Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004

S. McKeen,[1,2] J. Wilczak,[3] G. Grell,[2,4] I. Djalalova,[3,5] S. Peckham,[2,4] E.-Y. Hsie,[1,2] W. Gong,[6] V. Bouchet,[7] S. Menard,[7] R. Moffet,[7] J. McHenry,[8] J. McQueen,[9] Y. Tang,[10] G. R. Carmichael,[10] M. Pagowski,[4,11] A. Chan,[12] T. Dye,[12] G. Frost,[1,2] P. Lee,[9] and R. Mathur[13,14]

Received 8 February 2005; revised 30 May 2005; accepted 15 August 2005; published 8 November 2005.

[1] The real-time forecasts of ozone (O_3) from seven air quality forecast models (AQFMs) are statistically evaluated against observations collected during July and August of 2004 (53 days) through the Aerometric Information Retrieval Now (AIRNow) network at roughly 340 monitoring stations throughout the eastern United States and southern Canada. One of the first ever real-time ensemble O_3 forecasts, created by combining the seven separate forecasts with equal weighting, is also evaluated in terms of standard statistical measures, threshold statistics, and variance analysis. The ensemble based on the mean of the seven models and the ensemble based on the median are found to have significantly more temporal correlation to the observed daily maximum 1-hour average and maximum 8-hour average O_3 concentrations than any individual model. However, root-mean-square errors (RMSE) and skill scores show that the usefulness of the uncorrected ensembles is limited by positive O_3 biases in all of the AQFMs. The ensembles and AQFM statistical measures are reevaluated using two simple bias correction algorithms for forecasts at each monitor location: subtraction of the mean bias and a multiplicative ratio adjustment, where corrections are based on the full 53 days of available comparisons. The impact the two bias correction techniques have on RMSE, threshold statistics, and temporal variance is presented. For the threshold statistics a preferred bias correction technique is found to be model dependent and related to whether the model overpredicts or underpredicts observed temporal O_3 variance. All statistical measures of the ensemble mean forecast, and particularly the bias-corrected ensemble forecast, are found to be insensitive to the results of any particular model. The higher correlation coefficients, low RMSE, and better threshold statistics for the ensembles compared to any individual model point to their preference as a real-time O_3 forecast.

Citation: 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.

1. Introduction

[2] Real-time forecasts of surface level O_3 have been available from regional-scale air quality models (AQFMs) for several years (CHRONOS, 2001, see description below) [*Grell et al.*, 2002; *McHenry et al.*, 2004; *Vaughan et al.*, 2004; *Delle Monache et al.*, 2004]. As part of the International Consortium for Atmospheric Research on Transport and Transformation/New England Air Quality Study (ICARTT/NEAQS) field study conducted over New England during the summer of 2004, six operational and research institutions contributed their real-time forecast results to a central facility (the National Oceanic and Atmospheric Administration (NOAA) Aeronomy Laboratory). The NOAA Forecast Systems Laboratory Weather Research

[1]Aeronomy Laboratory, NOAA, Boulder, Colorado, USA.
[2]Also at Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA.
[3]Environmental Technology Laboratory, NOAA, Boulder, Colorado, USA.
[4]Forecast Systems Laboratory, NOAA, Boulder, Colorado, USA.
[5]Science and Technology Corporation, Hampton, Virginia, USA.
[6]Meteorological Service of Canada, Downsview, Ontario, Canada.
[7]Meteorological Service of Canada, Dorval, Quebec, Canada.
[8]Baron Advanced Meteorological Systems, Raleigh, North Carolina, USA.
[9]NOAA/National Weather Service National Centers for Environmental Prediction Environmental Modeling Center, Camp Springs, Maryland, USA.
[10]Center for Global and Regional Environmental Research, University of Iowa, Iowa City, Iowa, USA.
[11]Also at Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, USA.
[12]Sonoma Technology, Inc., Petaluma, California, USA.
[13]Air Resources Laboratory, NOAA, Silver Spring, Maryland, USA.
[14]On assignment to National Exposure Research Laboratory, Environmental Protection Agency, Research Triangle Park, North Carolina, USA.

0148-0227/05/2005JD005858$09.00

and Forecast model/Chemistry version (WRF/Chem) model provided three independent forecasts, though only one forecast is considered in this study. The Meteorological Services of Canada provided results from both their Canadian Hemispheric and Regional Ozone and NO_x System (CHRONOS) and A Unified Regional Air-Quality Modeling System (AURAMS) models. Also included are O_3 forecasts from the National Weather Service/National Centers for Environmental Prediction (NWS/NCEP) Community Multi-scale Air Quality Model/Eta (CMAQ/Eta) model, which are publicly available through the NWS/NCEP Internet based data archive facility. The Baron Advanced Meteorological System, Inc. (Baron AMS), corporation provided four AQ forecasts at three horizontal grid lengths (45, 15, and 5 km) with results from the 45 and 15 km models considered here. Forecasts from the University of Iowa 12 km horizontal grid spaced Sulfur Transport and Emissions Model–2003 (STEM-2K3) AQ model complete the ensemble of seven models that are utilized in this study.

[3] As part of a collective, informal model verification project within the ICARTT/NEAQS-2004 study, forecasts of several key meteorological, radiation, and gas-phase atmospheric constituents were gathered in near real time (typically 4- to 10-hour computational delay) from the nine AQFMs, and graphically compared with corresponding real-time measurements from various surface, upper-air, and ship-based platforms associated with ICARTT/NEAQS. The day-to-day time series of these comparisons, as collected in real time, can be viewed at the NOAA Environmental Technology Laboratory (ETL) Internet Web address http://www.etl.noaa.gov/programs/2004/neaqs/verification/. Although the focus of the NEAQS-2004 program and its model verification study is the northeast U.S. urban corridor, the majority of AQ forecasts predict surface O_3 concentrations over a much larger region of North America. The AIRNow O_3 monitoring network (http://www.epa.gov/airnow) provides real-time hourly average O_3 measurements at hundreds of stations in the eastern United States and Canada, allowing excellent spatial coverage for model O_3 forecast evaluations. The work presented here is limited to comparisons of O_3 from the real-time data collected through the AIRNow network. It should be noted that official, published 2004 AIRNow measurements might differ at particular locations from the real-time values used here through quality assurance/control (QA/QC) processing. Previous experience, however, based on real-time and final AIRNow data from 2002 suggest the influence of AIRNow QA/QC on the statistics of large samples of data is negligible.

[4] The cooperative efforts of the various forecast groups also allowed a unique opportunity to present one of the first ever real-time ensemble forecasts, as well as the bias-corrected ensemble forecasts, of surface O_3 at 13 locations in the northeast United States and southern Canada. These ensemble forecasts were also posted and can be viewed at the above-mentioned NOAA/ETL web address. Ensemble techniques are commonly used to improve the forecast ability of weather models [e.g., *Kalnay*, 2003], and have successfully been applied to dispersion forecasts of radionuclides and inert tracers [*Galmarini et al.*, 2004, and references therein]. The application of ensemble techniques to air quality forecasts is very recent, and usually available only in a retrospective sense for a few sites [*Delle Monache and Stull*, 2003]. The multimodel real-time O_3 forecasts presented here, and the ensemble O_3 forecasts of *Delle Monache et al.* [2004] for the summer of 2004 over the northwest United States and southwest British Columbia are probably the first ever documented attempts at real-time ensemble O_3 forecasts. The major intent of this study is to critically examine the veracity and usefulness of the ensemble forecast relative to its individual members, and to provide a reference and some guidance for future real-time AQ ensemble forecasts.

[5] A couple caveats accompany the analysis presented in this paper. First, model results are based solely on the forecasts collected in real time during the summer of 2004. All of the AQFMs have undergone modifications in numerical formulation or emissions that supersede the models used in this analysis. The model evaluations therefore represent a snapshot in time of the rapidly evolving field of air quality forecasting, and the relative performance of current modeling systems cannot be inferred from the results presented here. Second, the summer of 2004 exhibited very few occurrences of pollution episodes because of unseasonably cool weather associated with continental polar air masses during July and the influence of several hurricanes during August. An analysis of surface O_3 along the eastern U.S. coastline during the summer of 2002 shows more than a factor 10 increase in the 85 ppbv average 8-hour maximum, and 125 ppbv 1-hour maximum threshold exceedances compared to the summer of 2004. The model statistics presented here are therefore specific to the summer of 2004 and probably not climatologically representative.

2. Air Quality Forecast Models (AQFMs)

[6] Of the nine AQFMs displayed in real time, eight models provided nearly continuous forecasts of the eastern United States and southeastern Canada between 6 July and 30 August 2004. The domains of these models are shown in Figure 1a. The region of overlap for all of the models shown in Figure 1b includes the combination of the two Baron Advanced Meteorological System 15 km grid spaced models (hereafter referred to as BAMS-15 km), which accounts for the seven models referred within this work. The region of model overlap is determined by the STEM-12 km and BAMS-15 km model domain limits. The locations of the 358 AIRNow monitors used in the analysis are also shown in Figure 1b.

[7] It is beyond the scope of this work to describe the numerical details of all the AQFMs used here. Instead, a brief description of each AQFM with journal and Web-based references is provided for additional information. Table 1 summarizes some basic features (grid spacing, chemical mechanism, base year of the anthropogenic emissions inventory), and the Web address corresponding to each AQFM that links to further references, real-time forecast products, or ICARTT applications. The following AQFM descriptions provide a more detailed account of the basic meteorological framework, photochemical mechanism, and emissions of anthropogenic and biogenic O_3

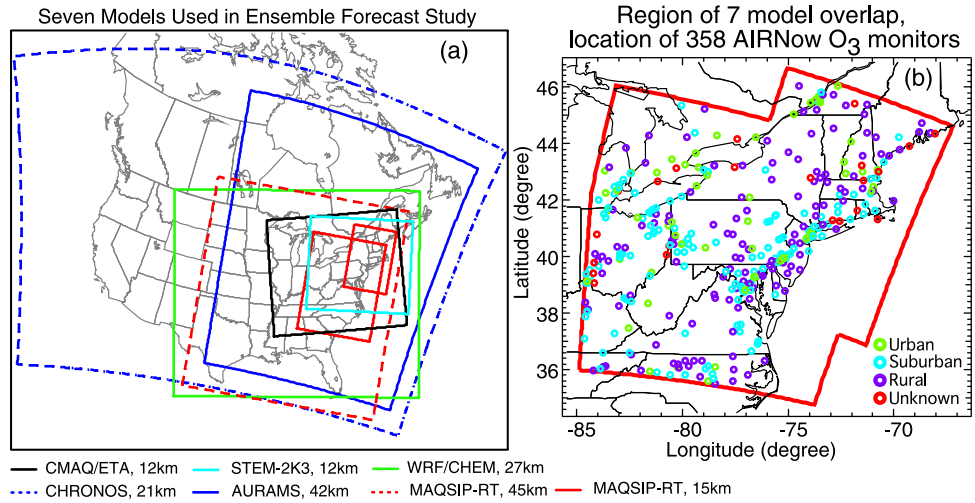


Figure 1. (a) Continental map of North America showing the model boundaries of the eight models used in the analysis and ensemble average. The abbreviation assigned to each AQFM on the abscissa is given in Table 1. (b) Domain of model overlap in Figure 1a and the location of 358 AIRNow monitors that collected real-time O_3 measurements during the summer of 2004.

precursors in order to highlight the model differences for these important O_3 forcing factors.

2.1. NOAA/FSL WRF/Chem

[8] The Weather Research and Forecasting (WRF) Chemical model is based upon the nonhydrostatic WRF community model developed at National Center for Atmospheric Research (NCAR; http://www.mmm.ucar.edu/wrf/users). Details of WRF/Chem are given by *Grell et al.* [2005]. This model is an extension of the earlier MM5/Chem [*Grell et al.*, 2000] regional-scale chemical transport model to the WRF architecture, and real-time forecasts can be found at Internet Web address http://www-frd.fsl.noaa.gov/aq/wrf. This model system is "online" in the sense that all processes affecting the gas phase and aerosol species are calculated in lock step with the meteorological dynamics. Meteorological initial conditions are taken from the Rapid Update Cycle (RUC) model analysis fields generated at NOAA/FSL, and lateral boundary conditions from the NCEP Eta model forecast. Hourly output from WRF/Chem 36-hour forecasts are started at 0000 and 1200 UT of each day using a horizontal grid spacing of 27 km. Gas-phase chemistry is based upon the RADM2 mechanism of *Stockwell et al.* [1990] with updates to the original mechanism [*Stockwell et al.*, 1995]. Lateral boundary conditions for O_3 and its precursors are prescribed identically to those given by *McKeen et al.* [2002] and are based on averages of midlatitude aircraft profiles from several field studies over the eastern Pacific.

[9] Anthropogenic emissions are treated similar to that of *McKeen et al.* [2002] with updates to the April 2002 release of the EPA NET-96 inventory version 3.12 [*U.S. Environmental Protection Agency*, 1998]. Hourly temporal allocation, VOC speciation, and spatial partitioning within a specified county are based on the older, yet detailed information within the National Acid Precipitation Assessment Program (NAPAP) emissions database [*U.S. Environmental Protection Agency*, 1989]. Canadian emissions are also taken directly from the NAPAP modeler's inventory. In order to adhere to the RADM2 mechanism, reactivity weighting of the various NAPAP-lumped VOC species is used to derive the emissions of the RADM2-lumped species according to *Stockwell et al.* [1990, 1997].

Table 1. Air Quality Forecast Model Name (Horizontal Grid Spacing), Abbreviated Mnemonic, Abbreviated Chemical Mechanism, Base Year of Anthropogenic Emissions Inventory, and Internet Web Addresses of the Seven Models[a]

AQ Forecast Model	Mnemonic	Chemical Mechanism	Base Year Emissions	Internet Web Address
AURAMS (42 km)	AUR	ADOM-II	1995, 1996	http://www.msc-smc.ec.gc.ca/research/icartt/aurams_e.html
BAMS (15 km)	BAM 15	CBM-4	2001, 2004	http://www.baronams.com/projects/SECMEP/index.html
BAMS (45 km)	BAM 45	CBM-4	2001, 2004	http://www.baronams.com/projects/SECMEP/index.html
CHRONOS (21 km)	CHR	ADOM-II	1995, 1996	http://www.msc-smc.ec.gc.ca/aq_smog/chronos_e.cfm
CMAQ/Eta (12 km)	CMQ	CBM-4	2001, 2004	http://wwwt.emc.ncep.noaa.gov/mmb/aq/
STEM-2K3 (12 km)	STM	SAPRC-99	1999	http://nas.cgrer.uiowa.edu/ICARTT/icartt-2k4.html
WRF/Chem (27 km)	WRF	RADM2	1996	http://www.wrf-model.org/WG11
Ensemble mean[b]	ENS mean			http://www.etl.noaa.gov/programs/2004/neaqs/verification/
Ensemble median	ENS median			this report

[a]Abbreviated mnemonic denotes individual models in figures. See text for details of the chemical mechanisms and anthropogenic emissions. The Web addresses are for further information or details regarding each AQFM.

[b]Ensemble results at 13 surface sites calculated in real time during the summer of 2004 are contained within this comprehensive ICARTT evaluation Web site.

[10] Biogenic emissions of isoprene, monoterpenes, other VOC (OVOC), and nitrogen emissions by the soil are specified at reference temperature and photosynthetic active radiation (PAR) according to *Guenther et al.* [1994] for deciduous, coniferous and mixed forest, and from *Schoenemeyer et al.* [1997] for agricultural and grassland. Temperature and light dependence of isoprene emissions are taken from *Guenther et al.* [1995], while the temperature dependence of monoterpenes, soil NO and OVOC are that of *Simpson et al.* [1995]. Emissions are applied over a surface grid according to the single WRF land use category assigned to that grid, and the temperature dependence of the emissions is tied to the surface temperature. Similar to the anthropogenic sources, the emissions of monoterpenes and OVOC are disaggregated into the RADM2 species classes.

2.2. Meteorological Service of Canada CHRONOS and AURAMS Models

[11] Canadian Hemispheric and Regional Ozone and NO_x System (CHRONOS) is the original Canadian national AQFM designed for O_3 forecasts that has been operational in real time since 2001 and is based upon the chemical transport model of *Pudykiewicz et al.* [1997]. Real-time forecasts and limited information about this model can be found at Internet address http://www.msc-smc.ec.gc.ca/aq_smog/chronos_e.cfm. CHRONOS is an offline model, which is currently driven by meteorological fields (at hourly intervals) calculated by the regional version of the Global Environmental Model (GEM), the operational weather prediction model of the Meteorological Service of Canada [*Côté et al.*, 1998a, 1998b]. Hourly output is obtained from 48-hour forecasts at a horizontal grid spacing of 21 km started at 0000 and 1200 UT of each day. CHRONOS uses the Acid Deposition and Oxidant Model-2 (ADOM-II) chemical mechanism, which is based upon the lumped approach of *Lurmann et al.* [1986] with updates to kinetic rates and reaction pathways from *Atkinson et al.* [1992].

[12] Anthropogenic emissions are prepared with the Canadian Emission Processing System (CEPS) [*Moran et al.*, 1997] for spatial and temporal allocations and chemical speciation. The emission fields are based on the 1990 Canadian and U.S. national criteria air contaminant inventories projected to 1995 and 1996 levels, respectively. Biogenic emissions are processed in line and based on the Biogenic Emissions Inventory System-2 (BEIS2) and the Biogenic Emissions Land-use Database (BELD), version 3, surface vegetation characterization described by *Pierce et al.* [1998].

[13] The AURAMS model is similar to CHRONOS in that it was built upon the CHRONOS AQFM. It was designed as an episodic, regional particulate matter modeling system. AURAMS employs the same ADOM II gas-phase chemical mechanism but, in addition, has a full, size-resolved and chemically resolved, representation of aerosol microphysics and gas-aerosol interaction processes [*Gong et al.*, 2003]. AURAMS is also an off-line model and is driven by the meteorological fields (at 15-minute intervals) from the limited-area version of GEM, which is in turn driven by the regional GEM forecast. One 48-hour forecast from AURAMS starts at 0000 UT of each day using a horizontal grid spacing of 42 km. The anthropogenic emissions of gaseous precursors are identical to that of CHRONOS. The biogenic emission processing in AURAMS currently uses the BEIS2 emission assignments between vegetative categories and specific biogenic VOC, but unlike CHRONOS the surface vegetation types are those of BEIS1 [e.g., *Lamb et al.*, 1993]. Primary PM emissions are also included in AURAMS on the basis of the bulk primary PM emissions included in the 1990 Canadian and U.S. national inventories, chemically and size segregated within AURAMS according to the primary source stream.

2.3. Baron AMS MAQSIP-RT Models

[14] Multiscale Air Quality Simulation Platform–Real-Time (MAQSIP-RT) [*McHenry et al.*, 2004] is an off-line chemical transport model that been applied for real-time ozone forecasting since 1998 [*McHenry et al.*, 1999]. MAQSIP-RT relies upon the MM5 (version 3.6) mesoscale model [*Grell et al.*, 1994] for meteorological information. MM5 forecasts of near-surface meteorology are also used to compute all emissions components that are meteorologically modulated within the Sparse Matrix Operator Kernel Emissions (SMOKE), version 1.4-RT, emissions processing/modeling system [*Coats*, 1996]. These include all biogenic emissions, point source plume rise, and all mobile source emissions, and evaporative VOC emissions from stationary (fuel storage tank) sources. During summer 2004, twice-daily forecasts were provided for the two 15-km domains and the 45-km domain shown in Figure 1a. The MM5 meteorological forecasts include customized physics for improved boundary layer, land surface, and explicit moisture treatments. MM5 initial and boundary conditions are derived from either NCEP's operational Eta or GFS models depending upon circumstance; initial and boundary conditions were switched from Eta to GFS on 19 August 2004 to account for more accurate tropical cyclone initialization in the GFS analysis fields. MAQSIP-RT is configured with a modified Carbon Bond 4 (CBM-4) chemistry mechanism [*Gery et al.*, 1989], which includes updated kinetic data for the CO + OH reaction [*DeMore et al.*, 1994], PAN chemistry [*Chang et al.*, 1996], an updated condensed isoprene chemical mechanism based on *Carter* [1996], and modifications to the chemical pathways of the universal peroxy radical operators (XO_2 and XO_2N) used in CBM-4, as discussed by *Kasibhatla et al.* [1997]. Upgrades to the actinic flux cloud interaction submodel [*McHenry and Coats*, 2003] are also incorporated in the forecasts.

[15] In preparation for summer 2004 significant effort was undertaken to update the foundational emissions inventories and data sets underlying the SMOKE V1.4-RT system. First, migration to the EPA NEI Version 3 (2001) point, area, and nonroad anthropogenic emission inventories was completed (U.S. EPA, Interstate air quality rule: Notice of data availability, 40 CFR Parts 51, 7, 73, 74, 77, 78, and 96, CAIR docket OAR-2003-0053, 2004, available at http://www.epa.gov/oar/interstateairquality/rule.html). This was augmented by application of major NO_x point source reduction factors based on U.S. Department of Energy (DOE) fuel use data trends for electric generating units obtained from EPA [*Pouliot*, 2005] (available at http://ams.confex.com/ams/Annual2005/

techprogram/programexpanded_257.htm), and, on a statewide basis, project the 2001 major point source NO_x emissions to the year 2004. Next, updates to the temporal and VOC speciation profiles and cross-references tables were implemented (U.S. EPA, EPA Clearinghouse for Inventories and Emissions Factors: Temporal allocation, 2004, available at http://www.epa.gov/ttn/chief/emch/temporal/index.html; U.S. EPA, EPA Clearinghouse for Inventories and Emissions Factors: Speciation, 2004, available at http://www.epa.gov/ttn/chief/emch/speciation/index.html). An updated Canadian point source inventory was also implemented to replace the year 1990 inventory used in previous forecast summers (J. Vukovich, personal communication, 2004). Additional changes to the area and nonroad source categories were implemented by utilizing updated spatial surrogates (U.S. EPA, EPA Clearinghouse for Inventories and Emissions Factors: Related spatial allocation files: "New" surrogates, 2004, available at http://www.epa.gov/ttn/chief/emch/spatial/newsurrogate.html), replacing the surrogate database derived from 1990 population census data. For on-road (mobile) emissions, SMOKEv1.4-RT includes the MOBILE5b (U.S. EPA, EPA Clearinghouse for Inventories and Emissions Factors: Emissions models, 2004, available at http://www.epa.gov/ttn/chief/emch/models/index.html) modeling system for estimating motor vehicle emissions, which was implemented with updated year 2002 vehicle miles traveled (VMT) and VMT-mix for seven Highway Performance Monitoring System roadway classes (interstate freeway, urban freeway, principal arterial, minor arterial, major collector, minor collector, and locals). Mobile source modeling also incorporates the ancillary data updates (spatial, speciation and temporal profiles) mentioned above, resulting in improved spatial allocation of mobile sources (e.g., location of recently constructed segments of major interstates and arteries). Biogenic emissions modeling utilized the Biogenic Emissions Inventory System (BEIS) version 3.9 [*Vukovich and Pierce*, 2002] with land use obtained from the Biogenic Emissions Land-use Database version 3 (BELD3) [*Pierce et al.*, 1998].

2.4. University of Iowa STEM-2K3 Model

[16] The University of Iowa Sulfur Transport and Deposition Model (STEM model) was initially developed for simulating sulfur dioxide (SO_2) transport and transformation [*Carmichael et al.*, 1986], and more recently adapted into a general regional air quality model [*Tang et al.*, 2003; *Carmichael et al.*, 2003]. STEM-2K3 is the latest version of this model, which employs the SAPRC-99 gas-phase mechanism [*Carter*, 2000], detailed treatment of aerosol thermodynamics and growth [*Tang et al.*, 2004] and an online treatment of photolysis rates and radiation that explicitly accounts for the influence of aerosols and clouds [*Tang et al.*, 2003]. During the ICARTT experiment, STEM-2K3 provided a multiscale forecast, including a primary domain with 60 km horizontal grid spacing covering the continental United States, southern Canada, and northern Mexico, the nested 12 km domain covering the northeast United States shown in Figure 1a, and an additional 4 km model domain for the New England region. Daily 48-hour forecasts from these three models for the summer of 2004 can be found at the Web address listed in Table 1. Off-line meteorological fields (1-hour resolution) from the Penn State/NCAR MM5 meteorological model are calculated using one-way nesting from larger to smaller domains for meteorological, photochemical and aerosol variables. The time-varying lateral and top boundary conditions of the 60 km grid spaced domain were provided by forecasts of the MOZART-2 global chemical transport model [*Horowitz et al.*, 2003].

[17] The STEM-2K3 uses anthropogenic emissions based on the EPA NEI-99 (version 3) inventory (U.S. EPA, EPA Clearinghouse for Inventories and Emissions Factors: 1999 National Emission Inventory documentation and data: Final version 3.0, 2004, http://www.epa.gov/ttn/chief/net/1999inventory.html) compiled at the NOAA Aeronomy Laboratory. If available within the inventory, ozone-season-day emissions are used, otherwise annual averages are assumed. These emissions are publicly available, and viewable through a graphics information system interface at http://map.ngdc.noaa.gov/website/al/emissions/viewer.htm. Spatial partitioning within U.S. counties and Canadian provinces (4 km resolution) is based on spatial surrogates and source classification code (SCC) assignments recommended by the U.S. EPA (EPA Clearinghouse for Inventories and Emissions Factors: Related spatial allocation files: "New" surrogates, 2004, available at http://www.epa.gov/ttn/chief/emch/spatial/newsurrogate.html). Daily temporal allocation also uses EPA SCC classifications, and is based upon summertime weekday profiles (U.S. EPA, EPA Clearinghouse for Inventories and Emissions Factors: Temporal allocation, 2004, available at http://www.epa.gov/ttn/chief/emch/temporal/index.html). Anthropogenic VOC speciation is taken from SCC based total organic carbon to total VOC partitioning (U.S. EPA, EPA Clearinghouse for Inventories and Emissions Factors: Speciation, 2004, available at http://www.epa.gov/ttn/chief/emch/speciation/index.html) followed by partitioning into SAPRC-99 VOC species assignments (http://pah.cert.ucr.edu/~carter/emitdb/). Canadian mobile and area sources (2000 base year) are also taken from EPA recommendations (U.S. EPA, EPA Clearinghouse for Inventories and Emissions Factors: North American emissions inventories: Canada, 2004, http://www.epa.gov/ttn/chief/net/canada.html#data), but Canadian point emissions are not included because of nondisclosure policies of private Canadian corporations. Biogenic emissions of isoprene, monoterpenes and other VOC are taken from the IGAC-GEIA archive [*Guenther et al.*, 1995] for July 1990 average emissions. The 0.5° latitude $\times$0.5° longitude inventory values are interpolated onto the various model grids, and daily invariant, diurnal profiles are assigned to each biogenic VOC.

2.5. NWS/NCEP CMAQ/Eta Model

[18] Surface O_3 forecasts from the CMAQ/Eta model are based on the off-line photochemical transport CMAQ model [*Byun and Ching*, 1999], meteorological fields derived from the NWS/NCEP Eta forecasts, and emissions processing also based on Eta forecast meteorological fields [*McQueen et al.*, 2004]. An interface component, PREMAQ, that facilitates the transformation of Eta-derived meteorological fields to conform with the CMAQ grid structure, coordinate system, and input data format has been developed. Since

both the Eta and CMAQ models use significantly different coordinate systems and grid structures, the interface component has been carefully designed to minimize effects associated with horizontal and vertical interpolation of dynamical fields in this initial implementation. Details on the methods employed and impacts of assumptions invoked are given by *Otte et al.* [2005]. For the 2004 ICARTT/NEAQS study, two 48-hour forecasts (beginning at 0600 and 1200 UT) from the 12 km resolution experimental domain, designed for the northeast United States, are used in the comparisons. The CMAQ model and its emissions processing are coupled to 12-km horizontal grid spaced Eta forecast data fields at hourly intervals, and a 6-hour Eta initialization cycle. Lateral boundary conditions for O_3 are based on an ozonesonde climatology for altitudes below 400 hPa, and NCEP's Global Forecast System (GFS) forecast O_3 fields for altitudes between 100 and 400 hPa. The CBM-4 mechanism is used in the photochemical calculations, which includes several improvements and additions to the original *Gery et al.* [1989] formulation mentioned previously and detailed by *Byun and Ching* [1999]. Additional aspects of the CMAQ model configuration used in the 2004 forecast applications are summarized by *Mathur et al.* [2004] (available at http://www.cmascenter.org/html/2004_workshop/abstracts_presentations.html).

[19] The emission processing and incorporation of both anthropogenic and biogenic emissions into the CMAQ model is based on the Sparse Matrix Operator Kernel Emissions (SMOKE) system, which is also used with the Baron AMS MAQSIP-RT model described above. The primary differences are that the mobile emissions are based on the MOBILE6 emissions model (U.S. EPA, EPA Clearinghouse for Inventories and Emissions Factors: Emissions models, 2004, available at http://www.epa.gov/ttn/chief/emch/models/index.html) and that the meteorologically dependent components of the SMOKE emissions modeling system have been incorporated into the single interface program PREMAQ, and use the Eta forecast fields to determine the meteorologically modulated emissions. The emission inventories used by the CMAQ-Eta system were updated to represent the 2004 forecast period. NO_x emissions from point sources were projected to 2004 (relative to a 2001 base inventory) using estimates derived from the annual energy outlook by the Department of Energy (http://www.eia.doe.gov/oiaf/aeo/index.html). Since MOBILE6 is computationally expensive and inefficient for real-time applications, mobile source emissions were estimated using approximations to the MOBILE6 model. In this approach MOBILE6 was used to create retrospective emissions over an eight-week period over the air quality forecast grid using the 1999 Vehicle Miles VMT data and 2004 vehicle fleet information. Least squares regressions relating the emissions to variations in temperature were then developed for each grid cell at each hour of the week and for each emitted species [*Pouliot*, 2005]. Consequently, mobile emissions could then be readily estimated in the forecast system using the temperature fields from the Eta model. Area source emissions were based on the 2001 National Emissions Inventory, version 3, while BEIS3.12 [*Pierce et al.*, 2002] was used to estimate the biogenic emissions. Additional details on the emissions processing system for the CMAQ/Eta model and evaluation of the mobile and point source emission estimates developed for 2004 are given by *Pouliot* [2005].

3. Observations and Details of Analysis

[20] Real-time, hourly updated O_3 data were provided by Sonoma Technology, Inc., to NOAA/FSL for display within the NOAA FX-NET weather information network. The location of the 358 stations within the domain of model overlap is shown in Figure 1b along with some information on surrounding population. Although hourly AIRNow data are available, the statistical analyses presented here are based on the daily maximum 1-hour O_3 levels and daily maximum 8-hour O_3 levels. The reported values from Sonoma Technology are used for these quantities, rather than recalculating maximums from the hourly data. The AIRNow procedure for accepting a 1-hour or 8-hour maximum concentration is quite rigorous. If more than 2 hourly averages are missing within a 24-hour period a missing value for that day is reported. Sixteen stations within Figure 1b failed to have more than 30 daily maximum values available for the sampling period considered, and are eliminated from the statistical comparisons, leaving 342 stations within the sample region.

[21] The 56-day period between 0000 UT, 6 July 2004, and 0000 UT, 30 August 2004, is the sampling period used in this analysis. The statistical evaluation is only for results from the first 28 hours of the 0000 UT forecasts. Daily values of maximum 8-hour average and maximum 1-hour average O_3 are calculated from the 0000 UT forecasts between 0400 UT and 28 hours into the 0000 UT forecast to match the definition of daily maximums used by Sonoma Tech. In the case of CMAQ/Eta (no 0000 UT forecast, instead starting at 0600 UT), the maximums are determined over the 22-hour period from 0600 to 0400 UT the next day. Three days are removed from the analysis because of one or more missing 0000 UT forecasts (1, 5, and 9 August) in one of the seven models, leaving 53 days with coincident and ensemble results.

[22] Latitudes and longitudes of 342 AIRNow O_3 monitoring stations falling within the domain of model overlap are mapped into the grid coordinates of each model. Comparisons with observations assume that model values are uniform over a model grid, and observed O_3 values are compared with model grid values that contain each monitor. Thus no spatial interpolation is performed on the models, but depending on model resolution, O_3 observations from several monitors could be evaluated against results from only one model grid. Results from the two 15-km grid spaced BAMS models are combined into one model. Because the larger domain provides more spatial coverage, the BAMS-15 km model used in this study takes results of the mid-Atlantic, larger domain model when and where data from that model is available; otherwise the results of the smaller, northeast U.S. domain are used.

[23] 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

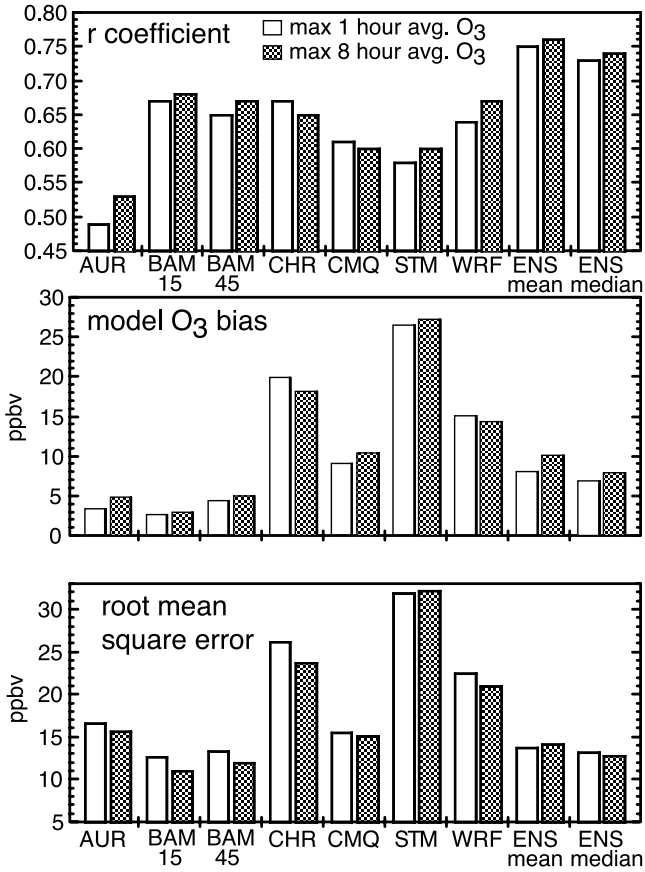


Figure 2. Median values of the (top) *r* correlation coefficient, (middle) mean bias in ppbv, and (bottom) root-mean-square error in ppbv for the seven AQFMs, the ensemble mean, and the ensemble median. Values for each model are derived from comparisons at the 342 monitor locations having 30 or more days of available data.

already averaged over these hourly periods. The WRF, STEM, 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. The BAMS results come as half-hourly snapshots, so hourly averages based on area weight (1/4, 1/2, 1/4) centered on the half hour are calculated in this case.

[24] Two ensemble model data sets are constructed from the results of the seven AQFMs on this common time base for each monitor location. The ensemble mean O_3 is calculated by taking the linear average of all seven model hourly O_3 forecasts, and this ensemble mean is the focus of much of the following analysis. The second ensemble is determined from the median value of the seven AQFM results for each monitor location and at each hour. The ensemble median should filter model forecasts that are significant outliers, and is often used in radionuclide and inert tracer dispersion studies [e.g., *Galmarini et al.*, 2004]. The benefits of ensemble median relative to the ensemble mean for air quality forecasts are a point of interest, and are therefore examined within the context of this ICARTT/NEAQS-2K4 evaluation study. Alternative ensemble formulations are possible, such as the approach presented by *Pagowski et al.* [2005], which uses the same model forecasts and AIRNow data set reported here. In that study the ensemble is constructed by a weighted sum of the seven AQFMs, where weights to each model are determined by a least squares minimization of ensemble error for a highly overdetermined set of linear equations.

4. Standard Statistics for the Models and Ensemble

[25] Standard statistical measures [see, e.g., *Tilmes et al.*, 2002] representing overall median conditions for the domain of model overlap are given in Figure 2, which shows the median values of the correlation coefficient

$$\mathrm{r(i)} = \frac{\sum_{days}\left(O_3^{modl}(i,day) - O_3^{modl}(i,avg)\right)\left(O_3^{obs}(i,day) - O_3^{obs}(i,avg)\right)}{\sqrt{\sum_{days}\left(O_3^{modl}(i,day) - O_3^{modl}(i,avg)\right)^2 \sum_{days}\left(O_3^{obs}(i,day) - O_3^{obs}(i,avg)\right)^2}} \quad (1)$$

the mean bias

$$\text{Mean Bias(i)} = \left(\frac{1.}{\mathrm{N}_{days}}\right)\sum_{days}\left[O_3^{modl}(i,day) - O_3^{obs}(i,day)\right] \quad (2)$$

and the root-mean-square error

$$\text{RMSE(i)} = \sqrt{\left(\frac{1.}{\mathrm{N}_{days}}\right)\sum_{days}\left(O_3^{modl}(i,day) - O_3^{obs}(i,day)\right)^2} \quad (3)$$

where i refers to O_3 monitor i (i = 1 to 342), N_{days} refers to number of observing days at each site, "obs" refers to observed, and "modl" refers to model values. These three quantities are shown for the seven AQFMs, the ensemble mean and ensemble median for both the maximum 1-hour average O_3 and the maximum 8-hour average O_3. First, the statistical measures between the 1-hour average and 8-hour average O_3 data are relatively similar from model to model. Further analysis is therefore restricted to the set of maximum 8-hour average O_3 data. The correlation coefficient for the ensemble mean (0.76) and ensemble median (0.74) are significantly larger than that of the nearest individual model (0.68). The square of the correlation coefficient is often considered a measure of the observed temporal variance that is described by a model. Under this definition the ensemble forecasts can explain more than half the temporal variance of the observations at more than half of the monitors, which cannot be said of any individual AQFM. However, the ensemble mean has a 10 ppbv bias, which is representative of the average of the median biases of the 7 individual models. Three models show relatively large model bias (median > 15 ppbv), while 3 models show much smaller positive bias (median < 5 ppbv). The ensemble median shows somewhat reduced biases (median 8 ppbv) compared to the ensemble mean. Finally, the median root-mean-square errors are directly proportional to the median biases, showing the expected importance of

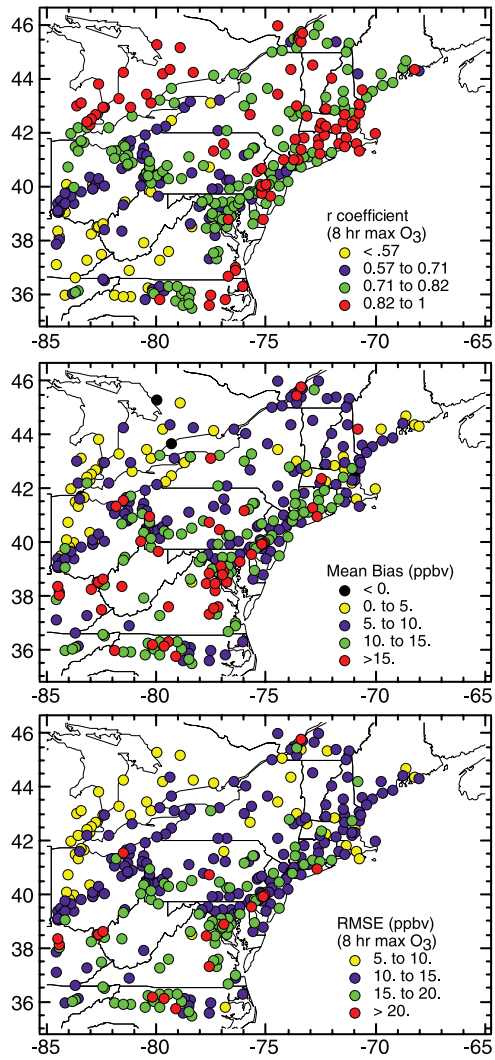


Figure 3. Spatial distribution of the (top) *r* correlation coefficient, (middle) mean bias in ppbv, and (bottom) root-mean-square error in ppbv for the seven-model ensemble mean.

positive bias errors in determining this measure of model agreement.

[26] Data from the aggregate of all 53 days and 342 monitor comparisons are used in a simple analysis of the confidence intervals and significance tests of the maximum 8-hour average O_3 ensemble *r* coefficient. For the aggregate statistics (16,487 comparison points) the *r* coefficient is 0.733 for the ensemble mean, 0.719 for the ensemble median, and 0.670 for the next highest model (BAMS-15 km). Standard statistical evaluations of the 99.9% confidence intervals [e.g., *Johnson and Leone*, 1977] for these three *r* coefficients yields a range between 0.721 and 0.745 for the ensemble mean *r* coefficient, 0.706 and 0.731 for the ensemble median, and 0.656 to 0.684 for the BAMS-15 km model. The improvement in *r* coefficients for both the ensemble mean and median are therefore highly significant for the aggregate of all possible comparisons, and the correlation for the ensemble mean is also statistically higher than for the ensemble median at the 99.9% confidence level. The limitation of the ensemble results as a forecast tool relative to any other model is due to the high bias and resulting high RMSE values.

[27] Though the ensembles show statistical improvement in temporal correlation, the ensemble is only a "quick fix" to providing a better forecast. Identifying and eliminating sources of model O_3 biases and errors in each individual AQFM is the subject of ongoing research, requiring improved treatments of emission estimates and various meteorological and chemical processes. No attempt is made here to diagnose or speculate on possible causes of bias or other imperfections within each of the individual models. The ensemble, on the other hand, represents a collective model perspective in which errors and biases result from a collective misrepresentation of the system. The spatial distribution of *r* correlation coefficients, mean bias and RMSE that comprise the ensemble mean results in Figure 2 are shown in Figure 3. For the *r* correlation coefficients a distinct spatial pattern emerges; a region that straddles the Ohio River Valley has lower *r* correlation coefficients than regions north of Lake Erie or along the U.S. coastline. Mean biases show a similar pattern with higher biases associated with proximity to the Ohio River Valley and lower biases to the north and northwest. This northwest to southeast gradient in O_3 bias also correlates with gradients in agricultural versus forested land use patterns. Analogous to the mean statistics shown in Figure 2, the RMSE values correspond spatially and proportionally to patterns of mean bias. The same spatial patterns of these three quantities are also observed in the individual models, albeit with varying degrees of magnitude. These patterns suggest that the search for the source of model biases should be focused on the Ohio River Valley region. O_3 production in this region is expected to be highly dependent on what the models assume for emissions of anthropogenic sources of NO_x (e.g., urban activity and large point sources from electrical generating units) and biogenic volatile organic carbon emissions from natural vegetation [*Ryerson et al.*, 2001], as well as the transport and intermixing of these two O_3 precursors [*Gillani and Pleim*, 1996].

[28] The ensemble models can also be diagnosed according to the designation by the AIRNow program of the surrounding population in order to look for systematic differences between the urban, suburban, or rural monitor locations that are differentiated in Figure 1b. Table 2 gives the same median statistics shown in Figure 2 for both ensemble results, for all data, and the three distinct location types. Most monitor types are designated either rural or suburban, and both the ensemble models have slightly higher biases and RMSE for the suburban locations compared to the rural location values, or for all data combined. However, the differences in median statistics between urban, suburban and rural are small, demonstrating that the ensemble statistics shown in Figure 2 are representative of all three monitor classifications, and by extension, independent of the local population density.

5. Bias Correction Techniques

[29] It is difficult to justify the analysis of bias-corrected statistics on the basis of trying to gain additional insight into model deficiencies or performance. However, from an

Table 2. Median Values of the *r* Correlation Coefficient, Mean Bias, and Root-Mean-Square Error for the Seven-Model Ensemble Mean and Ensemble Median of the Maximum 8-Hour Average O_3 Parsed According to the Monitor Location Classification (Urban, Suburban, Rural, or All Locations)

		Ensemble Mean			Ensemble Median		
	Number	*r* Coefficient	Bias, ppbv	RMSE, ppbv	*r* Coefficient	Bias, ppbv	RMSE, ppbv
All data	342	0.76	10.1	14.1	0.74	8.6	13.3
Urban	58	0.77	9.4	14.0	0.76	7.9	13.3
Suburban	129	0.76	10.6	14.4	0.73	8.9	13.4
Rural	132	0.76	9.8	13.9	0.75	8.4	12.9

operational forecast viewpoint, if a simple bias correction can lead to significantly improved forecasts, this is useful information. For the ensemble model there are two basic bias correction strategies from which to choose: bias correct each model at each hour before generating the ensemble, or generate the ensemble from the original models, and resulting average 1-hour or 8-hour maximum values, before applying bias correction. The analysis performed here is based on the latter approach, leaving a comparison to results based on the former approach for future study. Additionally, there are several numerical means by which bias correction can be applied. In this study we examine the effects of two bias correction approaches that are applied to each O_3 monitor site, an additive correction:

$$O_3^{corrected}(i,day) = C(i) + O_3^{model}(i,day) \tag{4}$$

where

$$C(i) = -Mean\ Bias(i),$$

and a multiplicative correction:

$$O_3^{corrected}(i,day) = C(i) \bullet O_3^{model}(i,day) \tag{5}$$

where

$$C(i) = \frac{\sum_{days} O_3^{obs}(i,day)}{\sum_{days} O_3^{model}(i,day)}.$$

Both of these corrections force the mean bias at each O_3 monitor to zero. It should also be noted that *r* correlation coefficients are independent of any linear transformation of the original set of model O_3, and therefore the *r* coefficients in Figure 2 are unaffected by either of these bias correction approaches. The first approach, hereafter referred to as "mean subtraction," is the standard, and most often used bias correction in meteorological analysis. The second approach, hereafter referred to as "ratio adjusted," is an alternative correction that may be more applicable to O_3 since corrected mixing ratios will always be nonnegative. This is not always guaranteed with the mean bias subtraction. Bias corrections based on linear combinations of equations (4) and (5) are also possible, but the intercept and slope coefficients determined from least square fitting depend on the metric chosen to minimize (RMSE, distance from 1-to-1 line, . . ., etc). We choose to focus the analysis on the two simplest one-parameter correction schemes to illustrate the basic effect of an additive versus multiplicative correction. Another important simplification is that bias corrections are calculated from comparisons over the entire 56-day period. Statistical evaluations for bias corrections based on training periods from one to tens of days are given elsewhere [*Pagowski et al.*, 2005; J. M. Wilczak et al., manuscript in preparation, 2005].

[30] Figure 4a shows the effect of the two bias correction strategies on median RMSE for the set of maximum 8-hour average O_3 comparisons. Also overlaid in Figure 4a is the RMSE determined by persistence. This persistence forecast is based only on the observations, and assumes that the forecast for any given O_3 monitor on any day is just the observations at that monitor from the previous day. The fractional number of sites having lower RMSE than the persistence forecast can be used as a measure of skill, and models having 50% or more points with lower RMSE

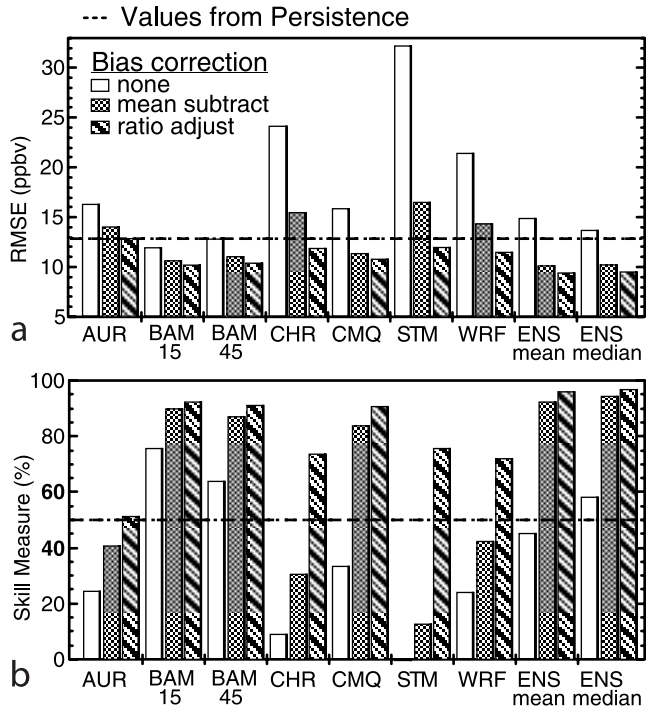
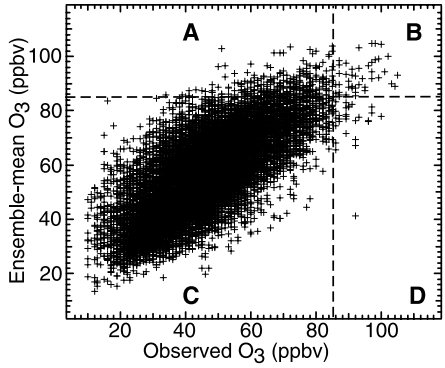


Figure 4. (a) Median RMSE of maximum 8-hour average O_3 for the uncorrected and the two bias-corrected cases. The dashed line is the median RMSE for the persistence forecast based on observed maximum 8-hour average O_3. (b) Skill score for maximum 8-hour average O_3 for the uncorrected and the two bias-corrected cases. The dashed line is the 50% value, or the break-even point with the persistence forecast.



$$\text{Probability of Detection} = \frac{B}{B+D}$$

$$\text{False Alarm Rate} = \frac{A}{B+A}$$

$$\text{Critical Success Index} = \frac{B}{B+A+D}$$

$$\text{Bias Ratio} = \frac{A+B}{B+D}$$

Figure 5. Maximum 8-hour average O_3 for the seven-model ensemble mean (no bias correction) versus observations for 16,487 total monitor days. The dashed lines represent the 85 ppbv threshold for the model and observations. The number of points in the four quadrants labeled A–D are used in the definitions of POD, FAR, CSI, and bias ratio.

compared to persistence are considered to have some skill. Model comparisons for this quantity are shown in Figure 4b, and are approximately inversely proportional to the RMSE values in Figure 4a. Under all cases, the skill score increases (RMSE decreases) when bias correction is applied. When the bias corrections are applied to the ensemble mean, the forecast goes from having no skill to more than a 90% skill score, but only slightly higher than the bias-corrected scores for three of the AQFMs. The ensemble median shows a small degree of skill without any bias correction, but nearly identical skill as the ensemble mean for the two bias correction cases. The ratio adjusted bias correction improves skill even further compared to the mean subtraction correction, especially for the three models with highest mean bias shown in Figure 2. As discussed further in section 7 below, these three models also display high O_3 variance compared to observed O_3 variance, and a ratio-adjusted bias correction reduces their variance to more closely match observed levels, while the mean subtraction correction has no affect on variance. One would naturally expect the bias correction that more closely reproduces observed O_3 variance to also have better bulk statistical scores.

[31] A general, valid question related to O_3 forecasting is: What fraction of O_3 forecast errors is due to limitations in describing the dynamic meteorology as opposed to uncertainties related to O_3 photochemistry and precursor emissions? Separating out these two broad categories of errors for any particular model is difficult because of the nonlinear dependence of O_3 formation on precursors and hence the PBL dynamics, meteorology, and coupled photochemistry that determine precursor distributions. However, the ensemble errors in Figure 4a are not particular to any given model, and can be generalized as a limitation of our collective understanding based on seven independent model representations. Figure 4a shows the median, uncorrected ensemble mean RMSE to be 14 ppbv, and an apparent minimum median RMSE of about 9 ppbv regardless of model configuration or bias correction. Simplistically, one can attribute the 9 ppbv RMSE that is independent of bias correction to be an upper limit of errors due to random or meteorological noise. This value is an upper limit since additional errors originating from fundamental model formulations such as grid resolution, physical and chemical parameterizations, and numerical approximations contribute to some unknown degree. The remaining 5 ppbv of the median, uncorrected ensemble mean RMSE would then be attributed to biases that could presumably be eliminated with proper adjustment of anthropogenic and biogenic O_3 precursor emissions. From this highly simplified ensemble perspective a 64% upper limit of O_3 forecast errors would be attributed to meteorological uncertainty and the remaining 36% a lower limit associated with O_3 precursors and photochemistry.

6. Threshold Statistics for the Models and Ensembles

[32] Figures 2–4 provide an analysis of median or bulk conditions for O_3 over the 2004 summer time period. However, a main justification for real-time O_3 forecasts is the information they provide for public pollution exposure and health advisories [e.g., *Dabberdt et al.*, 2004]. For health-related issues the useful information is not contained in the bulk statistics, but rather in the occurrence of maximum average O_3 values greater than a particular threshold value. Two threshold values that have previously been considered are the 125 ppbv maximum 1-hour average O_3 limit, and the 85 ppbv maximum 8-hour average O_3 limit [*McHenry et al.*, 2004; *Kang et al.*, 2005]. There were no exceedances of the maximum 1-hour average 125 ppbv O_3 threshold for the period and stations used in this analysis, so only threshold statistics for the 85 ppbv maximum 8-hour averages are presented.

[33] Figure 5 shows the ensemble mean model versus observations of the daily maximum 8-hour average O_3 with an 85 ppbv threshold limit superimposed on both axes. The number of model-measurement comparison points that lie within the four quadrants of the figure define four quantities that are used in the threshold statistical analysis; probability of detection (POD), false alarm rate (FAR), critical success index (CSI), and the bias ratio. For a perfect model the POD would be 100%, FAR would be zero, CSI would be 100%, and the bias ratio would be unity. Bias removal from the ensemble mean forecasts would tend to move the entire group of points vertically along the *y* axis, giving preference

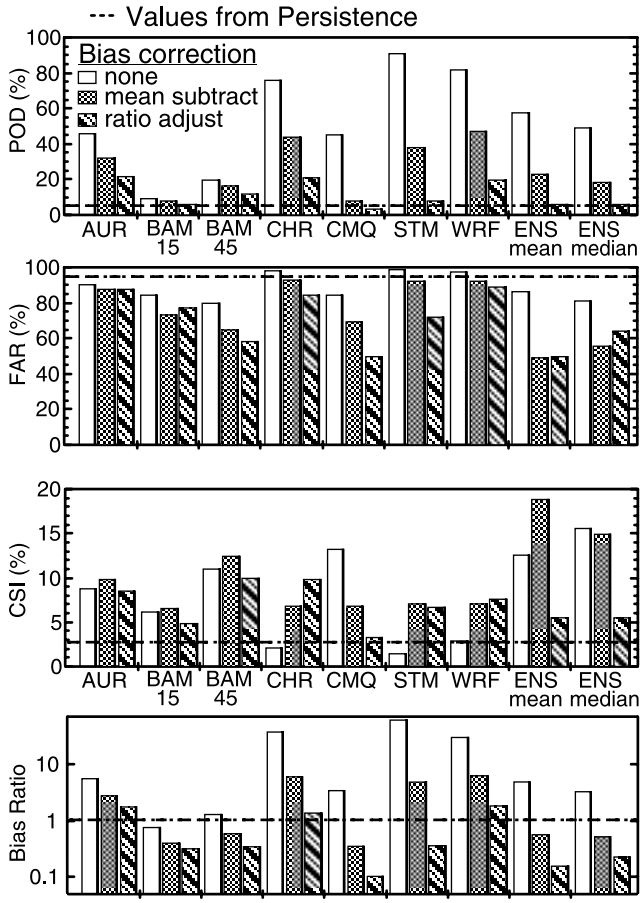


Figure 6a. A summary of the maximum 8-hour average O_3 threshold statistics referred to in Figure 5 for the uncorrected and the two bias-corrected cases, using a maximum 8-hour average O_3 threshold value of 85 ppbv. The dashed lines are results derived from the persistence forecasts.

to either a high POD or low FAR at the expense of the other. The CSI represents a balanced measure of the combined FAR and POD if the importance of model and observed exceedances are equally weighted. The bias ratio represents a bulk measure of the model threshold exceedance ratio relative to the observed ratio. It should be noted that only 87 observations in Figure 5 exceed the 85 ppbv maximum 8-hour average threshold (out of 16,487 total points), and thus the threshold analysis is for a relatively small sample at the very tail of the observed O_3 population distribution.

[34] The threshold statistics for the seven AQFMs, the ensembles, and the impact of the two bias correction techniques are shown in Figure 6a. The relative variation of POD values between the various model cases is nearly identical to that of the bias ratios. Both of these quantities are reduced to varying degrees by the mean subtraction bias correction. This decrease is even more pronounced for the case of ratio adjusted bias correction, with the degree of additional decrease related to the magnitude of bias correction associated with each model. Though bias corrections detrimentally influence the POD of each model, the FAR values all show beneficial reductions. The ensemble models yield the largest decrease in FAR with mean subtraction bias correction, which is probably due to the better overall correlation coefficient of the ensembles shown in Figure 2. The CSI values show an increase with the mean subtraction bias correction for all models except CMAQ/Eta and the ensemble median, while the ratio adjusted bias correction has a mixed effect. For the four models with the highest bias corrections (CHRONOS, CMAQ/Eta, STEM and WRF/Chem) there is a tendency for a particular model configuration to have a higher CSI when the bias ratio is closest to 1. For the ensemble mean model in particular a mean subtraction bias correction provides an optimal balance between POD and FAR. The ratio adjusted bias correction degrades bias ratio, POD and CSI substantially for both ensembles.

[35] It is difficult to assign a statistical significance to the POD, FAR, and CSI differences between models shown in Figure 6a since no average or median is calculated, and an extrapolation to a larger sample population is not assumed. Applying a sampling uncertainty of $1/\sqrt{N}$, where N is number of observations above the 85 ppbv threshold, yields a value of $\sim$11% with the sampling uncertainty for the model results even higher for bias ratios less than 1. We therefore test the veracity of the CSI results in Figure 6a by changing the threshold value, and recompute the sensitivity of the threshold measures. Figure 6b shows the CSI values of the various models and their bias correction cases when the maximum 8-hour average O_3 threshold is reduced to 80 ppbv (209 observations above the 80 ppbv threshold). POD and FAR values are less than 5% different than the corresponding values for a 85 ppbv threshold for all model cases, and are thus not shown. Figure 6b shows that with this lower threshold all CSI values (even for persistence) increase a couple percent, and two model cases increase as much as 6%. The relative patterns of the different model cases are nonetheless very similar to those for CSI in Figure 6a. Therefore the higher CSI values for the ensemble mean with mean subtraction bias correction in Figure 6a does not appear to be an artifact of low-population statistics.

7. Relationship Between O_3 Variance and Threshold Statistics

[36] Though the threshold statistics, specifically the CSI response, to the two bias correction techniques in Figure 6a are somewhat confusing, there is another quantity to consider that helps explain each model's sensitivity. We define the variance at a given monitor to be the square of the standard deviation about the average O_3:

$$\text{Variance(i)} = \left(\frac{1.}{\text{N}_{days}}\right) \sum_{days} (O_3(i, day) - O_3(i, average))^2 \quad (6)$$

where O_3 can either be observed or model daily maximum average O_3. This quantity is chosen because it represents the power of the O_3 signal about the mean from a purely signal processing point of view. Variance histograms (in fraction of monitor sites) for the observations, two AQFMs and the ensembles are shown in Figure 7. It is important to note that this variance quantity is unaffected by a mean subtraction bias correction. In contrast the variance can be very sensitive to ratio adjusted bias corrections, since the

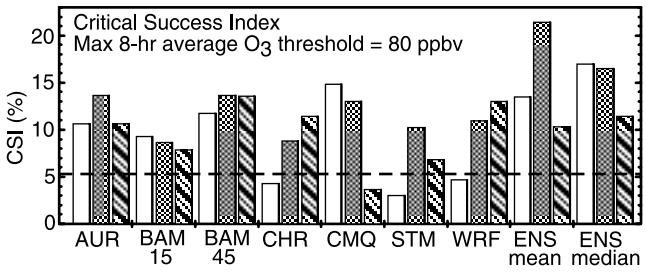


Figure 6b. Critical success index (CSI) as in Figure 6a, except the threshold value of the maximum 8-hour average O_3 is taken to be 80 ppbv instead of 85 ppbv. The meaning of the bar shading and dashed line is the same as in Figure 6a.

resulting model variance from this correction is equal to the original model variance multiplied by the square of the multiplicative correction factor (the $C(i)$ term in equation (5)). For the dominant case of positive model bias this correction term is less than 1, resulting in reduced model variance.

[37] The variance histogram in Figure 7 for the observations resembles a Gaussian profile centered at ~180 ppbv^2, corresponding to a median standard deviation for the observations of 13.3 ppbv. The variance histogram for the uncorrected or mean-subtraction-bias-corrected BAMS-15 km model is skewed toward lower values compared to the histogram for the observations. Since little bias correction is needed for this model, the histogram corresponding to the ratio-adjusted correction is nearly identical and is not shown. The variance histogram for WRF/Chem shows this model has too much variance relative to the observations without bias correction or with the mean subtraction correction. The ratio adjusted bias correction brings the histogram into better agreement with the observations, but still somewhat skewed to the high end. The variance histogram for the ensemble mean with mean subtraction correction (or no correction) appears to match the observed histogram rather well with no obvious skew to the high or low sides. When the ratio adjusted bias correction is applied to the ensemble, variance is reduced significantly and the histogram is skewed too far to the low end. Figure 7 suggests that the median variance within each histogram may be a useful quantity for comparing the models' variance with that of the observations.

[38] Figure 8 shows the median values of O_3 variance for the various bias correction cases of the seven AQFMs and the ensembles along with the median variance from the observations. With mean subtraction correction (or no correction) the seven models are divided between 4 that have too much variance and three that have too little. The ensemble mean model variance represents a balance between the various models, and the fact that the ensemble mean matches the observed variance appears to be a result of compensation, since none of the models match the observations as well. This good match is probably fortuitous, and may not generally apply to the ensemble of an arbitrary set of forecasts.

[39] The median variance values shown in Figure 8 help to explain the sensitivity of the threshold statistics shown in Figure 6a to changes in the bias correction method. In general the closer the model variance matches observed variance the closer the bias ratio in Figure 6a is to 1. This is expected, since the number of model forecasts exceeding the threshold intuitively should be correlated with model noise or variance. The CMAQ/Eta and ensemble median models are the only models in Figure 6a to show a decrease in CSI when the mean subtraction bias correction is applied to the uncorrected forecast. However, these models are unique in that they have significant bias, but have lower

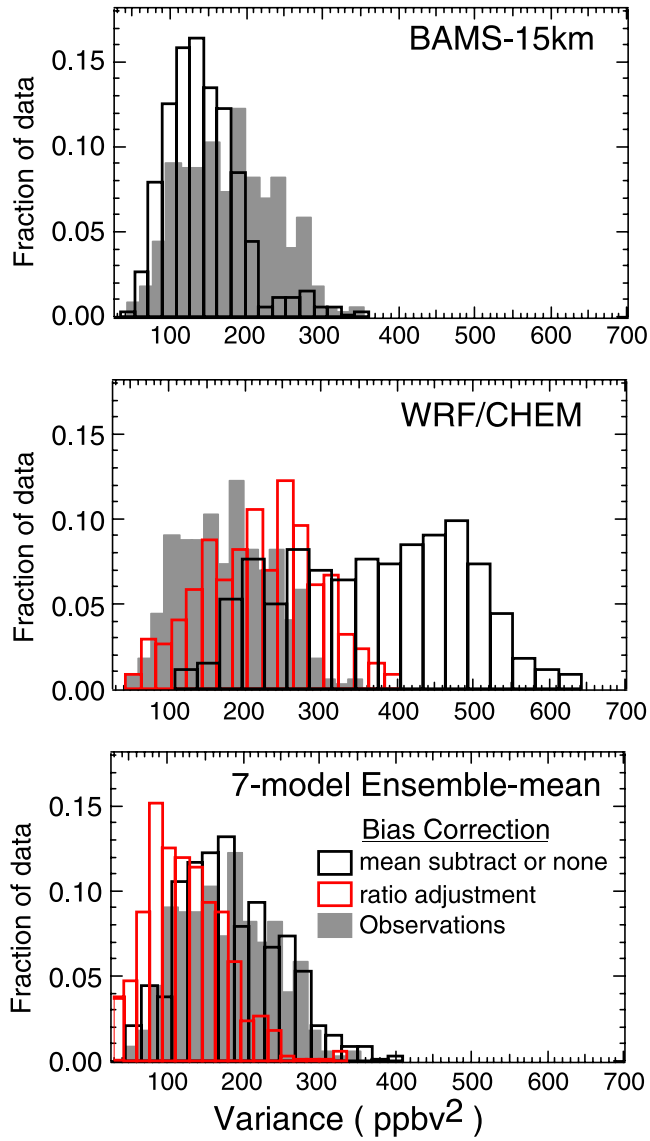


Figure 7. Histograms (in fraction of total data) of maximum 8-hour average O_3 variance ((standard deviation)^2) for the observations (gray shaded) and three models (unshaded). The case of no bias correction, which is the same as the case with mean subtraction correction, is in black. The ratio-adjusted bias correction case is in red. The ratio-adjusted case for the BAMS-15 km model is not shown since low biases make it nearly the same as the other case. The collection intervals are chosen so that 18 bins (square root of monitor number) cover the variance range of each model.

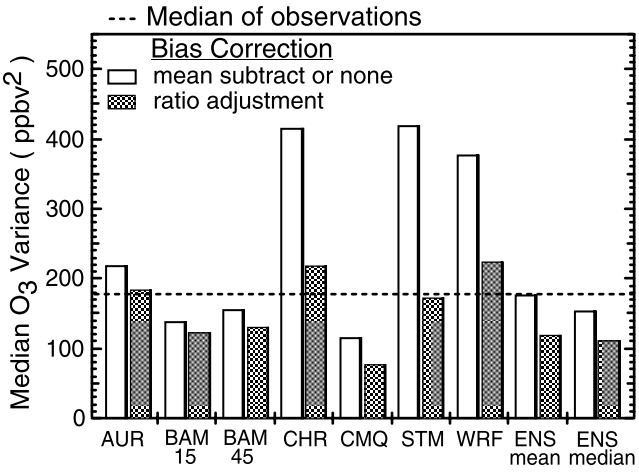


Figure 8. Median values of the maximum 8-hour average O_3 variance ((standard deviation)^2) for the seven AQFMs, the ensemble mean, and the ensemble median, for the two bias correction approaches. Median values for each model are taken from comparisons at the 342 monitor locations having 30 or more days of available data. The dashed line shows the median of the observed O_3 variance.

variance compared to the observations and the other models. Removal of the mean bias makes their bias ratios much less than 1, as expected from the low variance, which in turn seriously affects POD. Also, in Figures 6a and 6b the CSI values of all models tend to be higher the closer the corresponding median variance in Figure 8 is to the observations. However, the STEM model is a clear exception to this pattern and the WRF/Chem model shows only slight improvement in CSI with a large decrease in variance. The ratio adjusted bias correction is obviously affecting populations of points to the left or right of the 85 ppbv threshold in Figure 5 differently for each model. Further analysis of results for the particular monitors and episodes when observed O_3 exceeded the threshold is needed in order to examine model response differences further.

8. Sensitivity of the Ensemble to Each Model

[40] The contribution of each model to the statistics of the seven-model ensemble mean can be assessed by eliminating each model from the ensemble separately, and comparing statistics for the resulting six-member ensembles. The previous analysis suggests that there are clear advantages to the mean subtracted bias correction of the ensemble mean forecast. An analysis is therefore also done for the case where a model is removed from the ensemble mean, and the mean subtraction bias correction is applied to the resulting six-member ensemble.

[41] Figure 9 shows the resulting mean bias, RMSE, and skill measures for the various six-member ensembles along with the seven-model ensemble mean results overlaid as straight lines. It is important to note that the model name on the abscissa corresponds to the model removed from the seven-model ensemble, and therefore a decrease in model bias when a model is removed corresponds to a model with higher than average bias. When the model with highest mean bias in Figure 2 is removed from the uncorrected ensemble, the mean bias is reduced ~25%, but the resulting 7.5 ppbv bias is still appreciable, and the associated RMSE only decreases ~10%. As expected, for the uncorrected ensemble the sensitivity of RMSE and the skill measure correlate directly with individual model bias, and the removal of any given model has a minor effect compared to the bias correction itself. For the bias-corrected case, the removal of any particular model has very little effect on the resulting RMSE or skill measure compared to the uncorrected case. It therefore appears the combined effects of ensemble averaging and bias removal tend to diminish the influence of a particular model more than just the ensemble averaging operation itself.

[42] Figure 10 shows the threshold statistics for the various threshold scores in a manner similar to Figure 9. For both bias-corrected and uncorrected ensemble mean cases the only quantity that is somewhat sensitive to model removal is the bias ratio. Yet despite factor of two type

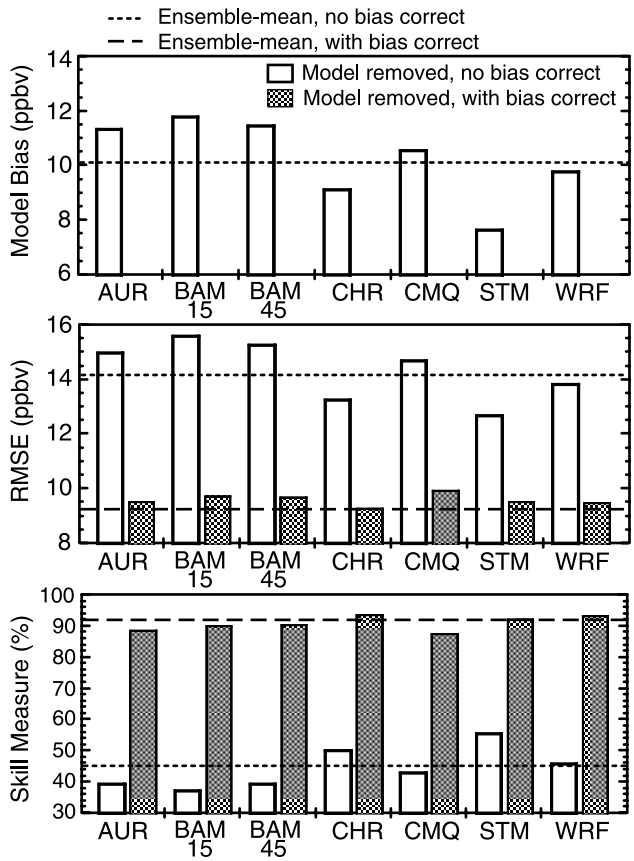


Figure 9. Median values of the six-member ensemble mean for the (top) r correlation coefficient, (middle) mean bias in ppbv, and (bottom) root-mean-square error in ppbv when each model is removed individually from the ensemble mean (model removed is on the x axis). Short-dashed lines are median values of the seven-model ensemble mean with no bias correction, and the long-dashed lines are the seven-model ensemble mean with mean subtraction as the bias correction. Unshaded bars correspond to the case with no bias correction, and shaded bars correspond to the bias-corrected case.

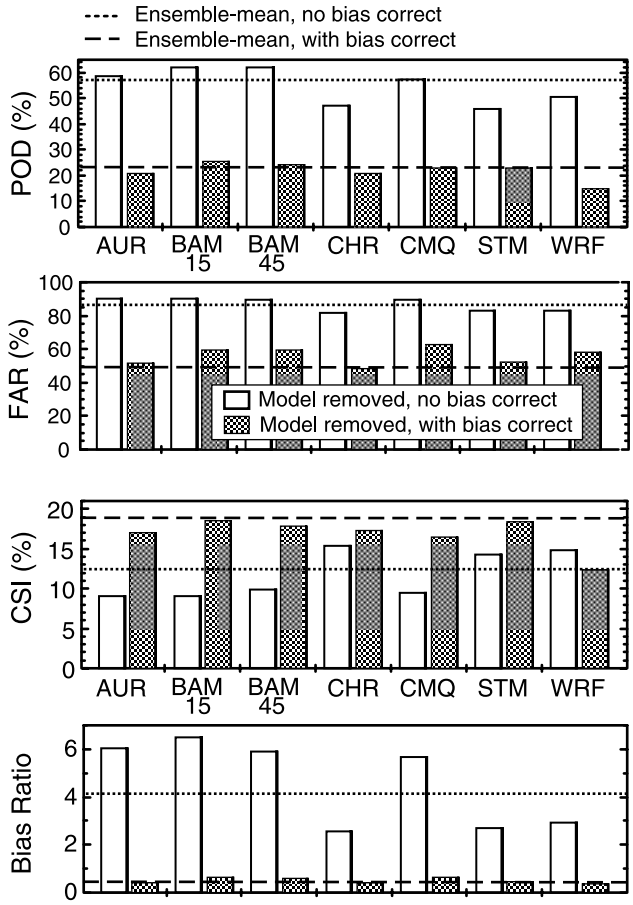


Figure 10. Median values of the six-member ensemble mean for the threshold statistics POD, FAR, CSI, and bias ratio when each model is removed individually from the ensemble mean (model removed is on the *x* axis). The meaning of short-dashed and long-dashed lines and shaded versus unshaded bars is the same as in Figure 9.

changes in the bias ratio, false alarm rates are particularly insensitive to removal of any particular model. It is interesting that for the bias-corrected ensemble mean FARs do not decrease if any model is removed, and likewise CSI values do not increase if any model is removed. For the bias-corrected case one model appears to have a disproportionately positive influence on CSI, but the significance of this influence is on the edge of the 4 to 6% uncertainty assigned to CSI from the threshold sensitivity calculations previously discussed in connection with Figures 6a and 6b.

9. Conclusions

[43] As part of the NEAQS-2004 model verification project, daily observations of surface O_3 from the AIRNow network are used to assess forecasts from a mean ensemble and a median ensemble, both determined from results of 7 real-time air quality models. Several statistical measures for the ensembles and individual models are evaluated, and include *r* correlation coefficients, mean bias, and RMSE. For median conditions over 342 O_3 monitors in the eastern United States and southern Canada, *r* correlation coefficients are significantly higher for the model ensembles than for any individual model. Because all of the models exhibit positive O_3 bias, the ensembles also possess significant positive bias. This large bias directly accounts for high RMSE values and low model skill for forecast daily maximum O_3. The sources of persistent positive model bias requires further analysis and could be due to a number of factors such as a common overestimate of O_3 precursor emissions or high O_3 boundary conditions during the relatively low O_3 summer of 2004. There are spatial patterns associated with *r* correlation coefficients, mean bias, and RMSE that show the ensembles perform much better along the eastern U.S. corridor, New England and southern Canada in comparison to a broad region that straddles the Ohio River Valley. The statistical analysis for maximum 8-hour average O_3 is reapplied to the model and ensemble forecasts using two bias correction techniques: a mean bias subtraction, and a multiplicative type ratio adjustment. Bias correction from both methods reduces RMSE and increases skill, but the ratio-adjusted method provides additional improvement over the mean subtraction method, particularly for the three AQFMs with the highest biases.

[44] Threshold statistics for 85 ppbv maximum 8-hour average O_3 limits are also presented for the uncorrected and bias-corrected ensemble and individual model results. These statistical scores are also related directly to O_3 bias for the uncorrected models. For the AQFMs with significant model bias, as well as the ensembles, the method of bias correction has a large impact on the threshold statistics. Somewhat surprisingly, the ratio-adjusted bias correction improves the bulk statistical scores under all cases but adversely affects the threshold statistics for most cases. The response of the threshold statistics to a particular bias correction method appears to be related to the degree to which a particular model's O_3 variance matches the observed variance. The ratio-adjusted bias correction always decreases model O_3 variance for the case of positive model bias, as well as the POD (probability of detection) statistic compared to a mean subtraction correction. Thus threshold statistics for uncorrected models that have too little O_3 variance, or variance that matches observations, are adversely affected by a ratio-adjusted correction. The O_3 variance of the ensemble mean model with a mean subtraction bias correction, or with no correction, matches observed O_3 variance well, and appears to represent a compensating balance between 3 AQFMs with too little O_3 variance and 4 AQFMs with too much variance.

[45] If one contrasts the standard statistical measures of *r* correlation coefficient, mean bias, and RMSE against the threshold statistics for the individual AQFMs, there is no preferred AQFM that stands out as best in both categories regardless of bias correction or not. Though the seven-model ensembles with mean subtraction bias correction have relatively low PODs (~20%) compared to some individual models, their false alarm rates (FAR) are the lowest of all models, giving them the highest value of the critical success index (CSI) threshold statistic, which is a balanced measure between FAR and POD. The ensemble median forecast without bias correction shows better bulk and threshold statistics for all measures except RMSE, RMSE related skill score, and POD compared to any uncorrected individual model, and the bias-corrected ensemble mean forecast shows better performance for all statistical measures except POD compared to any other

model case. The results of this study clearly points to the preference of the ensemble cases as a forecast over any individual forecast model. A simple model removal analysis shows that the statistics of the bias-corrected ensemble mean are quite insensitive to the removal of any particular AQFM, and that the mean subtraction bias correction improves all statistical measures (except for bias-correlated POD) much more than removing any AQFM from the uncorrected ensemble.

[46] Since the primary value of real-time O_3 forecasts is their guidance for issuing health advisories and alerts, it can be argued that the threshold statistics, rather than mean bias and associated RMSE, are the most important quantities to consider. The analysis presented here illustrates that matching observed O_3 variance, along with the elimination of model bias, are both necessary for useful real-time O_3 forecasts.

[47] It is important to note that all the forecast models presented here have undergone revisions that make any statistical inferences about a particular model obsolete. The focus is on establishing the utility of the O_3 ensemble, and the bias corrections required to yield significant statistical improvement over any single model forecast. The models used in the ensemble were chosen on the basis of their availability within a broader cooperative model evaluation project of the ICARTT/NEAQS-2K4 study. A better ensemble forecast could possibly be devised from forecasts designed for ensemble inclusion. For example the effects of older, less reliable emissions inventories would be eliminated if all ensemble elements used the most recent emission estimates available, but maintain a diverse set of meteorological foundations. Alternatively, an ensemble of model forecasts that explicitly reflect known uncertainties in O_3 precursor emissions all contained within the "best" possible meteorological framework may provide the optimum combination. Additional investigation into optimum ensemble design within a set of all possible combinations and permutations of AQFM configurations is obviously needed to address conceptual issues such as this. Likewise, the practical implementation of ensemble O_3 forecasts within an operational setting is not addressed here. One obvious issue is computational expense, which in this case was shared by seven individual institutions. This study has also taken advantage of a full season's data to derive bias correction terms, and thus represents a limiting case where biases are well characterized. However, this same seven-model data set has been used for further statistical evaluation of alternative ensemble strategies, training period requirements, and bias correction applications that are more applicable to an operational ensemble framework [*Pagowski et al.*, 2005; J. M. Wilczak et al., manuscript in preparation 2005].

[48] **Acknowledgments.** This research is partially funded by Early Start Funding from the NOAA/NWS Office of Science and Technology and the NOAA Office of Atmospheric Research Weather and Air Quality Program and would not be possible without the participation of the AIRNow program and participating stakeholders. Credit for program support and management is given to Paula Davidson (NOAA/NWS/OST), Steve Fine (NOAA/ARL), and Jim Meagher (NOAA/AL). Computational and logistic assistance from the following individuals and organizations is also gratefully appreciated: Amenda Stanley (NOAA/FSL), Ted Smith (Baron AMS), Jessica Koury (NOAA/ETL), Wendi Madsen (NOAA/ETL), Ann Keane (NOAA/ETL), Sophie Cousineau (MSC), L.-P. Crevier (MSC), Stephane Gaudreault (MSC), Mike Moran (AQRB/MSC), Paul Makar (AQRB/MSC), Balbir Pabla (AQRB/MSC), Dezsö Dévényi (FSL/NOAA), and the NOAA/FSL High Performance Computing Facility. The thoughtful suggestions and comments from three anonymous reviewers are also greatly appreciated. One author's (R.M.) contributions presented here were 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. Although it has been reviewed by EPA and NOAA and approved for publication, it does not necessarily reflect their policies or views.

References

Atkinson, R., D. L. Baulch, R. A. Cox, R. F. Hampson, J. A. Kerr, and J. Troe (1992), Evaluated kinetic and photochemical data for atmospheric chemistry: Supplement IV, *Atmos. Environ., Part A*, *26*, 1187–1230.

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

Carmichael, G. R., L. K. Peters, and T. Kitada (1986), A second generation model for regional-scale transport chemistry and deposition, *Atmos. Environ.*, *20*, 173–188.

Carmichael, G. R., et al. (2003), Regional-scale chemical transport modeling in support of intensive field experiments: Overview and analysis of the TRACE-P observations, *J. Geophys. Res.*, *108*(D21), 8823, doi:10.1029/2002JD003117.

Carter, W. P. L. (1996), Condensed atmospheric photooxidation mechanism for isoprene, *Atmos. Environ.*, *30*, 4275–4290.

Carter, W. (2000), Documentation of the SAPRC-99 chemical mechanism for VOC reactivity assessment, final report to California Air Resources Board, contract 92–329, Univ. of Calif., Riverside, 8 May.

Chang, J. S., Y. Li, M. Beauharnois, H.-C. Huang, C.-H. Lu, and G. Wojcik (1996), *SAQM User's Guide*, 500 pp., Calif. Air Resour. Board, Sacramento.

Coats, C. J., Jr. (1996), High-performance algorithms in the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system, paper presented at Ninth AMS Joint Conference on Applications of Air Pollution Meteorology With A&WMA, Am. Meteorol. Soc., Atlanta, Ga.

Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth (1998a), The operational CMC-MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation, *Mon. Weather Rev.*, *126*, 1373–1395.

Côté, J., J.-G. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth (1998b), The operational CMC/MRB Global Environmental Multiscale (GEM) model. Part II: Results, *Mon. Weather Rev.*, *126*, 1397–1418.

Dabberdt, W. F., et al. (2004), Meteorological research needs for improved air quality forecasting, *Bull. Am. Meteorol. Soc.*, *85*, 563–586.

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

Delle Monache, L., X. Deng, Y. Zhou, H. Modzelewski, G. Hicks, T. Cannon, R. B. Stull, and C. di Cenzo (2004), Ensemble air quality forecasts over the Lower Fraser Valley, British Columbia: A summer 2004 case study, paper presented at 13th AMS Conference on the Applications of Air Pollution Meteorology With A&WMA, Am. Meteorol. Soc., Vancouver, B. C.

DeMore, W. B., et al. (1994), Chemical kinetics and photochemical data for use in stratospheric modeling: Evaluation number 11, *JPL Publ., 84–26*, 273 pp.

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

Gery, M. W., et al. (1989), A photochemical kinetics mechanism for urban and regional scale computer models, *J. Geophys. Res.*, *94*, 12,295–12,356.

Gillani, N., and J. E. Pleim (1996), Sub-grid-scale features of anthropogenic emissions of NO_x and VOC in the context of regional Eulerian models, *Atmos. Environ.*, *30*, 2043–2059.

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.

Grell, G. A., J. Dudhia, and D. R. Stauffer (1994), A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5), *NCAR Tech. Note NCAR/TN-398+STR*, 122 pp., Natl. Cent. for Atmos. Res., Boulder, Colo.

Grell, G. A., S. Emeis, W. R. Stockwell, T. Schoenemeyer, R. Forkel, J. Michalakes, R. Knoche, and W. Seidl (2000), Application of a

multiscale, coupled MM5/chemistry model to the complex terrain of the VOTALP valley campaign, *Atmos. Environ.*, *34*, 1435–1453.
Grell, G. A., S. A. McKeen, J. Michalakes, J.-W. Bao, M. Trainer, and E.-Y. Hsie (2002), Real-time simultaneous prediction of air pollution and weather during the Houston 2000 field experiment, paper presented at 4th Conference on Atmospheric Chemistry, Am. Meteorol. Soc., Orlando, Fla., 13–17 Jan.
Grell, G. A., S. E. Peckham, R. Schmitz, S. A. McKeen, G. Frost, W. Skamarock, and B. Eder (2005), Fully coupled "online" chemistry within the WRF model, *Atmos. Environ.*, in press.
Guenther, A., P. Zimmerman, and M. Wildermuth (1994), Natural volatile organic compound emission rate estimates for U.S. woodland landscapes, *Atmos. Environ.*, *28*, 1197–1210.
Guenther, A. B., et al. (1995), A global model of natural volatile organic compound emissions, *J. Geophys. Res.*, *100*, 8873–8892.
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.
Johnson, N. L., and F. C. Leone (1977), *Statistics and Experimental Design in Engineering and the Physical Sciences*, vol. 1, 2nd ed., 592 pp., John Wiley, Hoboken, N. J.
Kalnay, E. (2003), *Atmospheric Modeling, Data Assimilation and Predictability*, 341 pp., Cambridge Univ. Press, New York.
Kang, D., B. K. Eder, A. F. Stein, G. A. Grell, S. E. Peckham, and J. McHenry (2005), The New England Air Quality Forecasting Pilot Program: Development of an evaluation protocol and performance benchmark, *J. Air Waste Manage. Assoc.*, in press.
Kasibhatla, P., W. L. Chameides, B. Duncan, M. Houyoux, C. Jang, R. Mathur, T. Odman, and A. Xiu (1997), Impact of inert organic nitrate formation on ground-level ozone in a regional air quality model using the Carbon Bond Mechanism 4, *Geophys. Res. Lett.*, *24*(24), 3205–3208.
Lamb, B., D. Gay, H. Westberg, and T. Pierce (1993), A biogenic hydrocarbon emission inventory for the USA using a simple forest canopy model, *Atmos. Environ., Part A*, *27*, 1673–1690.
Lurmann, F. W., A. C. Lloyd, and R. Atkinson (1986), A chemical mechanism for use in long-range transport/acid deposition computer modeling, *J. Geophys. Res.*, *91*, 10,905–10,936.
Mathur, R., et al. (2004), Adaptation and applications of the Community Multiscale Air Quality (CMAQ) modeling system for real-time air quality forecasting during the summer of 2004, paper presented at 3rd Annual CMAS Models-3 Users Workshop, CMAS Cent., Univ. of N. C., Chapel Hill.
McHenry, J. N., and C. J. Coats (2003), Improved representation of cloud/actinic flux interaction in multiscale photochemical models, paper presented at Fifth Conference on Atmospheric Chemistry: Gases, Aerosols, and Clouds, Am. Meteorol. Soc., Long Beach, Calif.
McHenry, J. N., N. Seaman, C. J. Coats, A. Lario-Gibbs, J. Vukovich, N. Wheeler, and E. Hayes (1999), Real-time nested mesoscale forecasts of lower tropospheric ozone using a highly optimized coupled model numerical prediction system, paper presented at Symposium on Interdisciplinary Issues in Atmospheric Chemistry, Am. Meteorol. Soc., Dallas, Tex.
McHenry, J. N., W. F. Ryan, N. L. Seaman, C. J. Coates Jr., J. Pudykiewicz, S. Arunachalam, and J. M. Vukovich (2004), A real-time Eulerian photochemical model forecast system, *Bull. Am Meteorol. Soc.*, *85*, 525–548.
McKeen, S. A., G. Wotawa, D. D. Parrish, J. S. Holloway, M. P. Buhr, G. Hubler, F. C. Fehsenfeld, and J. F. Meagher (2002), Ozone production from Canadian wildfires during June and July of 1995, *J. Geophys. Res.*, *107*(D14), 4192, doi:10.1029/2001JD000697.
McQueen, J., et al. (2004), Development and evaluation of the NOAA/EPA prototype air quality model prediction system, paper presented at 20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction, Am. Meteorol. Soc., Seattle, Wash., 11–15 Jan.
Moran, M. D., M. T. Scholtz, C. F. Slama, A. Dorkalam, A. Taylor, N. S. Ting, D. Davies, P. A. Makar, and S. Venkatesh (1997), An overview of CEPS1.0: Version 1.0 of the Canadian Emissions Processing System for regional-scale air quality models, paper presented at 7th A&WMA Emission Inventory Symposium, Air and Waste Manage. Assoc., Research Triangle Park, N. C., 28–30 Oct.
Otte, T. L., et al. (2005), Linking the Eta model with the Community Multiscale Air Quality (CMAQ) modeling system to build a national air quality forecasting system, *Weather Forecasting*, *20*, 367–384.
Pagowski, M., et al. (2005), A simple method to improve ensemble-based ozone forecasts, *Geophys. Res. Lett.*, *32*, L07814, doi:10.1029/2004GL022305.
Pierce, T., C. Geron, L. Bender, R. Dennis, G. Tonneson, and A. Guenther (1998), Influence of increased isoprene emissions on regional ozone modeling, *J. Geophys. Res.*, *103*, 25,611–25,629.
Pierce, T., C. Geron, G. Pouliot, E. Kinnee, and J. Vukovich (2002), Integration of the Biogenic Emission Inventory System (BEIS3) into the Community Multiscale Air Quality Modeling System, paper presented at 12th Joint Conference on the Applications of Air Pollution Meteorology With the A&WMA, Am. Meteorol. Soc., Norfolk, Va., 20–24 May .
Pouliot, G. A. (2005), The emissions processing system for the Eta/CMAQ air quality forecast system, paper presented at 7th Conference on Atmospheric Chemistry, 85th AMS Annual Meeting, Am. Meteorol. Soc., San Diego, Calif.
Pudykiewicz, J., A. Kallaur, and P. K. Smolarkiewicz (1997), Semi-Lagrangian modeliling of tropospheric ozone, *Tellus, Ser. B*, *49*, 231–258.
Ryerson, T., et al. (2001), Observations of ozone formation in power plant plumes and implications for ozone control strategies, *Science*, *292*, 719–723.
Schoenemeyer, T., K. Richter, and G. Smiatek (1997), Vorstudie über ein raumlich und zeitlich aufgelöstes Kataster anthropogener und biogener Emissionen für Bayern mit Entwicklung eines Prototyps und Anwendung für Immissionsprognosen, Abschluss bericht an das Bayerische Landesamt für Umweltschutz, Fraunhofer-Inst. für Atmos. Umweltforsch., Garmisch-Partenkirchen, Germany.
Simpson, D., A. Guenther, C. N. Hewitt, and R. Steinbrecher (1995), Biogenic emissions in Europe: 1. Estimates and uncertainties, *J. Geophys. Res.*, *100*, 22,875–22,890.
Stockwell, W. R., P. Middleton, J. S. Chang, and X. Tang (1990), The second generation regional acid deposition model chemical mechanism for regional air quality modeling, *J. Geophys. Res.*, *95*, 16,343–16,367.
Stockwell, W. R., P. Middleton, J. S. Chang, and X. Tang (1995), The effect of acetyl peroxy-pperoxy radical reactions on peroxyacetyl nitrate and ozone concentrations, *Atmos. Environ.*, *29*, 1591–1599.
Stockwell, W. R., F. Kirchner, M. Kuhn, and S. Seefeld (1997), A new mechanism for regional atmospheric chemistry modeling, *J. Geophys. Res.*, *102*, 25,847–25,879.
Tang, Y., et al. (2003), Impacts of aerosols and clouds on photolysis frequencies and photochemistry during TRACE-P: 2. Three-dimensional study using a regional chemical transport model, *J. Geophys. Res.*, *108*(D21), 8822, doi:10.1029/2002JD003100.
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.
Tilmes, S., et al. (2002), Comparison of five Eulerian air pollution forecasting systems for the summer of 1999 using the German ozone monitoring data, *J. Atmos. Chem.*, *42*, 91–121.
U.S. Environmental Protection Agency (1989), The 1985 NAPAP emissions inventory (version 2), Development of the annual data and modelers' tapes, *Rep. EPA-600/7-89-012a*, 692 pp., Natl. Tech. Inf. Serv., Springfield, Va.
U.S. Environmental Protection Agency (1998), National air pollutant emission trends, procedures document, 1900–1996, *Rep. EPA-454/R-98–008*, 148 pp., Off. of Air Qual. Plann. and Stand., Research Triangle Park, N. C.
Vaughan, J., et al. (2004), A numerical daily air quality forecast system for the Pacific Northwest, *Bull. Am Meteorol. Soc.*, *85*, 549–562.
Vukovich, J., and T. Pierce (2002), Implementation of BEIS3 within the SMOKE modeling framework, paper presented at EPA Emissions Inventory Conference, U.S. Environ. Prot. Agency, Atlanta, Ga., 16 April.

V. Bouchet, S. Menard, and R. Moffet, Meteorological Service of Canada, Canadian Meteorological Centre, 2121 Trans-Canada Highway, 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.
A. Chan and T. Dye, Sonoma Technology, Inc., 1360 Redwood Way, Suite C Petaluma, CA 94954, USA.
I. Djalalova and J. Wilczak, Environmental Technology Laboratory, NOAA, 325 Broadway R/ETL, Boulder, CO 80305-3328, USA.
G. Frost, E.-Y. Hsie, and S. McKeen, Aeronomy Laboratory, NOAA, 325 Broadway R/E/AL4, Boulder, CO 80305-3328, USA. (stu@al.noaa.gov)
W. Gong, Meteorological Service of Canada, 4905 Dufferin Street, Downsview, ON, Canada M3H 5T4.
G. Grell, M. Pagowski, and S. Peckham, Forecast Systems Laboratory, NOAA, 325 Broadway, Boulder, CO 80305-3328, USA.
P. Lee and J. McQueen, NCEP Environmental Modeling Center, 5200 Auth Road, Camp Springs, MD 20746-4304, USA.
R. Mathur, National Exposure Research Laboratory, EPA, 109 T. W. Alexander Drive, Research Triangle Park, NC 27709, USA.
J. McHenry, North Carolina Supercomputing Center, 3021 Cornwallis Road, Research Triangle Park, NC 27709, USA.