lett.*, *32*, L11809, doi:10.1029/2005GL022995.

Pilinis, C., K. P. Capaldo, A. Nenes, and S. N. Pandis (2000), MADM—A new multicomponent aerosol dynamics model, *Aerosol Sci. Technol.*, *32*, 482–502.

Pudykiewicz, J., A. Kallaur, and P. K. Smolarkiewicz (1997), Semi-Lagrangian modeling of tropospheric ozone, *Tellus, Ser. B*, *49*, 231–258.

Schwartz, S. E. (1988), Mass transport limitation to the rate of in-cloud oxidation of SO_2: Reexamination in the light of new data, *Atmos. Environ.*, *22*, 2491–2499.

Seigneur, C. (2001), Current status of air quality modeling for particulate matter, *J. Air Waste Manage. Assoc.*, *51*, 1508–1821.

Seigneur, C., and M. Moran (2004), Chemical-transport models, in *Particulate Matter Science for Policy Makers*, edited by P. McMurry, M. Shepherd, and J. Vickery, pp. 283–323, Cambridge Univ. Press, New York.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and D. Kim (2000), Parameterization of cold-season processes in the MAPS land-surface scheme, *J. Geophys. Res.*, *105*(D3), 4077–4086.

Spracklen, D. V., K. J. Pringle, K. S. Carslaw, M. P. Chipperfiled, and G. W. Mann (2005), A global off-line model of size-resolved aerosol microphysics: I. Model development and prediction of aerosol properties, *Atmos. Chem. Phys.*, *3*, 2227–2252.

Sun, Q., and A. S. Wexler (1998), Modeling urban and regional aerosols near acid neutrality—Application to the June 24–25 SCAQS episode, *Atmos. Environ.*, *32*, 3533–3545.

Tang, Y., et al. (2004), Three-dimensional simulations of inorganic aerosol distributions in east Asia during spring 2001, *J. Geophys. Res.*, *109*, D19S23, doi:10.1029/2003JD004201.

Warneke, C., et al. (2006), Biomass burning and anthropogenic sources of CO over New England in the summer 2004, *J. Geophys. Res.*, *111*, D23S15, doi:10.1029/2005JD006878.

Weber, R. J., D. A. Orsini, Y. Duan, Y.-N. Lee, P. J. Klotz, and A. Brechtel (2001), Particle-into-liquid collector for rapid measurement of aerosol bulk chemical composition, *Aerosol Sci. Technol.*, *35*, 718–727.

Wilczak, J., et al. (2006), Bias-corrected ensemble and probabilistic forecasts of surface ozone over eastern North America during the summer of 2004, *J. Geophys. Res.*, *111*, D23S28, doi:10.1029/2006JD007598.

Yu, S. C., P. S. Kasibhatla, D. L. Wright, S. E. Schwartz, R. McGraw, and A. Deng (2003), Moment-based simulation of microphysical properties of sulfate aerosols in the eastern United States: Model description, evaluation and regional analysis, *J. Geophys. Res.*, *108*(D12), 4353, doi:10.1029/2002JD002890.

Yu, S., R. L. Dennis, P. V. Bhave, and B. K. Eder (2004), Primary and secondary organic aerosols over the United States: Estimates on the basis of observed organic carbon (OC) and elemental carbon (EC), and air quality modeled primary OC/EC ratios, *Atmos. Environ.*, *38*, 5257–5268.

Yu, S. C., R. Dennis, S. Roselle, A. Nenes, J. Walker, B. Eder, K. Schere, J. Swall, and W. Robarge (2005), An assessment of the ability of three-dimensional air quality models with current thermodynamic equilibrium models to predict aerosol NO_3^-, *J. Geophys. Res.*, *110*, D07S13, doi:10.1029/2004JD004718.

Yu, S. C., P. V. Bhave, R. L. Dennis, and R. Mathur (2007), Seasonal and regional variations of primary and secondary organic aerosols over the continental United States: Semi-empirical estimates and model evaluation, *Environ. Sci. Technol*, in press.

Zhang, Y., C. Seigneur, J. H. Seinfeld, M. Z. Jacobson, and F. S. Binkowski (1999), Simulation of aerosol dynamics: A comparative review of algorithms used in air quality models, *Aerosol Sci. Technol.*, *31*, 487–514.

Zhang, Y., C. Seigneur, J. H. Seinfeld, M. Z. Jacobson, S. L. Clegg, and F. S. Binkowski (2000), A comparative review of inorganic aerosol thermodynamics equilibrium aerosol modules: Similarity, differences, and their likely causes, *Atmos. Environ.*, *34*, 117–137.

Zhang, Y., B. Pun, K. Vijayaraghavan, S. Y. Wu, C. Seigneur, S. N. Pandis, M. Z. Jacobson, A. Nenes, and J. H. Seinfeld (2004), Development and application of the model of aerosol dynamics, reaction, ionization, and dissolution (MADRID), *J. Geophys. Res.*, *109*, D01202, doi:10.1029/2003JD003501.

V. Bouchet and R. Moffet, Meteorological Service of Canada, Dorval, QC, Canada H9P 1J3.

G. R. Carmichael and Y. Tang, Center for Global and Regional Environmental Research, University of Iowa, Iowa City, IA 52242, USA.

S. H. Chung and S. McKeen, Chemical Sciences Division, Environmental Science Research Laboratory, NOAA, Boulder, CO 80305, USA. (stuart.a.mckeen@noaa.gov)

I. Djalalova and J. Wilczak, Physical Sciences Division, Environmental Science Research Laboratory, NOAA, Boulder, CO 80305, USA.

W. Gong, Meteorological Service of Canada, Downsview, ON, Canada M3H 5T4.

G. Grell and S. Peckham, Global Systems Division, Environmental Science Research Laboratory, NOAA, Boulder, CO 80305, USA.

R. Mathur and S. Yu, NREL/ASMD, U.S. EPA, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA.