Science of the Total Environment 473–474 (2014) 275–285

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Evaluation of a seven-year air quality simulation using the Weather Research and Forecasting (WRF)/Community Multiscale Air Quality (CMAQ) models in the eastern United States

Hongliang Zhang [a,1], Gang Chen [a], Jianlin Hu [b], Shu-Hua Chen [c], Christine Wiedinmyer [d], Michael Kleeman [b], Qi Ying [a,*]

[a] Zachry Department of Civil Engineering, Texas A&M University, College Station, TX 77843, USA
[b] Department of Civil & Environmental Engineering, University of California, Davis, CA 95616, USA
[c] Department of Land, Air, and Water Resources, University of California, Davis, CA 95616, USA
[d] Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO, USA

HIGHLIGHTS

- A 7-year WRF/CMAQ simulation for the eastern US with 4-km spatial resolution
- WRF model performance is comparable with other short-term studies.
- Ozone, $PM_{2.5}/PM_{10}$, elemental carbon, sulfate and nitrate are well reproduced.
- CMAQ (v4.7.1 with AERO5) still under-predicts organic carbon in the summer.

ARTICLE INFO

Article history:
Received 21 March 2013
Received in revised form 20 November 2013
Accepted 20 November 2013
Available online 27 December 2013

Keywords:
Model performance
Meteorology modeling
Air quality modeling
Eastern United States

ABSTRACT

The performance of the Weather Research and Forecasting (WRF)/Community Multi-scale Air Quality (CMAQ) system in the eastern United States is analyzed based on results from a seven-year modeling study with a 4-km spatial resolution.

For 2-m temperature, the monthly averaged mean bias (MB) and gross error (GE) values are generally within the recommended performance criteria, although temperature is over-predicted with MB values up to 2 K. Water vapor at 2-m is well-predicted but significant biases (>2 g kg^{-1}) were observed in wintertime. Predictions for wind speed are satisfactory but biased towards over-prediction with $0 < MB < 0.5$ m s^{-1} and root mean square error (RMSE) around 1.5 to 2 m s^{-1}. Wind direction, predicted without observation nudging, is not well-reproduced with GE values as large as 50° in summertime. Performance in other months is better with RMSE around 20–30° and MB within ±10°.

O_3 performance meets the EPA criteria of mean normalized bias (MNB) within ±0.15 and accuracy of unpaired peak (AUP) within 0.2. Normalized gross error (NGE) is mostly below 0.25, lower than the criteria of 0.35. Performance of PM_{10} is satisfactory with mean fractional bias (MFB) within ±0.6, but a large under-prediction in springtime was frequently observed. Performance of $PM_{2.5}$ and its components is mostly within performance goals except for organic carbon (OC), which is universally under-predicted with MFB values as large as −0.8. The predicted frequency distribution of $PM_{2.5}$ generally agrees with observations although the predictions are slightly biased towards more frequent high concentrations in most areas. Elemental carbon (EC), nitrate and sulfate concentrations are also well reproduced. The other unresolved $PM_{2.5}$ components (OTHER) are significantly overestimated by more than a factor of two. No conclusive explanations can be made regarding the possible cause of this universal overestimation, which warrants a follow-up study to better understand this problem.

1. Introduction

Ozone (O_3) and particulate matter (PM) have been shown to have adverse effects on human health (Lippmann, 1991; Poschl, 2005). However, spatial heterogeneity in individual or population exposure to air pollution cannot be correctly represented based on monitoring data alone (Bell, 2006). Furthermore, air monitoring data is limited by

* Corresponding author. Tel.: +1 979 845 9709; fax: +1 979 862 1542.
E-mail address: qying@civil.tamu.edu (Q. Ying).

[1] Current address: Department of Civil and Environmental Engineering, University of California, Davis, USA.

0048-9697/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.scitotenv.2013.11.121

analytical methods and may not provide sufficient temporal resolution or detailed chemical composition information to support detailed health outcome analysis.

State-of-the-art 3D chemical transport models (CTMs) can provide detailed gaseous and PM concentrations and their source and chemical composition information at one-hour resolution over large areas, which naturally fill in the spatial and temporal gaps in the exposure estimation solely based on air quality measurements at fixed monitors (Bell, 2006; Bravo et al., 2012). Among the many CTMs, the Community Multiscale Air Quality (CMAQ) model (Byun and Schere, 2006) is one of the most widely used regional air quality modeling systems in recent years (Simon et al., 2012) to evaluate pollution control measures, test new atmospheric mechanisms and processes that control air pollution, and determine source contributions to air pollutants. The CMAQ model has also been used to provide air pollution exposure estimations in a few recent studies (Arunachalam et al., 2011; Chang et al., 2012; Grabow et al., 2012; Tong et al., 2009).

The accuracy of the exposure estimation using modeled air quality data is affected by the ability of the model to reproduce observed air pollutant concentrations. Previous studies have shown that although regional model predictions of O_3 are not significantly improved by using a higher spatial resolution above 12 km (Simon et al., 2012), higher spatial resolution models better capture the peak concentrations of PM (Garcia-Menendez et al., 2010; Stroud et al., 2011). Model performance has also been reported to vary between short term and long term averages. Bravo et al. (2012) reported large seasonal variations in the CMAQ model performance of O_3 and $PM_{2.5}$ in the eastern United States (US) in 2002 although the annual model performance is satisfactory. Hogrefe et al. (2008) reported that the CMAQ model captures the long-term temporal variations of air pollutants better than their spatial variations over the northeastern US during 1988–2000.

In addition to the effects of spatial and temporal resolutions, the quality of the meteorological parameters such as ambient temperature, water vapor content, boundary layer height, and wind speed and direction also has impacts on the downstream prediction of air pollutant concentrations and needs to be carefully evaluated. While the meteorology performance of episodic simulations has been widely reported in the literature, long-term meteorological simulations and their model performance statistics are rarely reported.

In this study, a 7-year (2000–2006) air quality simulation is conducted over eastern US using the CMAQ model to provide exposure estimation databases of criteria gas pollutants, $PM_{2.5}$, PM_{10} and $PM_{2.5}$ chemical components for participants from seven cities in two longitudinal cohort studies. Spatial resolutions of the CMAQ simulations are as high as 4 km in order to better represent different exposure levels of participants. Although long term high resolution simulations have been reported for primary PM in California (Hu et al., under review-a; under review-b), to the author's knowledge, this is the first long term application of a CTM that covers eastern US with such a fine resolution. This paper focuses on the evaluation of model performance, which will provide a foundation for companion papers that estimate the O_3 and PM exposure and associate air pollution with human health outcomes.

2. Model application

2.1. Model settings

The chemical transport model used in this study is CMAQ version 4.7.1. The standard CMAQ model with the lumped, non-toxic version of the SAPRC99 photochemical mechanism (Carter, 2000) and the fifth generation aerosol module (AERO5) (Foley et al., 2010) is used in this work. Air quality simulations are performed for a period of 7 years from 2000 to 2006 using three-level nested domains with horizontal resolutions of 36-km, 12-km and 4-km, respectively. All three-level domains use the same vertical domain set up with 16 vertical layers that reach approximately 20 km above ground. The first layer thickness is approximately 30 m. The 36-km (160 × 124 grid cells) domain covers the entire continental US and the 12-km domain covers the eastern US (159 × 111). Four 4-km domains cover seven cities selected from two longitudinal cohort studies: New York City, NY (NYC), Pittsburgh, PA (PIT), Baltimore, MD (BAL), Chicago, IL (CHI), Detroit, MI (DET), St. Paul, MN (S-P), and Winston-Salem, NC (W-S). The 4 km NPB domain (185 × 119) includes NYC, PIT, and BAL and the 4 km CD domain (159 × 94) includes CHI and DET. The SP and WS domains (both are 50 × 50) include S-P and W-S, respectively. Five of the seven cities are from the Multi-Ethnic Study of Atherosclerosis (MESA) and two are from the Women's Health Initiative Observational Study (WHI-OS). Locations of the domains and the cities of interest are shown in Fig. 1.

2.2. Meteorological inputs

The meteorological inputs are generated by the Weather Research and Forecasting (WRF) model v3.2.1 with 29 vertical layers. The initial and boundary conditions for the WRF simulations are prepared using the 1° × 1° resolution NCEP (National Centers for Environmental Prediction) FNL (Final) Operational Global Analysis dataset (available at http://rda.ucar.edu/datasets/ds083.2/). The land use/land cover and topographical data are from the default WRF input dataset. Two-way nested runs are conducted to generate meteorological fields for the 36-km, 12-km, and 4-km domains simultaneously. The WRF domains are much larger than their CMAQ counterparts to ensure minimum influence from the boundary conditions to the air quality simulation domains. The WRF simulations for each year are divided into multiple runs. Each WRF run simulates 7 days with fresh initial conditions based on the NARR (NCEP North American Regional Reanalysis) reanalysis processed by WRF Preprocessing System (WPS). The first day of each run, which overlaps the last day of the previous run, is considered as a spin-up day and is discarded to avoid the influence of initial conditions on model results. The physics options used to drive the WRF simulations are listed in Table S1.

2.3. Emissions and other inputs

The US Environmental Protection Agency (US EPA) 2001 Clean Air Interstate Rule (CAIR) emission inventory is used to generate anthropogenic emissions from area, non-road, mobile and point sources for 2000 to 2004. 2000 is the base year, and emissions of year 2001 to 2004 are adjusted based on the average annual emissions treads of criteria pollutants (download from http://www.epa.gov/ttn/chief/trends/). The National Emissions Inventory (NEI) in 2005 is used to generate anthropogenic emissions for 2005 and 2006. The Sparse Matrix Operator

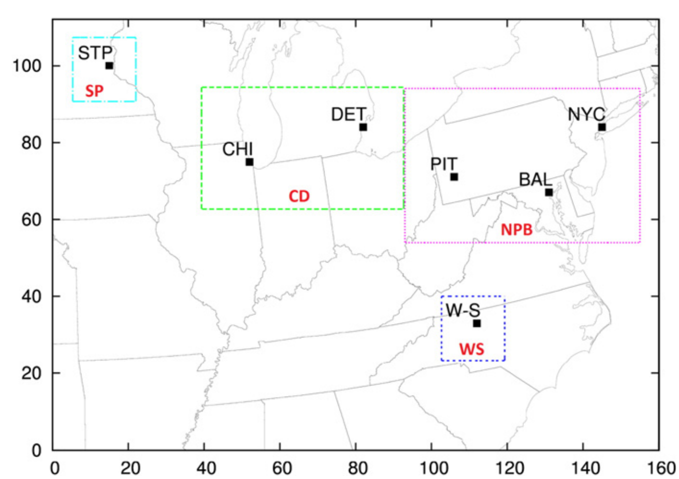

Fig. 1. The 12-km and 4-km air quality modeling domains used in this study. The 36-km domain, covering the entire continental US and part of Canada and Mexico, is not shown here. Numbers on the axes are grid cell index of the 12-km eastern US domain.

Kernel Emissions (SMOKE) emission processing model (version 2.7) from US EPA is used to process the raw emission inventory to generate emissions of gases and PM. Table 1 shows the annual emissions of major pollutants of 4-km domains for base case years 2000 and 2005.

Biogenic emissions are generated using the Biogenic Emissions Inventory System, v3.14 (BEIS3.14) incorporated in SMOKE, which includes a 1-km resolution land cover database with 230 different cover types (Vukovich and Pierce, 2002). Open biomass burning emissions for years other than 2000 and 2001 are generated using a home-made tool based on the Fire INventory from NCAR (FINN), an inventory converted from satellite observation (Wiedinmyer et al., 2011). Open burning emissions for 2000 and 2001 are generated based on the annual fire emission inventory from NEI. The FINN inventory provides SAPRC99 speciated daily emissions of gaseous and particulate emissions (elemental carbon (EC), organic carbon (OC), $PM_{2.5}$ and PM_{10}) based on satellite observations of open burning events. Each open burning event is allocated to model grid cells of each domain based on the reported longitude/latitude of the event and the area burned. The emissions are evenly distributed vertically to the layers under the planetary boundary layer (PBL). A temporal variation profile for open burning from a Western Regional Air Partnership (WRAP) report (WRAP, 2005) is used to distribute emissions to each hour of the day. Emissions from Canada and Mexico are also generated based on inventories provided by US EPA for 2000 and are not adjusted for different years.

Initial conditions for the first day are generated using CMAQ program based on default values. The last hour results of previous day provide initial conditions for the next day. Boundary conditions for the 36-km domain are generated using CMAQ default program as well, while boundary conditions for finer domains are generated from coarser domains. For all subsequent analyses, only simulations from the 4-km domains are utilized.

3. Model performance evaluation

3.1. Meteorological variables

Mean bias (MB; see Table 2 for definition of MB and other statistical measures), gross error (GE) and root mean square error (RMSE) are commonly used for evaluation of meteorology variables. Based on various simulations of PSU/NCAR mesoscale model (MM5) in Texas, Emery et al. (2001) proposed benchmarks for temperature (MB within ±0.5 K and GE < 2.0 K), water content (water vapor mixing ratio, MB within ±1.0 g/kg and GE < 2.0 g/kg), wind speed (MB within ±0.5 m s^{-1} and RMSE < 2 m s^{-1}) and wind direction (MB within ±10° and GE < 30°). McNally (2009) suggested an alternative set of benchmarks for temperature (MB within ±1.0 K and GE < 3.0 K) and identical performance criteria for water vapor. However, it is concluded that model performance of wind speed/direction varies significantly among regions with different terrain conditions and thus a common model performance criteria might not be appropriate. Although these proposed benchmarks are based on MM5 simulations, a recent study that compares WRF and MM5 over North America showed that WRF performance is similar to that of MM5 (Gilliam and Pleim, 2009). Thus, these benchmarks were actually used to evaluate WRF model performance in a number of recent studies (Fast et al., 2006; Misenis and Zhang, 2010; Wang et al., 2010; Zhang et al., 2012). It should be noted that due to the limited temporal and spatial coverage of the studies, the proposed benchmarks are intended to be guidelines rather than definitive tests of passing or failing of a particular meteorological model application.

Table. 1
Emissions of CO, SO_2, NO_x, VOC and $PM_{2.5}$ for 4 km domains at 2000 and 2005. Units are thousand tons $year^{-1}$.

		CO	SO_2	NO_x	VOC	$PM_{2.5}$
NPB	2000	13616.93	3305.53	2439.53	1780.65	390.40
	2005	11332.49	2964.61	2112.45	1798.48	354.81
CD	2000	9288.77	1030.19	1604.39	1412.72	291.35
	2005	7698.96	946.03	1350.57	1259.81	219.66
SP	2000	1284.71	88.99	242.06	196.20	54.57
	2005	1132.23	93.62	254.46	174.48	40.32
WS	2000	1263.62	323.26	250.53	223.09	40.41
	2005	986.63	359.14	192.96	180.13	34.80

Table 2
Definition of statistics metrics used in this study.

Metrics	Definition	
Mean bias	$MB = \frac{1}{N}\sum_{i=1}^{N}(C_{m,i}-C_{o,i})$	(1)
Gross error	$GE = \frac{1}{N}\sum_{i=1}^{N}\lvert C_{m,i}-C_{o,i}\rvert$	(2)
Root mean square error	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_{m,i}-C_{o,i})^2}$	(3)
Mean normalized bias	$MNB = \frac{1}{N}\sum_{i=1}^{N}\frac{C_{m,i}-C_{o,i}}{C_{o,i}}$	(4)
Normalized gross error	$NGE = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert C_{m,i}-C_{o,i}\rvert}{C_{o,i}}$	(5)
Unpaired predict-to-observed peak ozone ratio	$AUP = \frac{C_{p,ppeak}-C_{o,opeak}}{C_{o,opeak}}$	(6)
Mean fractional bias	$MFB = \frac{2}{N}\sum_{i=1}^{N}\frac{C_{m,i}-C_{o,i}}{C_{m,i}+C_{o,i}}$	(7)
Mean fractional error	$MFE = \frac{2}{N}\sum_{i=1}^{N}\frac{\lvert C_{m,i}-C_{o,i}\rvert}{C_{m,i}+C_{o,i}}$	(8)

Note: C_m is the model-predicted concentration i, C_o is the observed i, and N equals the number of prediction–observation pairs drawn from all monitoring stations. The subscripts *ppeak* and *opeak* are the hours when predicted and observed peak concentrations occur.

In this study, surface weather observations produced by National Climatic Data Center (NCDC) (available at http://rda.ucar.edu/datasets/ds463.3/) are used to compare with the model predictions. The statistics of WRF results are compared with the suggested benchmarks. The so-called "soccer field" plots are used to illustrate the meteorology model performance. The goal box in the plot is determined by the performance criteria for bias and error measures, which allows a clear demonstration whether a data point meets the bias and error criteria simultaneously. The ultimate model performance goal of zero bias and error is appropriately located at the middle of the goal-line. For domains NPB and CD, approximately 70–200 stations contain valid data for each domain. For domains SP and WS, only 5–15 stations are available for each domain. The exact number of stations used in the monthly performance analysis varies depending on data availability.

Fig. 2 shows the WRF performance of 2-m temperature (T2) for all the months from 2000 to 2006. As shown in Fig. 2(a), the MB of T2 for the NPB domain is within −1.0 to 1.5 K. Under-predictions of T2, as represented by negative MB values, occur much less frequently than over-predictions and a number of the under-predictions are for summer months in 2005. The GE values are narrowly distributed within 1.5 and 2.5 K with slightly higher values that occur in late spring and early summer months. Overall, for the NPB domain, most of the points are within the suggested model performance criteria with only a few exceptions. Model performance of T2 for the CD domain is similar to that of the NPB domain although there are a few more data points with MB exceeding the suggested criteria. For the two smaller domains, SP and WS, although the GE values are still within the suggested range, a significant number of data points exhibit larger MB values than the recommend criteria of 1.0 K. However, since there are only a limited number of observation stations within each domain, the performance statistics might not represent the actual domain-wide model performance.

Fig. 3 summarizes the WRF model performance of 2-m water vapor content (Q2) for the 4-km domains. For the NPB domain, approximately 20% of the data points, mostly winter months, do not meet the proposed criteria (MB within ±1.0 g/kg and GE < 2.0 g/kg). All of these data points show significant over-predictions with large MB values. For the CD domain, there are also approximately 20% of the data points

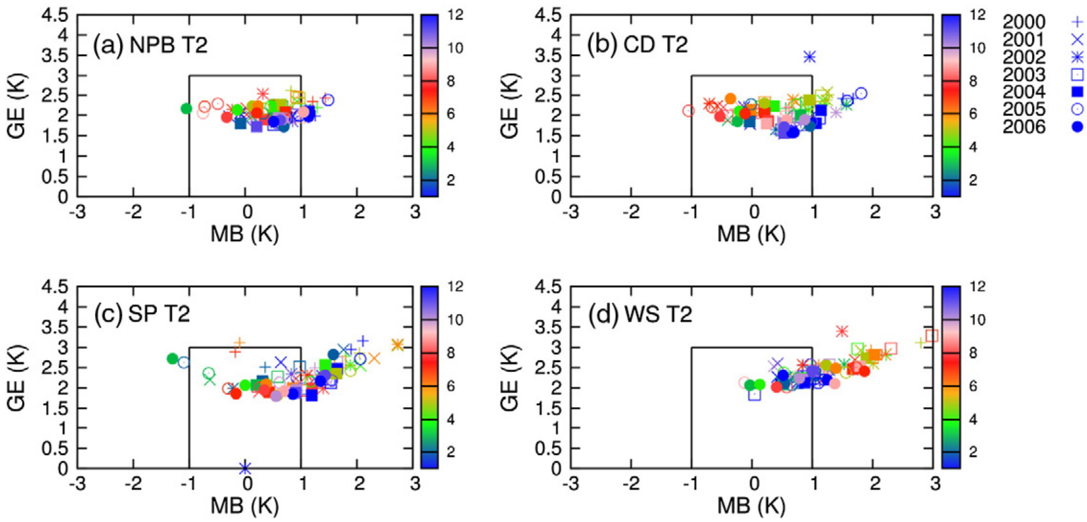

Fig. 2. WRF performance of 2-m temperature (T2) for all the months from 2000 to 2006 for 4-km domains: (a) NPB, (b) CD, (c) SP, and (d) WS. Different point types represent model performance criteria for different years. The points are color-coded by the month of the year. The box in each plot shows the range of the gross error (GE) and mean bias (MB) benchmark of McNally (2009).

exceeding the MB performance criteria. The relatively poor performance of water vapor content for the two domains contrasts that of T2, which exhibits excellent model performance. For the SP domain, most of the exceedances still occur in winter months but model predictions are lower than observations most of the time. Q2 model performance in the WS domain is slightly worse than other three domains with approximately 30% of the data points located out of the model performance region. The poor performance is mostly due to under-predictions during the winter months, though a few data points for summer months also show large MB values.

Fig. 4 illustrates the model performance on 10-m wind speed (WSPD). Fig. 4(a) shows a clear bias of over-predicting WSPD. For majority of the data points, the MB values are between 0 and 0.5 m s^{-1}. Although approximately half of the data points have RMSE values greater than the criteria of 2 m s^{-1}, they are mostly below 2.5 m s^{-1}. Similar behavior is observed for the CD domain in Fig. 4(b) but more values are within the goal box. For the SP domain as shown in Fig. 4(c), MB values are generally below zero, indicating a general under-prediction of WSPD. The data points that have MB values less than -0.5 m s^{-1} are usually for winter and spring months. Majority of RMSE values are within benchmark, even for the data points that are out of the goal box. Fig. 4(d) shows that WSPD values in the WS domain are generally overestimated. Most of the data points with MB values greater than 0.5 are for spring and summer months. With a few exceptions, the RMSE values are below the suggested criteria of 2 m s^{-1}. In summary, the WRF model appears to generally over-predict wind speed in three of the study domains and under-predict wind speed in the SP domain. The opposite trend in wind speed in SP is likely caused by the fact that the S-P domain is relatively small compare to the two large CD and NPB domains and only a few meteorology sites were included in the domain near Saint Paul/Minneapolis area. Influence of the Mississippi River and urban terrain may cause a change in wind speed and wind direction locally. The statistics might not represent the overall model performance in the SP domain.

Model performance of wind direction (WDIR) is shown in Fig. 5. When calculating MB and GE values, a wind speed cutoff value of 0.5 m s^{-1} is applied to exclude data with low predicted wind speed. MB values for all the months are within the range except two months in domain NPB and CD. 48% of GE for domain NPB is larger than 30° with a maximum of approximately 60°. For domain CD, only one point of MB is out of the benchmark criteria but 36% of the GEs are larger than 30° and the maximum is slight higher than 40°. For SP in

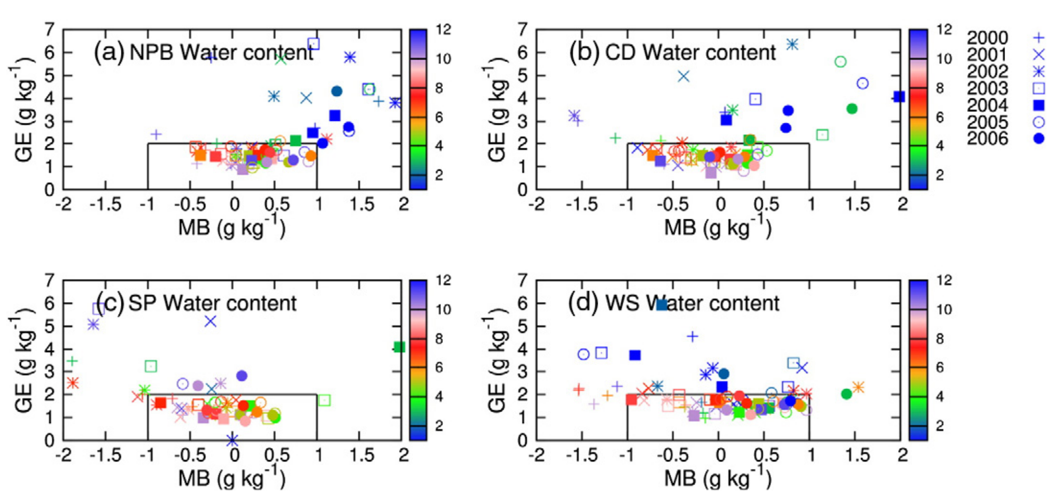

Fig. 3. WRF performance of 2-meter water vapor content (Q2) for all the months from 2000 to 2006 for 4-km domains: (a) NPB, (b) CD, (c) SP, and (d) WS. Different point types represent model performance criteria for different years. The points are color-coded by the month of the year. The box in each plot shows the range of the GE and MB benchmark of Emery et al. (2001).

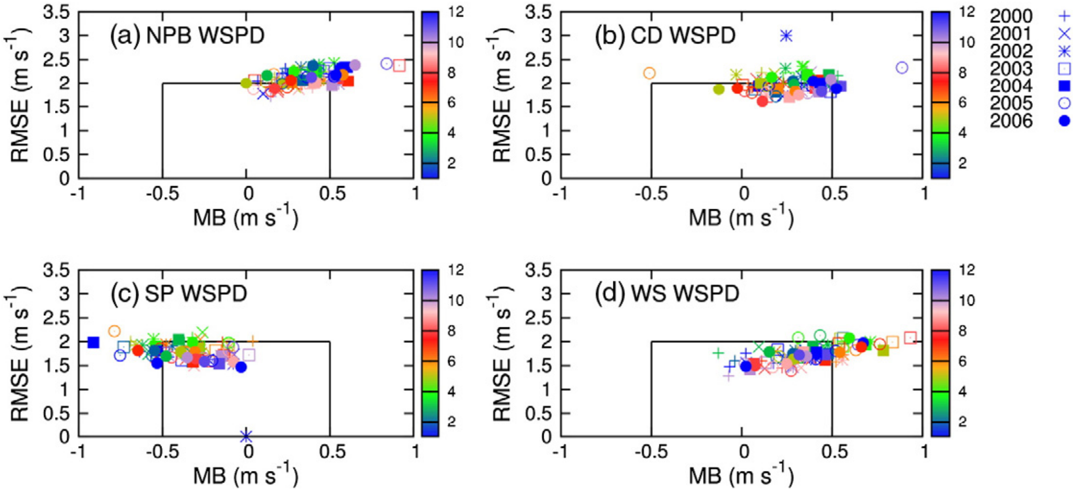

Fig. 4. WRF performance of wind speed (WSPD) for all the months from 2000 to 2006 for the 4-km domains: (a) NPB, (b) CD, (c) SP, and (d) WS. Different point types represent model performance criteria for different years. The points are color-coded by the month of the year. The box in each plot shows the range of the MB and RMSE benchmark of Emery et al. (2001).

Fig. 4(c), the MBs are exclusively larger than 0 showing over-prediction. RMSE values are all within or close to benchmark. For WS, 6 data points are out of MB benchmark range. The RMSE values are all close or higher than 30° with a maximum of up to 60°. Most months with large RMSE are summer months when wind speed is lower. The higher-than-benchmark RMSE values for wind direction and their seasonal variations are consistent with another study for western US in spring and summer 2008 (Morris and Tai, 2012).

Generally, although the WRF performance for temperature, water content, wind speed, and wind direction does not meet strict benchmarks for all the months in the seven year simulation, it is comparable with other studies, and reflects WRF's current ability in reproducing the observed meteorological conditions.

3.2. Gas phase species

Observed hourly concentrations of O_3, CO, SO_2 and NO_x are retrieved from the Air Quality System (AQS, http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm) maintained by the US EPA. US EPA recommended three statistical measures, mean normalized bias (MNB), normalized gross error (NGE) and unpaired predict-to-observed peak O_3 ratio (AUP), to evaluate model performance for O_3. The criteria are $\pm 15\%$ for MNB, $<35\%$ for NGE and $\pm 20\%$ for AUP with a threshold O_3 value of 60 ppb, based on the EPA recommendation (U.S.EPA, 2007).

Fig. 6 shows the monthly MNB, NGE and AUP for the 4-km domains from April to October of each year. For the other months, there are not enough data points with O_3 concentrations higher than the threshold value to calculate meaningful statistics. For the NPB domain, MNB values are within the $\pm 15\%$ range except October 2002 and 2003. The MNB values are closest to zero in the summer months, with over-predictions of O_3 in summer from June to August and under-predictions in other months. NGE values for the NPB domain are all within the range of 35%. For AUP, no values are lower than -20% and only July and August of 2000, 2002 and 2003 and October of 2002 are slightly higher than 20%. These show that the O_3 performance is fairly good except that the model slightly over-predicts summer peak O_3. For the CD domain, MNB values of April, May, September, and October of a few years are less than the lower criteria while all other months are within the criteria. NGE and AUP values are all within the EPA criteria. For the SP domain, approximately half of the months fall out of the MNB criteria, indicating that O_3 concentrations are generally underestimated. A significant under-prediction is observed for June

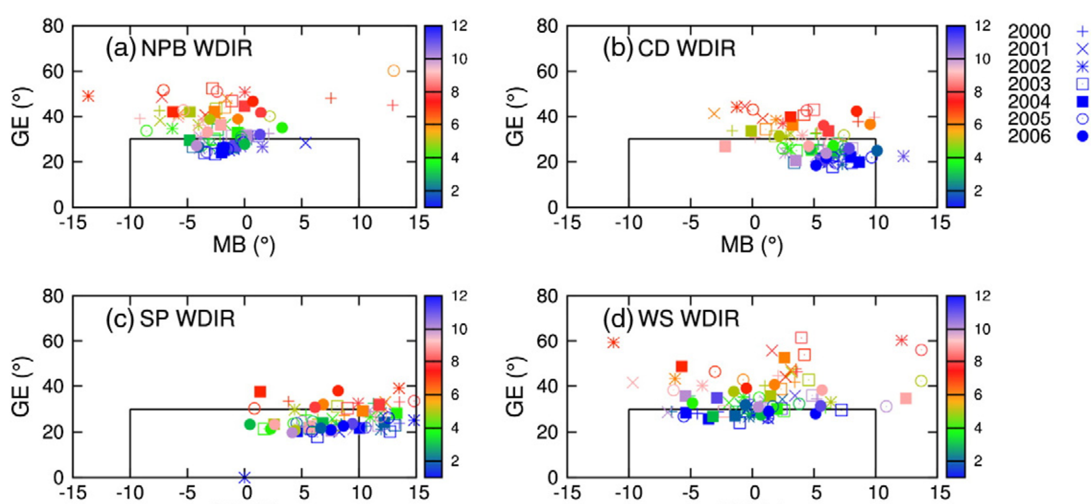

Fig. 5. WRF performance of wind direction (WDIR) for all the months from 2000 to 2006 for the 4-km domains: (a) NPB, (b) CD, (c) SP, and (d) WS. Different point types represent model performance criteria for different years. The points are color-coded by the month of the year. The box in each plot shows the range of the MB and GE benchmark of Emery et al. (2001).

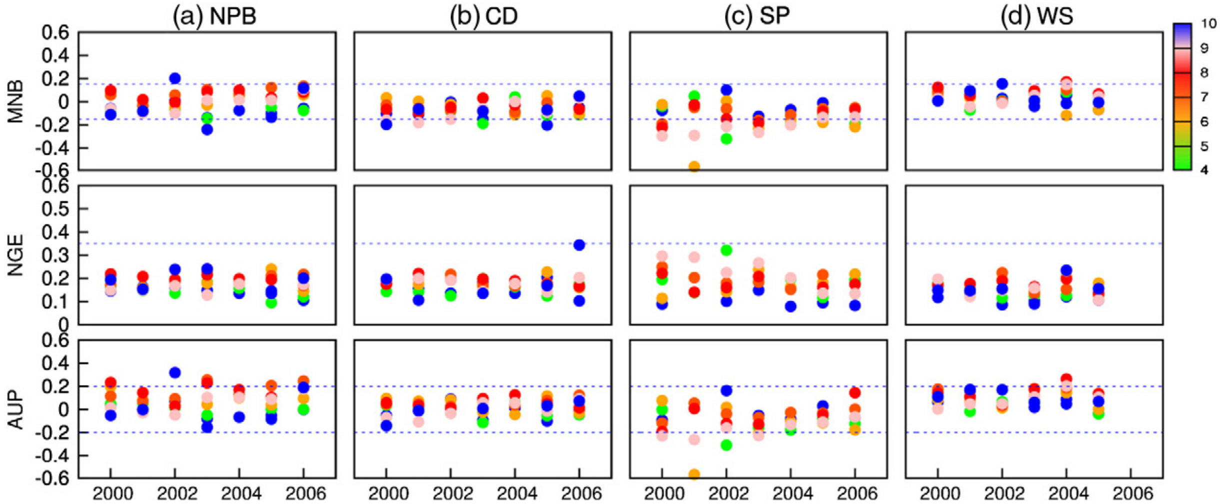

Fig. 6. Monthly mean normalized bias (MNB), normalized gross error (NGE), and unpaired predict-to-observed peak ozone ratio (AUP) for O_3 of 4-km domains NPB, CD, SP, and WS. Blue dashed lines show the criteria recommended by US EPA. Data points are color-coded by month.

2001. All months are within the criteria for NGE. Two months in 2001 and one month in 2002 do not meet the AUP criteria in the SP domain. Overall, under-prediction happens often in SP and mostly in 2000 to 2003. For the WS domain, all months are within the criteria of all three metrics except for one month in 2004 which does not meet the AUP criteria. No observation is available in the WS domain for 2006. In summary, O_3 performance is generally good and easily meets the EPA recommendations.

EPA does not provide performance criteria for CO, SO_2 and NO_x to evaluate regional air quality models. Model performance statistics for CO, SO_2 and NO_x are calculated for concentrations higher than threshold values of 100, 4 and 5 ppb, respectively. The MNB results for CO, NO_x, and SO_2 are shown in Fig. S1 in the Supplementary materials.

3.3. PM_{10} and $PM_{2.5}$ mass concentrations

Mean factional bias (MFB) and mean fractional error (MFE) are widely used in PM performance evaluation studies (Zhang and Ying, 2010). MFB and MFE are both bounded and symmetrical metrics so that usually no concentration threshold values are necessary for their calculations. Boylan and Russell (2006) proposed concentration dependent MFB and MFE performance goals and criteria, realizing that lower concentrations are more difficult to accurately predict. The performance goals are the level of accuracy close to the best model that can be expected to achieve, while performance criteria are the level of accuracy acceptable for standard modeling applications. Only MFB values are reported in the manuscript. MFE values are also calculated and included in the Supplementary materials (Figs. S2–S5). Observed 24-h averaged PM_{10} and $PM_{2.5}$ data used in these comparisons were downloaded from the US EPA's AQS.

Fig. 7 shows the monthly MFB values for PM_{10} in all 4-km domains. For the NPB domain (Fig. 7(a)), all the statistical metrics are within the suggested criteria and more than half are within the proposed goal, indicating that the domain wide CMAQ performance of PM_{10} is very good. PM_{10} performance of the CD domain (Fig. 7(b)) is similar to that of the NPB domain. Fig. 7(c) and (d) show PM_{10} performance at domain SP

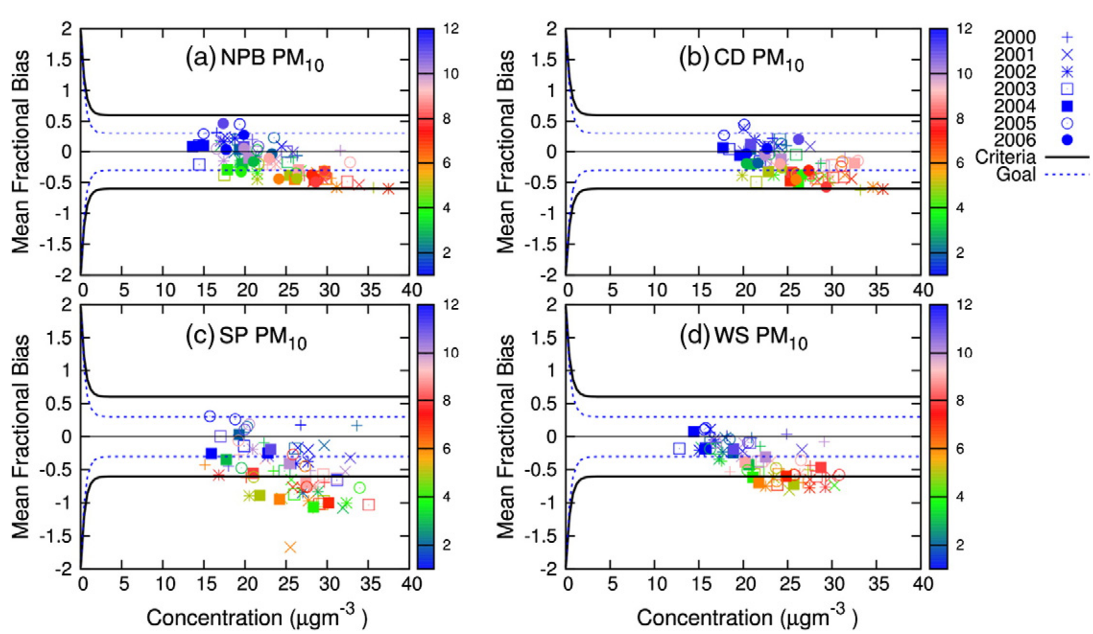

Fig. 7. Mean fractional bias (MFB) of monthly averaged PM_{10} for the 4-km domains: (a) NPB, (b) CD, (c) SP, and (d) WS along with the performance goals and criteria suggested by Boylan and Russell (2006). Data points are color-coded by month.

and WS. These two domains cover much smaller areas compared to the NPB and CD domains and only a few observation sites are in each domain. Several data points, mostly representing spring months, do not meet the performance criteria due to significant under-predictions. Overall, the model predictions for PM_{10} are generally good for most seasons. The simulations tend to under-predict high PM_{10} concentrations in spring, which is likely due to under-estimation of windblown dust or fugitive dust from agriculture activities such as soil tillage. As the current standard CMAQ modeling application does not treat agriculture emissions mechanistically using actual meteorological inputs and does not include windblown dust emissions, under-prediction of PM_{10} in spring time is expected. Under-predictions in the summer months are likely due to a combination of under-prediction of SOA and dust contributions. Over-prediction in other months is likely due to the overestimate of fugitive dust by the annual emission inventory (Pace, 2005b).

Fig. 8 shows the monthly MFB values of $PM_{2.5}$ for all 4-km domains. Generally, model predictions agree well with observation. MFB values of all months for all domains are within the criteria except two points for NPB (see Fig. 8(a)) and several points for SP (see Fig. 8(c)). Domain averaged $PM_{2.5}$ concentrations for NPB, CD, and WS domains are approximately 10–20 μg m^{-3} and 5–10 μg m^{-3} in the SP domain. The MFB values for the model indicate slight over-predictions of $PM_{2.5}$ during the wintertime while the MFB values for summertime are closer to 0. Under-predictions of $PM_{2.5}$ are obvious in the NPB and WS domains during summertime. The summertime under-predictions of $PM_{2.5}$ in the NPB domain appear to be related to under-predictions of the organic carbon components (as illustrated in Fig. 10).

Frequency plots have been used to assess long term performance of air quality simulations for O_3 (Winner and Cass, 1999). Fig. 9 shows the frequency distributions of predicted and observed 24-h averaged $PM_{2.5}$ in 4 domains used in the current study. All available observations and the matching model predictions within each 4-km domain are used to generate the frequency distribution plots. For NPB, the predicted frequency distributions generally agree well with the observations although they tend to slightly under-predict the occurrence of lower concentration events and over-predict the occurrence of higher concentration events. The predictions also show longer tails in the frequency distributions. This suggests that using the predicted concentrations likely leads to an overestimation of the yearly cumulative population ambient $PM_{2.5}$ exposure in this domain. The observed frequency distributions show an evident trend shifting towards lower concentrations between 2000 and 2006. Although model predictions show similar left-shifting trends from 2000 to 2004, the predicted peak in the concentration distribution increases from slightly below 10 μg m^{-3} to about 12 μg m^{-3} between 2005 and 2006, leading to a significant bias towards higher concentrations. As the meteorology model performance does not vary much between 2000 and 2006, the reduction in the emissions in 2005 might not be represented sufficiently by the 2005 NEI in this region. The long term frequency distributions for CD, as shown in the second column of Fig. 9, are similar to those for NPB. The modeled frequency distributions tend to have longer tails to higher concentrations and under-predict the frequency of lower concentrations. For SP (third column of Fig. 9), the predicted frequency distributions agree with the observations from 2000 to 2004. For 2005 and 2006, the predicted peak in the concentration distribution increases to approximately 15 μg m^{-3} while the measured peak decreases to approximately 5 μg m^{-3}. For the WS domain, predicted frequency distributions show excellent agreement with observations for all the years.

3.4. Major $PM_{2.5}$ chemical component concentrations

While $PM_{2.5}$ mass concentration has been reported to associate with adverse health outcomes, it has been hypothesized that the actual mechanism of injury might be related with PM chemical components (such as EC or trace metals). Thus it is essential to evaluate the major PM chemical components before they are used in epidemiology studies to associate exposure with adverse health outcomes. EC, OC, sulfate (SO_4^{2-}) and nitrate (NO_3^-) are compared to measurements compiled in US EPA's AQS system in the current study. AQS includes data from the Chemical Speciation Network (CSN) and the Interagency Monitoring of Protected Visual Environments network (IMPROVE), both providing 24-h average speciated $PM_{2.5}$ every 1, 3 or 6 days. The CSN and IMPROVE sites are mostly located in urban and national park/rural areas, respectively (Simon et al., 2012). Blank correction was applied to the measured concentrations. Although some previous studies adjust the EC and OC measurements (e.g. Held et al., 2004; Ying et al., 2008) to account for the differences in the observed EC and OC concentrations determined by the IMPROVE protocol (used by the IMPROVE sites) and the predicted EC and OC concentrations based on source profiles derived using the NIOSH protocol, this adjustment is not applied in

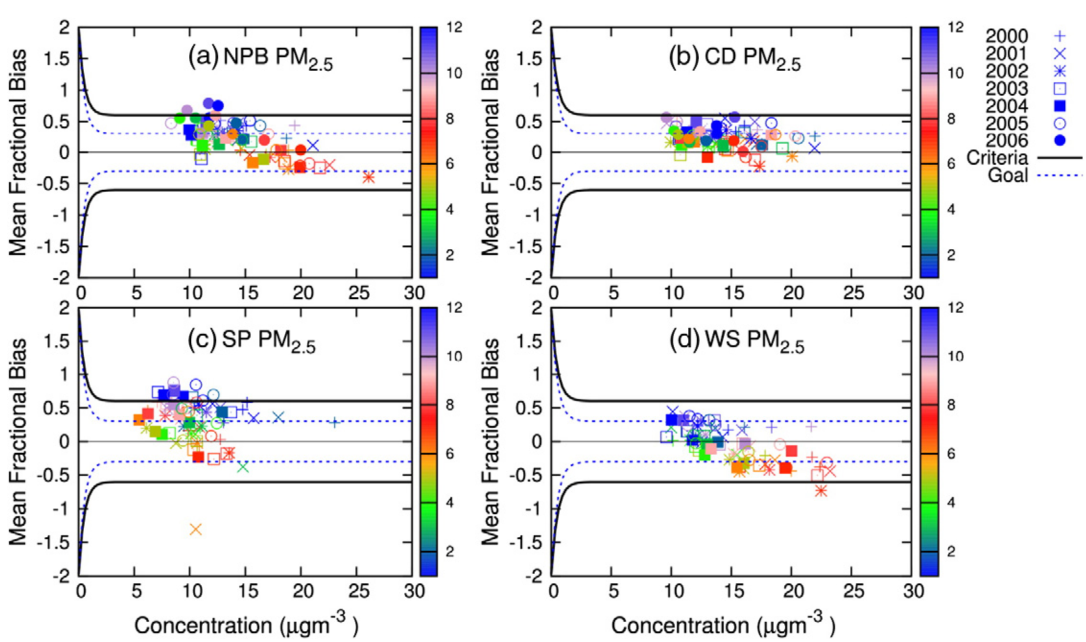

Fig. 8. Mean fractional bias (MFB) of monthly averaged $PM_{2.5}$ for the 4-km domains: (a) NPB, (b) CD, (c) SP, and (d) WS along with the performance goals and criteria suggested by Boylan and Russell (2006). Data points are color-coded by month.

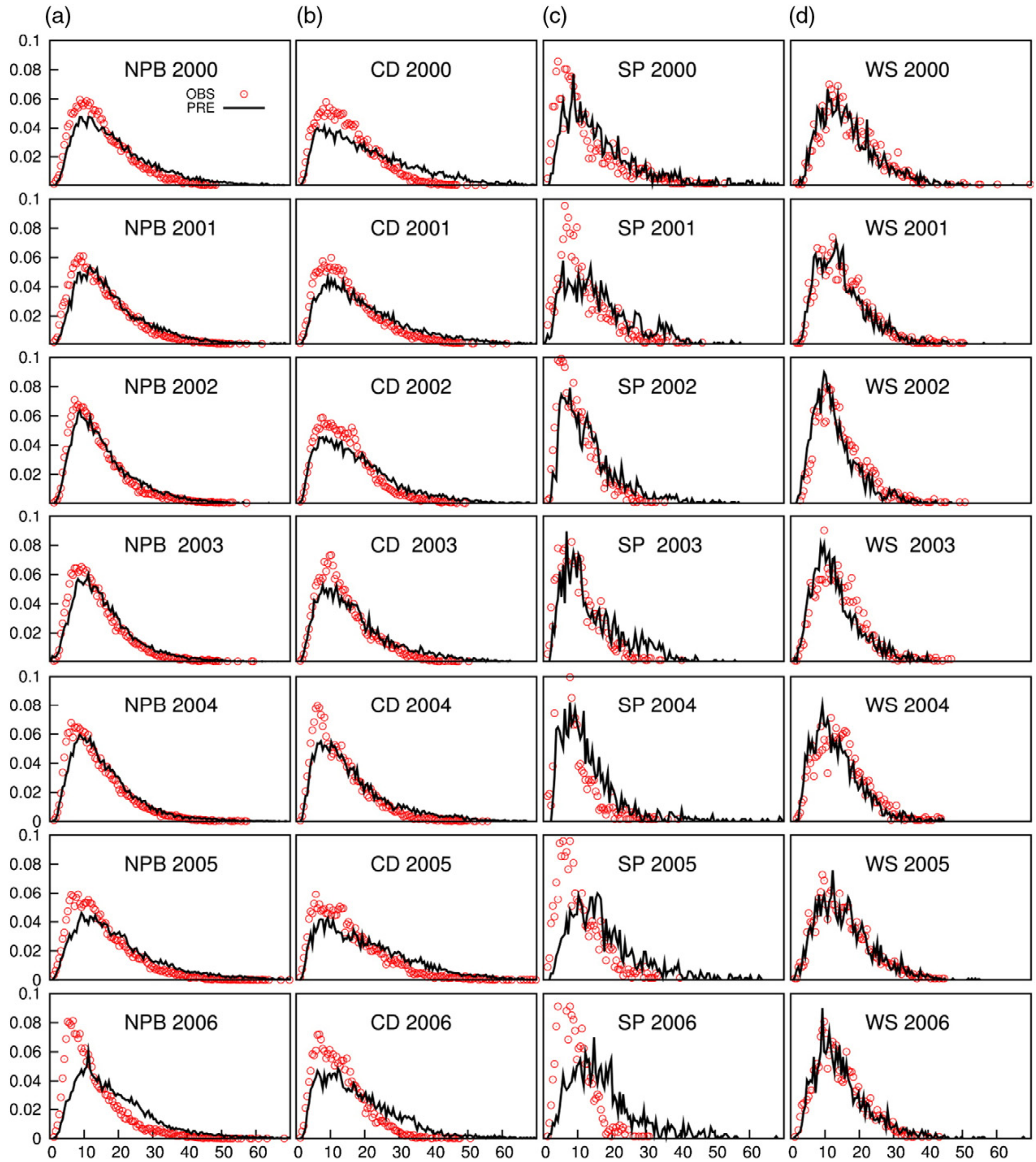

Fig. 9. Frequency distribution of predicted and observed daily $PM_{2.5}$ as a function $PM_{2.5}$ mass concentration for each domain: (a) NPB, (b) CD, (c) SP, and (d) WS. X-axis is the concentration with unit of $\mu g\ m^{-3}$.

this study as there are more CSN sites with many more observations and their EC/OC measurements are based on the NIOSH protocol before 2007 (Simon et al., 2012). Predicted organic mass (OM) concentrations are converted to OC using an OM/OC ratio of 1.4 (Turpin and Lim, 2001 and the references therein) to compare with observed OC concentrations.

Fig. 10 shows the monthly MFB values of 24-h average EC and OC for the NPB and CD domains. Model performance for the SP and WS domains is not shown due to lack of observations. As shown in Fig. 10(a) and (b), monthly averaged $PM_{2.5}$ EC values for both domains are in the range of 0.5 to 1 $\mu g\ m^{-3}$. All of the MFB values are within the model performance criteria and there is no obvious bias towards positive or negative MFB. This demonstrates that the model generally captures EC for all the years. Monthly average $PM_{2.5}$ OC concentrations are in the range of 2 to 5 $\mu g\ m^{-3}$. Although majority of the MFB values are within the model performance criteria (Fig. 10(c) and (d)), it is clear that OC is under-predicted by CMAQ with MFB values biased towards negative values. More significant underestimations, typically occurring in summertime, are likely associated with underestimation of secondary organic aerosol (SOA), as reported in many studies (for example, see Matsui et al. (2009), Volkamer et al. (2006), and Zhang and Ying (2011)).

Fig. 11(a) and (b) show that monthly averaged $PM_{2.5}$ sulfate concentrations are in the range of 2 to 9 $\mu g\ m^{-3}$. The majority of the MFB values fall in the proposed performance goals and only one data point in 2001 does not meet the model performance criteria. For the NPB domain, slight under-predictions of sulfate occur in summer months,

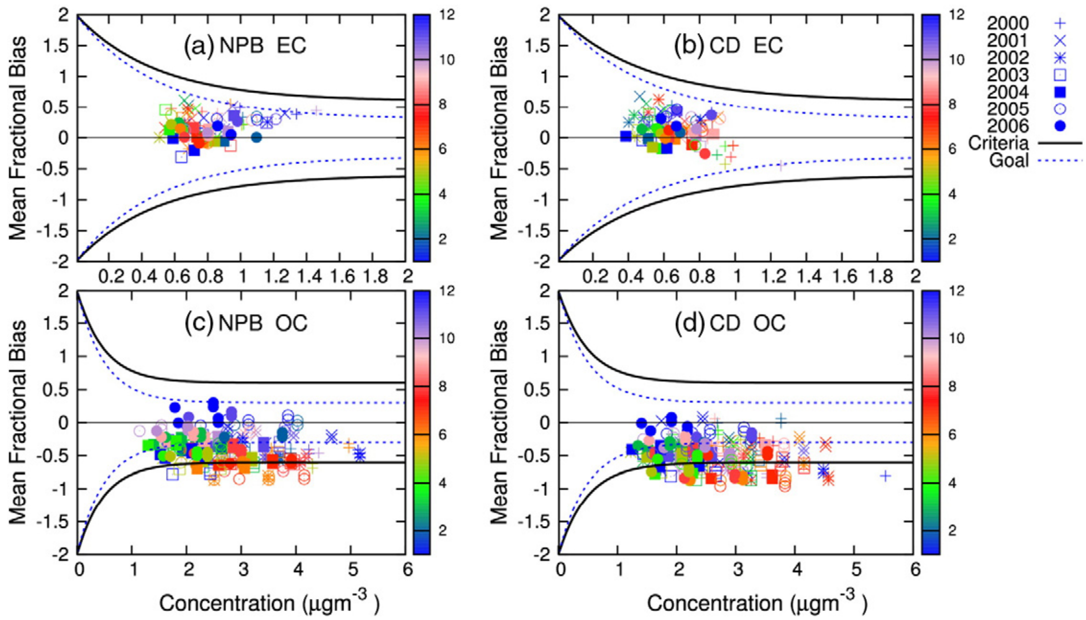

Fig. 10. Mean fractional bias (MFB) of monthly averaged $PM_{2.5}$ elemental carbon (EC) and organic carbon (OC) for the 4-km domains: (a) NPB, (b) CD, (c) SP, and (d) WS along with the performance goals and criteria suggested by Boylan and Russell (2006). Data points are color-coded by month.

especially when the observed sulfate concentrations are high. This is in agreement with another study which shows ubiquitous sulfate under-prediction in the continental US in summertime (Luo et al., 2011), which is attributed to the overestimation of wet scavenging by the CMAQ cloud module. As illustrated in Fig. 11(c) and (d), there are fewer points for $PM_{2.5}$ nitrate, especially for domain CD which only has valid data for 2003, 2004 and 2005. Most of the data points are within the performance criteria and majority of the data points are within the range of the performance goal. CMAQ slightly over-predicts the nitrate concentrations in wintertime and under-predicts the concentrations in summertime for NPB, although the observed concentrations are quite low throughout the year (less than 2 μg m^{-3} in wintertime and 1 μg m^{-3} in summertime). For the CD domain, nitrate concentrations are slightly higher. There are four data points that do not meet the model performance criteria, indicating under-prediction in those months.

Fig. 10 and Fig. 11 show model performance of individual $PM_{2.5}$ components. To further illustrate the ability of the model in $PM_{2.5}$ predictions, comparisons of observed and predicted 7-year average $PM_{2.5}$ components for NPB and CD based on all available measurements and the corresponding predictions are shown in Fig. 12. Due to lack of observations, domains SP and WS are not included in the analysis. The predictions are slightly higher in EC and ammonium ion concentrations, and lower in sulfate concentrations in the NPB domain. Predicted sulfate

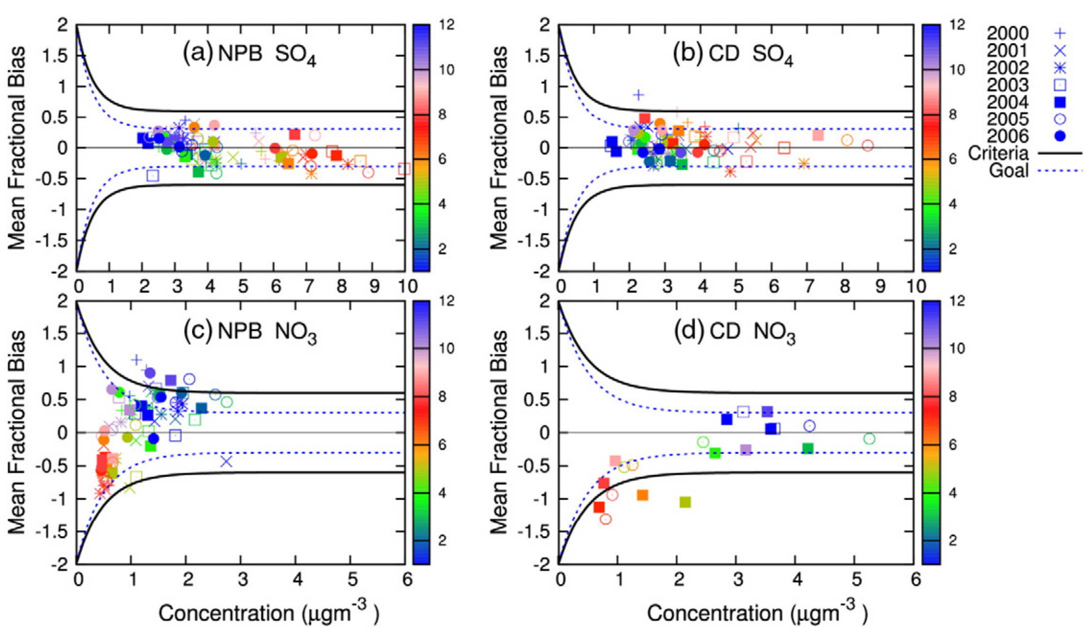

Fig. 11. Mean fractional bias (MFB) of monthly averaged $PM_{2.5}$ elemental carbon (EC) and organic compounds (OC) for the 4-km domains: (a) NPB, (b) CD, (c) SP, and (d) WS along with the performance goals and criteria suggested by Boylan and Russell (2006). Data points are color-coded by month.

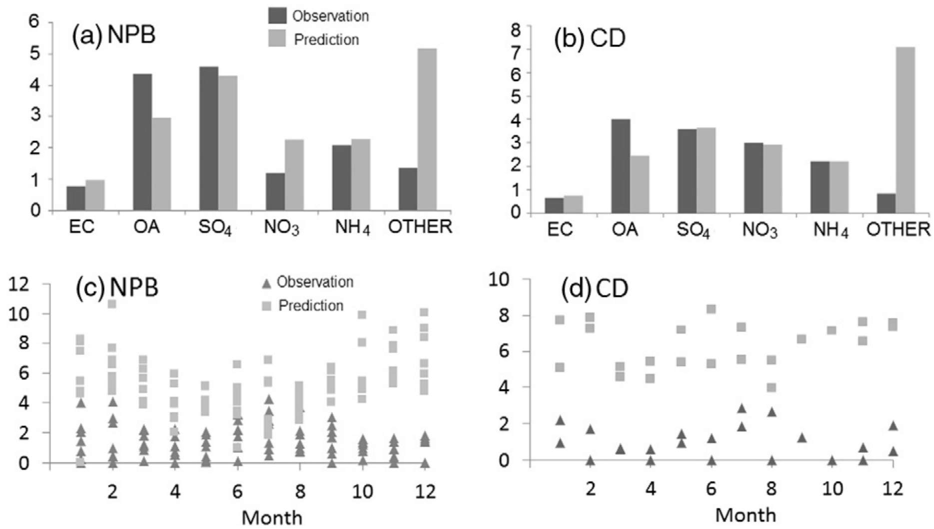

Fig. 12. Comparison of observed and predicted 7-year averaged $PM_{2.5}$ components for (a) NPB and (b) CD and comparison of observed and predicted monthly averaged $PM_{2.5}$ OTHER for (c) NPB and (d) CD. Units are μg m^{-3}.

and organic aerosol (OA) mass concentrations in the NPB domain are lower than observations by approximately 1.5 μg m^{-3} and 1 μg m^{-3}, respectively, and predicted nitrate concentration is higher than observation by approximately 1 μg m^{-3}. These results agree with results shown in Figs. 10 and 11. For the CD domain, CMAQ model predictions of average EC, sulfate, nitrate, and ammonium concentrations agree closely with observations but OA is under-predicted by 1.5 μg m^{-3}. In both domains, mass concentration of other $PM_{2.5}$ aerosol components combined (OTHER) is significantly over-predicted by 3.5 to 6 μg m^{-3}.

Several fugitive dust related factors likely contribute to the over-prediction: 1) over estimation of the $PM_{2.5}$ fraction of PM_{10} in fugitive dust emission calculations (Pace, 2005a). Based on Pace (2005a) the fractions might have been overestimated by approximately 70% on average; 2) overestimation of PM_{10} fugitive dust emissions. This could be caused by improper treatment of near-source removal of fugitive dust or overestimation of raw emissions before correction for the near-source removal. Although county-specific transport fractions (Pace, 2005b) have been applied in both 2001 CAIR and 2005 NEI to account for the near-source removal, the transportable fraction approach lacks monthly variability, which would be expected due to seasonal changes in vegetative cover. Further, the variability due to soil moisture, precipitation, and wind speeds is not accounted for by the methodology. Fig. 12(c) shows that for the NPB domain, observations do not show the obvious seasonal variation predicted by the model, and are universally lower than predictions at all times. However, Fig. 12(d) shows that for the CD domain, observations are significantly lower than predictions but there is no clear seasonal trend in both predictions and observations. It appears that correct simulation of the OTHER component requires improvement in $PM_{2.5}/PM_{10}$ split and PM_{10} emission estimations.

Although fugitive dust is usually the major source of the OTHER component, many other sources, such as biomass burning, can also have significant fractions of the OTHER component. Overestimation of $PM_{2.5}$ emission from these sources can contribute to the overestimation of the OTHER components as well. In addition, the OTHER component in $PM_{2.5}$ emission profiles is usually estimated by the difference between measured $PM_{2.5}$ mass emission rate and the resolved $PM_{2.5}$ components. Error in the PM component measurements such as aerosol water content estimation and organic aerosol mass (which requires estimated organic carbon to organic mass ratio) can lead to artificial inflation of the OTHER component (Pace, 2005a). A study of the sources of the PM components and PM speciation profiles is needed to understand the underlying causes of the over-prediction of the OTHER PM components.

4. Discussions and conclusions

The WRF model generally over-predicts temperature in all of the four study domains, with MB values as high as 2 K, although the monthly averaged MB and GE values are generally within the recommended performance criteria by McNally (2009). WRF model performance for water vapor content is satisfactory for spring, summer, and fall, but not wintertime when significant biases and errors occur. Deviation in the water vapor concentrations could affect predictions of OH radical formation during the day. It can also affect particle hygroscopic growth. However, it is not obvious in this study that wintertime deviation in the water vapor content has affected gas or PM model performance. Wind speed is adequately predicted by the WRF model and generally falls in the model performance criteria recommended by Emery et al. (2001) with a general trend of over-prediction, as reflected in the positive MB values in three of the four domains. The over prediction of wind speed, typically occurs at the low wind range by the WRF model, has been previously noted in a number of studies (see Hu et al., under review-b and the references therein). Wind directions, especially in summer months, are not well reproduced with GE values as large as 50°. However, observation nudging is not used in the WRF simulations, while the meteorology simulations reported in Emery et al. (2001) study applied observation nudging, which improves the model performance at observation locations. Inaccuracy in the wind direction can affect CMAQ model performance of all species, especially in high spatial resolution simulations.

The pollutant concentrations predicted by CMAQ are generally within the model performance criteria. O_3 performance meets the EPA criteria (MNB within ± 0.15, AUP within 0.2 and NGE lower than 0.25) except for under-predictions of a few months for the SP domain. Performance of PM_{10} is generally satisfactory with mean fractional bias (MFB) within ± 0.6, but a large under-prediction in springtime was frequently observed. Performance of $PM_{2.5}$ and its measured components is mostly within performance goals suggested by Boylan and Russell (2006) except for organic carbon (OC), which is universally under-predicted with MFB values as large as -0.8. The predicted frequency distribution of $PM_{2.5}$ generally agrees with observations although the predictions are slightly biased towards more frequent high concentrations in most areas. Elemental carbon (EC), nitrate and sulfate concentrations are also well reproduced. The other unresolved $PM_{2.5}$ components (OTHER) are significantly overestimated by more than a factor of 2. No conclusive explanations can be made regarding the possible cause of this universal overestimation, which warrants a follow-up study to better understand this problem.

Based on the authors' knowledge, this is the first study for the eastern United States that uses a spatial resolution higher than 12 km in long term air quality simulations for health effect analysis. Detailed performance statistics reported in this study can be served as a reference for future high spatial resolution CTM applications.

Acknowledgment

This project is partially supported by a grant from the U.S. Environmental Protection Agency's Science to Achieve Results (STAR) program through grant #RD-83386501. The authors want to acknowledge the Texas A&M Supercomputing Facility (http://sc.tamu.edu) and the Texas Advanced Computing Center (http://www.tacc.utexas.edu/) for providing computing resources essential for completing the research reported in this paper. Although the research described in the article has been funded in part by the U.S. Environmental Protection Agency's STAR program, it has not been subjected to any EPA review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.scitotenv.2013.11.121.

References

Arunachalam S, Wang BY, Davis N, Baek BH, Levy JI. Effect of chemistry-transport model scale and resolution on population exposure to $PM_{2.5}$ from aircraft emissions during landing and takeoff. Atmos Environ 2011;45:3294–300.

Bell ML. The use of ambient air quality modeling to estimate individual and population exposure for human health research: a case study of ozone in the Northern Georgia Region of the United States. Environ Int 2006;32:586–93.

Boylan JW, Russell AG. PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. Atmos Environ 2006;40:4946–59.

Bravo MA, Fuentes M, Zhang Y, Burr MJ, Bell ML. Comparison of exposure estimation methods for air pollutants: ambient monitoring data and regional air quality simulation. Environ Res 2012;116:1–10.

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

Carter WPL. Documentation of the SAPRC-99 chemical mechanism for VOC reactivity assessment. Report to the California Air Resources Board; 2000 [Available at http://cert.ucr.edu/ carter/absts.htm#saprc99 and http://www.cert.ucr.edu/ carter/reactdat.htm.].

Chang HH, Reich BJ, Miranda ML. Time-to-event analysis of fine particle air pollution and preterm birth: results from North Carolina, 2001–2005. Am J Epidemiol 2012;175:91–8.

Emery C, Tai E, Yarwood G. Enhanced meteorological modeling and performance evaluation for two Texas episodes. In: Corp International, editor. Report to the Texas Natural Resources Conservation Commission, p.b.E; 2001. [Novato, CA.].

Fast JD, Gustafson WI, Easter RC, Zaveri RA, Barnard JC, Chapman EG, et al. 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 - Atmos 2006;111.

Foley KM, Roselle SJ, Appel KW, Bhave PV, Pleim JE, Otte TL, et al. Incremental testing of the Community Multiscale Air Quality (CMAQ) modeling system version 4.7. Geosci Model Dev 2010;3:205–26.

Garcia-Menendez F, Yano A, Hu YT, Odman MT. An adaptive grid version of CMAQ for improving the resolution of plumes. Atmos Pollut Res 2010;1:239–49.

Gilliam RC, Pleim JE. Performance assessment of new land surface and planetary boundary layer physics in the WRF-ARW. J Appl Meteorol Clim 2009;49:760–74.

Grabow ML, Spak SN, Holloway T, Stone B, Mednick AC, Patz JA. Air quality and exercise-related health benefits from reduced car travel in the midwestern United States. Environ Health Perspect 2012;120:68–76.

Held T, Ying Q, Kaduwela A, Kleeman M. Modeling particulate matter in the San Joaquin Valley with a source-oriented externally mixed three-dimensional photochemical grid model. Atmos Environ 2004;38:3689–711.

Hogrefe C, Lynn B, Knowlton K, Goldberg R, Rosenzweig C, Kinney PL. Long-term regional air quality modelling in support of health impact analyses. Nato Sci Peace Secur 2008;605–613.

Hu J, Zhang H, Chen S-H, Ying Q, Wiedinmyer C, Vandenberghe F, et al. Identifying $PM_{2.5}$ and $PM_{0.1}$ sources for epidemiological studies in California. Environ Sci Tech 2013a. [Under review].

Hu J, Zhang H, Chen S, Wiedinmyer C, Vandenberghe F, YIng Q, et al. Predicting primary $PM_{2.5}$ and $PM_{0.1}$ trace composition for epidemiological studies in California. Environ Sci Technol 2013b. [Under review].

Lippmann M. Health effects of tropospheric ozone. Environ Sci Technol 1991;25:1954–62.

Luo C, Wang YH, Mueller S, Knipping E. Diagnosis of an underestimation of summertime sulfate using the Community Multiscale Air Quality model. Atmos Environ 2011;45:5119–30.

Matsui H, Koike M, Takegawa N, Kondo Y, Griffin RJ, Miyazaki Y, et al. Secondary organic aerosol formation in urban air: temporal variations and possible contributions from unidentified hydrocarbons. J Geophys Res - Atmos 2009;114.

McNally DE. 12 km MM5 performance goals. Presentation to the Ad-hov Meteorology Group; 2009.

Misenis C, Zhang Y. An examination of sensitivity of WRF/Chem predictions to physical parameterizations, horizontal grid spacing, and nesting options. Atmos Res 2010;97:315–34.

Morris R, Tai E. Preliminary ozone model performance evaluation for the Denver 2008 episode. Draft report for the Denver Regional Air Quality Council. Novato, California: ENVIRON International Corporation; 2012.

Pace TG. Examination of the multiplier used to estimate $PM_{2.5}$ fugitive dust emissions from PM_{10}. 14th Emission Inventory Conference. Las Vegas, NV: United States Environmental Protection Agency; 2005a.

Pace TG. Methodology to estimate the transportable fraction (TF) of fugitive dust emissions for regional and urban scale air quality analyses; 2005b [U. S. EPA].

Poschl U. Atmospheric aerosols: composition, transformation, climate and health effects. Angew Chem Int Ed 2005;44:7520–40.

Simon H, Baker KR, Phillips S. Compilation and interpretation of photochemical model performance statistics published between 2006 and 2012. Atmos Environ 2012;61:124–39.

Stroud CA, Makar PA, Moran MD, Gong W, Gong S, Zhang J, et al. Impact of model grid spacing on regional- and urban-scale air quality predictions of organic aerosol. Atmos Chem Phys 2011;11:3107–18.

Tong DQ, Mathur R, Kang DW, Yu SC, Schere KL, Pouliot G. Vegetation exposure to ozone over the continental United States: assessment of exposure indices by the Eta-CMAQ air quality forecast model. Atmos Environ 2009;43:724–33.

Turpin BJ, Lim H. Species contributions to $PM_{2.5}$ mass concentrations: revisiting common assumptions for estimating organic mass. Aerosol Sci Technol 2001;35:602–10.

U.S. EPA. Guidance on the use of models and other analyses for demonstrating attainment of air quality goals for ozone, $PM_{2.5}$ and regional haze. In: U.S.E.P Agency, editor. North Carolina: Research Triangle Park; 2007.

Volkamer R, Jimenez JL, San Martini F, Dzepina K, Zhang Q, Salcedo D, et al. Secondary organic aerosol formation from anthropogenic air pollution: rapid and higher than expected. Geophys Res Lett 2006;33.

Vukovich JM, Pierce T. The implementation of BEIS3 within the SMOKE modeling framework. MCNC-Environmental Modeling Center. Research Triangle Park and National Oceanic and Atmospheric Administration; 2002.

Wang LT, Jang C, Zhang Y, Wang K, Zhang Q, Streets DG, et al. Assessment of air quality benefits from national air pollution control policies in China. Part I: background, emission scenarios and evaluation of meteorological predictions. Atmos Environ 2010;44:3442–8.

Wiedinmyer C, Akagi SK, Yokelson RJ, Emmons LK, Al-Saadi JA, Orlando JJ, et al. The Fire INventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning. Geosci Model Dev 2011;4:625–41.

Winner DA, Cass GR. Modeling the long-term frequency distribution of regional ozone concentrations. Atmos Environ 1999;33:431–51.

WRAP. 2002 fire emission inventory for the WRAP region —phase II. Air Sciences Inc; 2005.

Ying Q, Lu J, Allen P, Livingstone P, Kaduwela A, Kleeman MJ. Modeling air quality during the California Regional PM_{10}/$PM_{2.5}$ Air Quality Study (CRPAQS) using the UCD/CIT source-oriented air quality model —part I. Base case model results. Atmos Environ 2008;42:8954–66.

Zhang H, Ying Q. Source apportionment of airborne particulate matter in Southeast Texas using a source-oriented 3D air quality model. Atmos Environ 2010;44:3547–57.

Zhang H, Ying Q. Secondary organic aerosol formation and source apportionment in Southeast Texas. Atmos Environ 2011;45:3217–27.

Zhang H, Li J, Ying Q, Yu JZ, Wu D, Cheng Y, et al. Source apportionment of $PM_{2.5}$ nitrate and sulfate in China using a source-oriented chemical transport model. Atmos Environ 2012;62:228–42.