Atmospheric Environment 61 (2012) 124–139

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

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

Review

# Compilation and interpretation of photochemical model performance statistics published between 2006 and 2012

Heather Simon*, Kirk R. Baker, Sharon Phillips

*U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, 109 TW Alexander Dr., Research Triangle Park, NC 27711, USA*

## HIGHLIGHTS

- Compilation of operational air-quality model evaluations published from 2006 to 2012.
- Model performance summarized for ozone and $PM_{2.5}$.
- Model performance also shown for wet deposition of sulfate, nitrate, ammonium, and Hg.
- Benefits of common performance metrics are discussed and evaluated.
- Recommendations given on how to perform evaluations for regulatory applications.

## ARTICLE INFO

*Article history:*
Received 21 February 2012
Received in revised form
2 July 2012
Accepted 6 July 2012

*Keywords:*
Model performance evaluation
CMAQ
CAMx
Particulate matter
Ozone
Wet deposition
Operational evaluation
$PM_{2.5}$
Mercury

## ABSTRACT

Regulatory and scientific applications of photochemical models are typically evaluated by comparing model estimates to measured values. It is important to compare quantitative model performance metrics to a benchmark or other studies to provide confidence in the modeling results. Since strict model performance guidelines may not be appropriate for many applications, model evaluations presented in recent literature have been compiled to provide a general assessment of model performance over a broad range of modeling systems, modeling periods, intended use, and spatial scales. Operational model performance is compiled for ozone, total $PM_{2.5}$, speciated $PM_{2.5}$, and wet deposition of sulfate, nitrate, ammonium, and mercury. The common features of the model performance compiled from literature are photochemical models that have been applied over the United States or Canada and use modeling platforms intended to generally support research, regulatory or forecasting applications. A total of 69 peer-reviewed articles which include operational model evaluations and were published between 2006 and March 2012 are compiled to summarize typical model performance. The range of reported performance is presented in graphical and tabular form to provide context for operational performance evaluation of future photochemical model applications. In addition, recommendations are provided regarding which performance metrics are most useful for comparing model applications and the best approaches to match model estimates and observations in time and space for the purposes of metric aggregations.

Published by Elsevier Ltd.

## 1. Introduction

Eulerian photochemical models implement numeric algorithms to predict air pollutant concentrations and deposition on local to continental scales. These models calculate the effects of emissions, transport, chemistry, particle physics, and deposition to estimate concentrations of air pollutants in space and time. Photochemical models are applied for a range of purposes such as evaluation of air pollution control scenarios by state and local governments, development of national air pollution rules, forecasting of air quality for public health and safety, investigations of scientific questions about atmospheric chemistry and physics, and research on the health effects of air pollution. All of these applications require credible science and acceptable model performance.

Regulating agencies must demonstrate that the modeling platform used to support potential control implementation compares well with observations. However, there is no strict guideline for model performance given the large variability in model applications and intended uses (United States Environmental Protection Agency, 2007). Operational performance evaluation is the most common

* Corresponding author. Tel.: +1 919 541 1803; fax: +1 919 541 3613.
*E-mail address:* simon.heather@epa.gov (H. Simon).

1352-2310/$ – see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.atmosenv.2012.07.012

approach to compare model estimates to the corresponding measured pollutant concentrations. Diagnostic evaluations investigate specific model processes frequently using specialized measurements and modeling tools such as process analysis and the direct decoupled method which isolate and track the affects of individual chemical and physical processes. Dynamic evaluations look at model response to perturbations of the inputs such as the predicted relative change in ozone resulting from the reduction of nitrogen oxide ($NO_x$) emissions (Napelenok et al., 2011). Diagnostic and dynamic evaluations provide useful insight into model formulation and parameterization, whereas operational performance evaluations against measurements taken at routine monitor networks provide more broadly contextual information for comparison with new modeling applications (Hogrefe et al., 2008).

Modeling studies for regulatory purposes need to provide context about how well that modeling application compares to measured values in relation to other independent studies. This builds confidence that the modeling study appropriately captures processes which lead to high pollution episodes and thus is a reliable tool to support pollution control strategy development. Currently, there is no article in the peer-reviewed literature that summarizes model performance for ozone, speciated $PM_{2.5}$, and wet deposition from recently published studies that could provide a useful benchmark against which regulatory modelers can gauge their model performance. Model performance reviews for specific pollutants have been published in the past. An ozone model performance review paper (Tesche, 1988) similarly compiles ozone photochemical model performance from studies up through the 1980s. While the performance metrics reported in previous decades may fall within ranges presented in this review paper, the recommendation from that era are not relevant today given the increasing complexity in photochemical model chemistry and physics, domain size and resolution, configuration options, and length of application.

This work presents a compilation of operational model evaluations for ozone, $PM_{2.5}$, and wet deposition of sulfate, nitrate, ammonium, and mercury based on model evaluations published in peer reviewed journals between 2006 and March 2012. While operational model performance metrics are simple to estimate, the literature review revealed a general tendency by researchers to provide little detail regarding fundamental assumptions made for aggregate metric calculations. In addition, this review uncovered a general lack of consistency in evaluation methodologies hindering our ability to compile a complete summary of comparable evaluation results. Consequently this study also presents recommendations for operational model performance evaluation and reporting so that future studies can provide a clearer picture of the state-of-the-science in air-quality model performance.

## 2. Methods

A total of 69 peer-reviewed articles published between 2006 and March 2012 have been compiled. These studies report quantifiable metrics from operational evaluations of regional photochemical models applied over some portion of the United States or Canada. Data from studies that reported results in graphical form only were not included in this review unless metrics could be quantitatively determined from these graphs. Statistical metrics are compiled on model performance for both ambient pollutant concentrations and wet deposition mass. Table 1 summarizes the general characteristics of the studies identified in this synthesis. Many studies reported multiple values for various performance metrics of a specific pollutant. When these values came from separate model simulations or were separated by season or region of the country, they were included as separate data points in our review. When separate performance metrics were reported for the same model simulation but for different monitoring networks, for different sites within a small region, or for different days within a season, the best reported model performance is used in this analysis.

The studies compiled for this project present a variety of model performance metrics: mean bias/error (MB/ME), root mean square error (RMSE), mean normalized bias/error (MNB/MNE), normalized mean bias/error (NMB/NME), fractional bias/error (FB/FE), unpaired peak accuracy (UPA), index of agreement (IofA), and the coefficient of determination ($r^2$). Table 2 shows the formulas for calculating each statistical metric summarized in this paper. Metrics given in the same units as the measurements (absolute metrics) are MB, ME, and RMSE. MB quantifies the tendency of the model to over- or under-estimate values while ME and RMSE measure the magnitude of the difference between modeled and observe values regardless of whether the modeled values are higher or lower than observations. For these metrics to be meaningful, the evaluator must be familiar with typical observed magnitudes in order to understand whether bias and error are “large”.

One disadvantage of absolute metrics is that they make intercomparisons of model performance in clean and polluted environments or across different pollutants difficult to interpret. Consequently, a range of relative metrics are often used. These metrics are presented either in fractional or percentage units. In this analysis all normalized values are given as percentages. MNB, MNE, NMB, and NME all normalize by observed values. For MNB, the normalization is paired in space and time with the reported bias/error while NMB first finds the mean absolute bias/error and then normalizes this value based on the mean observed value. MNB and MNE have the tendency to be weighted towards larger percentages because bias/error at very small observed values is often very large in percentage units (Boylan and Russell, 2006). This behavior means that these metrics often do a poor job of describing model performance in polluted conditions because they are so heavily weighted by modeled to measured differences at low observed values. In addition, this sometimes leads to a positive MNB when the absolute MB is negative. NMB and NME do not suffer from these problems. FB and FE normalize bias and error by the average of the observed and modeled concentration. Some researchers prefer the FB bias metric because it is symmetrical around 0 (Boylan and Russell, 2006). The range of possible FB values are from −200% to 200%, while the range of MNB and NMB values is from −100% to +infinity. Therefore, anyone interpreting the MNB or NMB metrics must understand that the magnitude of positive values in fractional units is equivalent to one minus the inverse of the magnitude of negative values in fractional units. For instance a factor of ten over-prediction leads to a NMB = 1000% (10 in fractional units) while a factor of ten under-prediction leads to a NMB = −90% (−0.9 in fractional units). Two studies (Appel et al., 2008; Foley et al., 2010) report median instead of mean values for bias, error, and normalized bias and error. Since only two studies used median values, reported MdnB, MdnE, NMdnB, and NMdnE values are grouped with MB, ME, NMB, and NME values.

Metrics such as IofA, $r$, and $r^2$ provide a sense of the strength of relationship between model estimates and observations that have been paired in time and space. The coefficient of determination ($r^2$) is 1 when modeled estimates and observations have a perfect linear relationship and 0 where no linear relationship exists. The coefficient of determination can be physically interpreted as indicating the portion of variability in the model prediction which can be accounted for by variability in the observed values. The correlation coefficient ($r$) does not have the same strict operational interpretation as the coefficient of determination but is often used because

**Table 1**
Summary of operational evaluation studies included in this review.

| Reference | Models evaluated | Pollutants evaluated | Seasons evaluated | Regions evaluated |
|---|---|---|---|---|
| (Appel et al., 2007) | CMAQ | Ozone (8-h max) | Summer | Eastern US |
| (Appel et al., 2008) | CMAQ | $PM_{2.5}$, ammonium, nitric acid, nitrate, $TNO_3$, sulfate, OC, EC, TC | Annual, summer | Eastern US |
| (Appel et al., 2010) | CMAQ | Ozone, ozone (8-h max), $PM_{2.5}$, ammonium wetdep, nitrate, $TNO_3$, nitrate wetdep, sulfate, sulfate wetdep, TC | Summer, winter | Eastern US |
| (Appel et al., 2011) | CMAQ | Ammonium wetdep, nitrate wetdep, sulfate wetdep | Annual, fall, spring, summer, winter | Eastern US, North America |
| (Arnold and Dennis, 2006) | CMAQ | Ozone (1-h max) | Summer | Southeast US |
| (Baker and Scheff, 2007) | CAMx | Ammonia, ammonium, nitric acid, nitrate, $TNO_3$, $SO_2$, sulfate, $TSO_4$ | Annual | Midwest |
| (Baker and Scheff, 2008) | CAMx | Ammonia wetdep, nitrate wetdep, sulfate wetdep | Fall, spring, summer, winter | Midwest |
| (Baker and Bash, 2012) | CAMx, CMAQ | Total Hg wetdep | Annual, fall, spring, summer, winter | Eastern US, Western US |
| (Bullock et al., 2009) | CMAQ, REMSAD, TEAM | Total Hg wetdep | Annual | Continental US |
| (Byun et al., 2007) | CAMx, CMAQ | Ozone | Summer | Houston area |
| (Carlton and Baker, 2011) | CMAQ | Formaldehyde, isoprene | Summer | Midwest |
| (Chen et al., 2008) | CMAQ | Ozone (8-h max), $PM_{2.5}$, ammonium, nitrate, sulfate, OC, EC | Summer/fall, summer | Pacific Northwest |
| (Cho et al., 2009) | AURAMS | Sulfate | Fall | Edmonton Canada |
| (Eder and Yu, 2006) | CMAQ | Ozone (1-h max), ozone (8-h max), $PM_{2.5}$, ammonium, nitrate, sulfate, OC, EC | Annual | North America |
| (Eder et al., 2006) | CMAQ | Ozone (8-h max) | Summer | California, Lower Midwest, Northeastern US, Pacific Northwest, Rocky Mountains, southeast US, Upper Midwest |
| (Eder et al., 2009) | CMAQ | Ozone (8-h max) | Summer | North America |
| (Foley et al., 2010) | CMAQ | Ozone (8-h max), $PM_{2.5}$, ammonium, ammonium wetdep, nitrate, $TNO_3$, nitrate wetdep, sulfate, sulfate wetdep, OC, EC, TC, $PM_{other}$ | Winter, summer | Eastern US |
| (Gaydos et al., 2007) | PMCAMx | Ozone, NO, $NO_2$, $PM_{2.5}$, ammonium, $TNH_4$, nitrate, $SO_2$, sulfate, $TNO_3$, OC, EC, TC | Summer | Eastern US, Pittsburgh |
| (Gego et al., 2006) | CMAQ, REMSAD | Nitrate, sulfate | Fall, spring, summer, winter | Central US, Great Lakes, Northeastern US, Southeast US, Western US |
| (Gong et al., 2006) | AURAMS | Ammonium, $HNO_3$, nitrate, $TNO_3$, $SO_2$, sulfate, total S | Summer | Eastern US |
| (Gorline and Lee, 2009) | CMAQ | $PM_{2.5}$ | Summer | Great Lakes, Northeastern US, Rocky Mountains |
| (Grell et al., 2005) | MM5/Chem, WRF/Chem | Ozone, ozone (avg 11am–7pm), $NO_y$, $SO_2$ | Summer | Northeastern US |
| (Hogrefe et al., 2007) | CMAQ | Ozone (8-h max), $PM_{2.5}$ | Summer, winter | New York |
| (Hogrefe et al., 2008) | CAMx, CMAQ | Ozone (8-h max) | Summer | Northeastern US |
| (Hogrefe et al., 2011) | CMAQ | Ozone (8-h max) | Summer | Northeastern US |
| (Jin et al., 2010) | CMAQ | Ozone, ozone (1-h max), ozone (8-h max), CO, NMHC, $NO_y$ | Summer | California |
| (Kang et al., 2010) | CMAQ | Ozone (8-h max), $PM_{2.5}$ | Summer, winter | Lower Midwest, Northeastern US, Pacific coast, Rocky Mountains, Southeast US, Upper Midwest |
| (Karamchandani et al., 2006) | CMAQ-MADRID, CMAQ-MADRID-APT | $PM_{2.5}$, ammonium, $HNO_3$, nitrate, $SO_2$, sulfate, OC, EC | Summer, winter | Southeast US |
| (Kim et al., 2009) | WRF/Chem | $NO_2$ – column | Summer | North America |
| (Kim et al., 2010) | CMAQ | Ozone (8-h max) | Summer | Southeast US |
| (Lee et al., 2011) | CMAQ | $PM_{2.5}$ | Summer | Central US, Eastern US, North America, Western US |
| (Lin et al., 2007) | CMAQ | Total Hg wetdep | Summer, winter | North America |
| (Lin et al., 2012) | CMAQ | Total Hg wetdep | Annual | North America |
| (Liu et al., 2010) | CMAQ | Ozone (1-h max), ozone (8-h max), $PM_{2.5}$, ammonium, ammonium wetdep, nitrate, nitrate wetdep, sulfate, sulfate wetdep, OM, EC | Summer, winter | North Carolina |
| (Makar et al., 2010) | AURAMS | Ozone, ozone (1-h max), $PM_{2.5}$, $PM_{2.5}$ (1-h max) | Summer | Canada, North America, Ontario |

**Table 1** (*continued*)

| Reference | Models evaluated | Pollutants evaluated | Seasons evaluated | Regions evaluated |
|---|---|---|---|---|
| (Marmur et al., 2009) | CMAQ | Ozone, CO, NO, $NO_y$, $PM_{2.5}$, ammonium, $HNO_3$, nitrate, $SO_2$, sulfate, SOIL, OC, EC | Annual | Southeast US |
| (Misenis and Zhang, 2010) | WRF/Chem | Ozone, CO, NO, $NO_2$, $PM_{2.5}$ | Summer | Houston area |
| (Molders et al., 2010) | WRF/Chem | $PM_{10}$, $PM_{2.5}$ | Summer | Alaska |
| (Morris et al., 2006) | CAMX, CMAQ | OC, TC | Summer | Southeast US, Central US, Midwest, Northeastern US, Western US |
| (Napelenok et al., 2011) | CMAQ | Ozone (8-h max) | Summer | Eastern US |
| (Otte, 2008) | CMAQ | Ozone (1-h max) | Summer | Eastern US |
| (Park et al., 2010) | AURAMS | $PM_{2.5}$, SOIL | Fall, spring, summer, winter | Eastern US, Western US |
| (Pun et al., 2006) | CMAQ-MADRID | Ozone, $PM_{2.5}$, ammonium, nitrate, $SO_2$, sulfate, OM, EC | Summer | Texas and neighboring states |
| (Queen and Zhang, 2008a, b) | CMAQ | Ammonium, ammonium wetdep, nitrate, nitrate wetdep, sulfate, sulfate wetdep | Summer, winter | North Carolina |
| (Rodriguez et al., 2009) | CAMx | Ozone | Annual | Western US |
| (Rodriguez et al., 2011) | CAMx | Ozone, $NO_x$, ammonia, ammonium, nitrate, $SO_2$, sulfate | Annual | Western US |
| (Roy et al., 2007) | CMAQ | Ammonium, nitrate, sulfate, OC, EC, TC | Spring, summer | Rocky Mountains, North America, Western US |
| (Sakulyanontvittaya et al., 2008) | CMAQ | OM | Summer | North America |
| (Seigneur et al., 2006) | TEAM | Total Hg wet deposition | Annual | North America |
| (Smyth et al., 2006) | CMAQ | Ozone, ozone (1-h max), $PM_{2.5}$, ammonium, nitrate, sulfate, OM | Summer, winter | North America |
| (Smyth et al., 2009) | AURAMS, CMAQ | Ozone, ozone (1-h max), ozone (1-h min), $PM_{2.5}$, ammonium, nitrate, sulfate, OM, EC | Summer | Pacific Northwest |
| (Spak and Holloway, 2009) | CMAQ | $PM_{10}$, $PM_{2.5}$, ammonium, nitrate, sulfate, OM, EC | Summer | North America |
| (Stroud et al., 2011) | AURAMS | OM | Fall, spring, summer, winter | Upper Midwest |
| (Tang et al., 2011) | CMAQ | Ozone (8-h max), $NO_x$ | Summer | Eastern US |
| (Tarasick et al., 2007) | AURAMS, CHORNOS | Ozone | Summer | Houston area |
| (Tesche et al., 2006) | CMAQ | Ammonium, nitrate, sulfate, OC, EC | Summer | Eastern US, North America |
| (Tong and Mauzerall, 2006) | CMAQ | Ozone | Annual | Southeast US |
| (Vijayaraghavan et al., 2007) | CMAQ-MADRID | Total Hg wet deposition | Summer | North America |
| (Vijayaraghavan et al., 2008) | CMAQ-AMSTERDAM | Total Hg wet deposition | Summer | North America |
| (Wu et al., 2008) | CMAQ | Ozone (1-h max), ozone (8-h max), $PM_{2.5}$, ammonium, nitrate, sulfate, OC, EC | Annual | Continental US |
| (Ying et al., 2008) | UCD/CIT | Ozone, CO, NO, $NO_2$, $PM_{2.5}$, ammonium, nitrate, sulfate, OC, EC | Winter | California |
| (Yu et al., 2006) | CMAQ | Ozone, ozone (1-h max), ozone (8-h max), CO, NO, $NO_2$, $NO_y$, PAN, $SO_2$ | Summer | Northeastern US |
| (Yu et al., 2007) | CMAQ | Ozone, ozone (1-h max), ozone (8-h max), CO, NO, $NO_y$, $SO_2$ | Summer | Northeastern US, Eastern US |
| (Yu et al., 2008) | CMAQ | $PM_{2.5}$, $PM_{2.5}$ (hourly), ammonium, nitrate, $SO_2$, sulfate, OC, EC, TC | Summer | Eastern US |
| (Zhang et al., 2006) | CMAQ | Ozone, ozone (1-h max), ozone (8-h max), $PM_{10}$, $PM_{2.5}$, ammonium, nitrate, sulfate, OC, EC | Summer | North America, Southeast US |
| (Zhang et al., 2007) | CMAQ | Ozone (8-h max), OC | Fall, spring, summer, winter | Eastern US, Southeast US, Western US, North America |
| (Zhang et al., 2009) | CMAQ | Ozone (1-h max), ozone (8-h max), $PM_{2.5}$, ammonium, ammonium wetdep, nitrate, nitrate wetdep, sulfate, sulfate wetdep, OC, EC | Annual, fall, spring, summer, winter | North America |
| (Zhang et al., 2010) | WRF/Chem-MADRID | Ozone, $PM_{2.5}$ | Summer | Texas |

it provides an indication of the strength of linear relationship and is signed positive or negative based on the slope of the linear regression. For the purpose of this synthesis, all $r$ values were converted to $r^2$ to increase comparability between studies. The UPA metric is intended to measure a model's ability to capture peak pollutant concentrations, but does not pair the model estimates with observations in time or space.

Performance is best when bias and error metrics approach zero and when the coefficient of determination approaches 1. However, as shown in Section 3.1, perfect agreement for any metric alone may not be indicative of good model performance. Multiple metrics must be considered when evaluating model performance. In addition, perfect performance is not possible given that observations themselves are subject to uncertainties related to measurement technique and analytical approach. Models cannot be expected to achieve accuracy beyond the measurement uncertainty of the instruments. It is important to know which, if any, corrections have been made for measurement artifacts. For example, $PM_{2.5}$ is measured gravimetrically from Teflon filters which likely have measurement artifacts from nitrate volatilization, organic carbon (OC) positive artifact, OC negative artifact, and water absorption (Frank, 2006; Simon et al., 2011). In addition, some $NO_x$

**Table 2**
Definitions of performance metrics.

| Abbreviation | Term | Definition[a] |
|---|---|---|
| MB | Mean bias | $\frac{1}{N}\sum(M_i - O_i)$ |
| ME | Mean error | $\frac{1}{N}\sum\lvert M_i - O_i\rvert$ |
| RMSE | Root mean squared error | $\sqrt{\frac{\sum(M_i - O_i)^2}{N}}$ |
| FB | Fractional bias | $100\% \times \frac{2}{N}\sum\frac{(M_i - O_i)}{(M_i + O_i)}$ |
| FE | Fractional error | $100\% \times \frac{2}{N}\sum\frac{\lvert M_i - O_i\rvert}{(M_i + O_i)}$ |
| NMB | Normalized mean bias | $100\% \times \frac{\sum(M_i - O_i)}{\sum O_i}$ |
| NME | Normalized mean error | $100\% \times \frac{\sum\lvert M_i - O_i\rvert}{\sum O_i}$ |
| MNB | Mean normalized bias | $100\% \times \frac{1}{N}\sum\left(\frac{M_i - O_i}{O_i}\right)$ |
| MNE | Mean normalized error | $100\% \times \frac{1}{N}\sum\left\lvert\frac{M_i - O_i}{O_i}\right\rvert$ |
| UPA | Unpaired peak accuracy | $100\% \times \frac{(M_{peak} - O_{peak})}{O_{peak}}$ |
| I of A | Index of agreement | $1 - \frac{\sum(M_i - O_i)^2}{\sum(\lvert M_i - \overline{O}\rvert + \lvert O_i - \overline{O}\rvert)^2}$ [2] |
| $r^2$ | Coefficient of determination | $\left(\frac{\sum_1^N((M_i - \overline{M}) \times (O_i - \overline{O}))}{\sqrt{\sum_1^N(M_i - \overline{M})^2 \sum_1^N(O_i - \overline{O})^2}}\right)$ |

monitors show interference from $NO_y$ species such as PAN and nitric acid (Dunlea et al., 2007). If no artifact corrections have been made then measured and modeled values may represent different pollutant species. Finally, photochemical grid model estimates are grid cell averages and may not always be commensurate with observations that represent a finite space around a point measurement.

## 3. Results

The 69 articles identified for this analysis used a variety of metrics to characterize model performance for estimating ambient concentrations and wet deposition of many different chemical pollutants. Fig. 1 shows the frequency of use for each statistical metric. The most commonly reported metrics are mean bias, normalized mean bias, normalized mean error, and $r/r^2$. Both ME and RMSE are frequently used as non-normalized error metrics.

Figs. 2 and 3 show the frequency of evaluation of different pollutants broken down by model and by metric. Evaluations of eight different regional photochemical models are included in this literature review: A Unified Regional Air-quality Modeling Systems (AURAMS) (Zhang et al., 2002), the Comprehensive Air Quality Model with extensions (CAMx) (ENVIRON, 2010), the Canadian Hemispheric and Regional Ozone and $NO_x$ System (CHRONOS) (Pudykiewicz and Koziol, 2001), the Community Multiscale Air Quality (CMAQ) model (Foley et al., 2010), the fifth generation PSU/NCAR Mesoscale Model coupled with Chemistry (MM5/Chem) (Grell et al., 2000), the Regional Modeling System for Aerosols and Deposition (REMSAD) (SAI, 2002), the UC Davis/California Institute of Technology (UCD/CIT) model (Kleeman and Cass, 2001), and the Weather Research and Forecasting model coupled with Chemistry (WRF/Chem) (Grell et al., 2005). The CHRONOS and AURAMS models are developed by Environment Canada, CAMx and REMSAD are developed by private companies (Environ and ICF consulting), CMAQ is developed at the US EPA, UCD/CIT is developed at CalTech and UC-Davis, and MM5/Chem and WRF/Chem are developed by a team of international researchers from universities and Federal agencies. Additional models included for mercury evaluation include TEAM (Trace Element Analysis Model) and CMAQ derivative models CMAQ-MADRID (Model of Aerosol Dynamics, Reaction, Ionization, and Dissolution) and CMAQ-AMSTERDAM (Advanced Modeling System for Transport, Emissions, Reactions, and Deposition of Atmospheric Matter). Fig. 2 shows that CMAQ was the most commonly evaluated model in the literature between 2006 and March 2012.

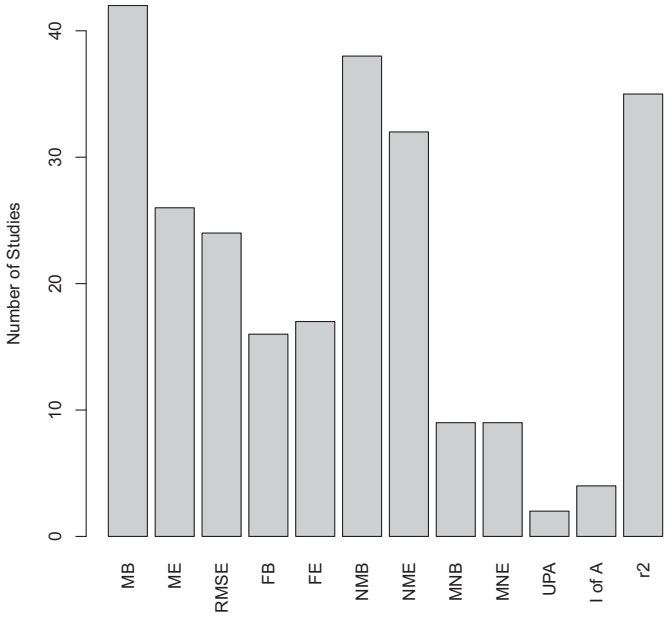


**Fig. 1.** Number of studies using various metrics in operation evaluation studies compiled in this review.

Making intercomparison of model evaluation studies is challenging because studies evaluate pollutants using different time averages, different pollutant classifications, and different metrics.

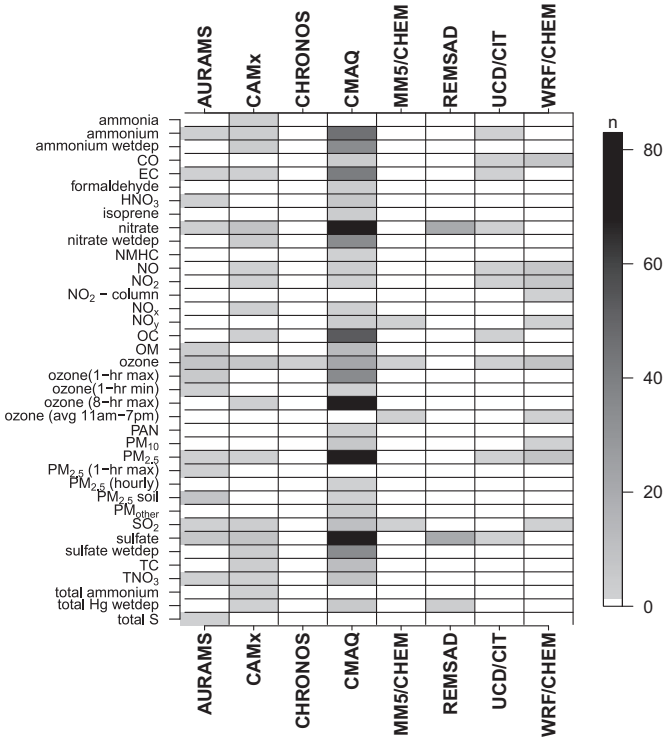


**Fig. 2.** Number of studies evaluating each paired pollutant and photochemical model.

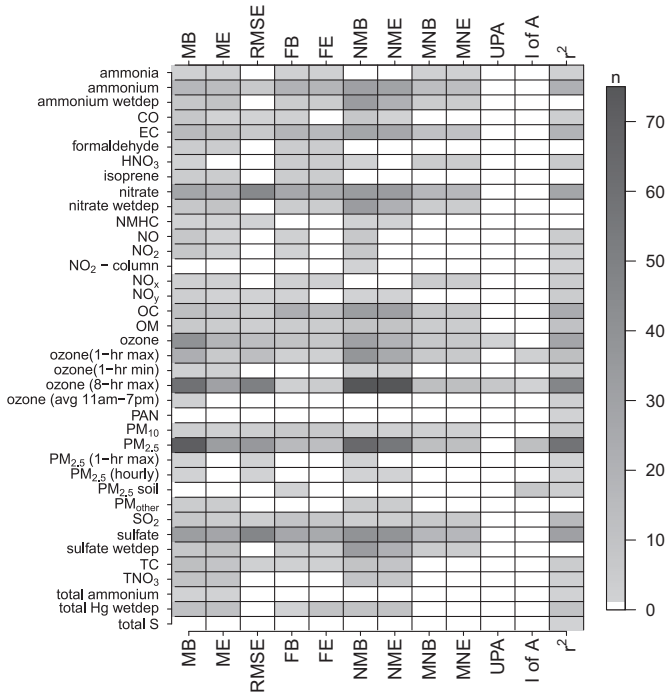


**Fig. 3.** Number of studies reporting each performance metric for every pollutant evaluated.

Also, these studies employed a large range of model set-ups which include varying horizontal and vertical grid resolutions, land surface models, planetary boundary layer schemes, chemical mechanisms, thermodynamic partitioning modules, numerical solvers, and others. In addition to different model parameterizations, many of the key inputs differ in these studies: initial conditions, boundary conditions, meteorology and emissions. Finally, evaluations are often performed on different temporal scales (several day episode to multi-year) and spatial scales (urban to continental; 1 km grid resolution to 36 km grid resolution). The intention of this study is to present context for future photochemical model performance, not to evaluate models, certain studies, or model formulations.

### 3.1. Ozone

Ozone is routinely measured on an hourly basis at monitors in several networks in the United States: EPA's Air Quality System (AQS) monitoring network, the SouthEastern Aerosol Research and Characterization Study (SEARCH) network, and the Clean Air Status Trends Network (CASTNET). In addition, surface and upper air ozone measurements are made during special studies. Modeled estimates are often compared to measurements at different time scales: hourly, daily 1-h maximum and daily 8-h maximum. 1-h daily maximum evaluations are intended to be relevant to the form of the 1979 ozone National Ambient Air Quality Standard (NAAQS) (a daily maximum 1-hr average of 0.12 ppm). 8-h daily maximum evaluations are intended to be relevant to the form of the 1997 and 2008 ozone NAAQS (a daily maximum 8-h average of 0.08 ppm and 0.075 ppm respectively).

Fig. 4 shows a summary of ozone performance reported in the identified literature. Performance for one metric cannot be directly compared to any other metric because each metric includes a unique subset of studies. The apparent tendency of ozone overestimation is largely due to the approach taken by investigators of averaging performance metrics over high and low ozone hours and days. Low observed ozone concentrations occur much more frequently than high observed ozone concentrations, so metrics calculated over all observed values are dominated by performance at low concentrations. Studies showing ozone performance with a minimum threshold or by bins of observed ozone concentration show that the highest observed ozone concentrations are generally underestimated (Foley et al., 2010).

When studies evaluate hourly ozone estimates at all hours of the day, they aggregate day and nighttime formation and destruction periods and may not group physically meaningful ozone concentrations together. Therefore, studies intended for regulatory applications should also evaluate bias and error for hourly or daily maximum ozone when ambient concentrations are above a minimum cutoff value or by discrete groups of observed ozone concentration. However, very few of the studies included in this review used such a cutoff. Out of 20 studies which reported hourly ozone performance, only 4 evaluated ozone above a cutoff. In addition, out of the 21 and 13 studies which reported performance statistics for 8-h daily maximum ozone and 1-h daily maximum ozone respectively, only one study in each case evaluated ozone performance using a cutoff. Cutoffs used included 40 ppb (Pun et al., 2006; Yu et al., 2007; Zhang et al., 2006), 60 ppb (Zhang et al., 2006), and 5th and 95th percentile values (Hogrefe et al., 2011). Appel et al. (2007) and Foley et al. (2010) showed ozone performance in 10 ppb bins of observed concentration.

To demonstrate the impact of differentiating metric estimates based on levels of ambient concentration, we re-evaluate modeled ozone outputs from Appel et al. (2011). Paired 8-h daily maximum ozone observations and model predictions were aggregated over high and low pollution episodes using no ambient concentration cutoff, a 60 ppb cutoff, and a 75 ppb cutoff. Separate statistics were calculated for the ozone season (May–September) of each year between 2002 and 2006 and for four regions of the U.S.: Northeast, Midwest, Southeast, and Central. The ranges of the statistical metrics for 8-h daily maximum ozone among these years and regions are shown in Table 3 (Table S1 shows this same set of information for hourly ozone). The statistics confirm that 8-h maximum ozone is overestimated when all data is included, but similar to Foley et al. (2010), modeled 8-h max ozone is underestimated when statistics are only calculated for ambient values above cutoffs of 60 and 75 ppb. In addition to the difference in the direction of ozone bias for high and low ambient ozone levels, it is clear that overall performance varies with ozone concentration. For instance, the $r^2$ values are substantially lower for high ozone days (between 0.05 and 0.28 for ambient 8-h max ozone above 75 ppb) than for all days (between 0.56 and 0.73). The fractional error and normalized mean error are lowest when a cutoff of 60 ppb is applied indicating that the model performs best for mid-range ambient ozone values. This analysis showed that the modeling performed by Appel et al. (2011) demonstrates an ability to predict ozone concentrations and variability for mid-range values but is less skilled at capturing these characteristics on either very high or very low ozone days. Consequently, in addition to providing aggregate statistics about model performance across all conditions, any model evaluation of ozone for regulatory purposes should also focus on 8-h max ozone estimates for days when ambient concentrations are high. Current EPA modeling guidance (United States Environmental Protection Agency, 2007) recommends calculating performance statistics both without a threshold and with a 60 ppb threshold. The analysis presented here supports the use of a 60 ppb threshold for showing performance at high observed ozone concentrations where modeling is intended for regulatory purposes.

Fig. 5 shows reported ozone performance by grid resolution and indicates that mean bias, as reported in the literature, does not

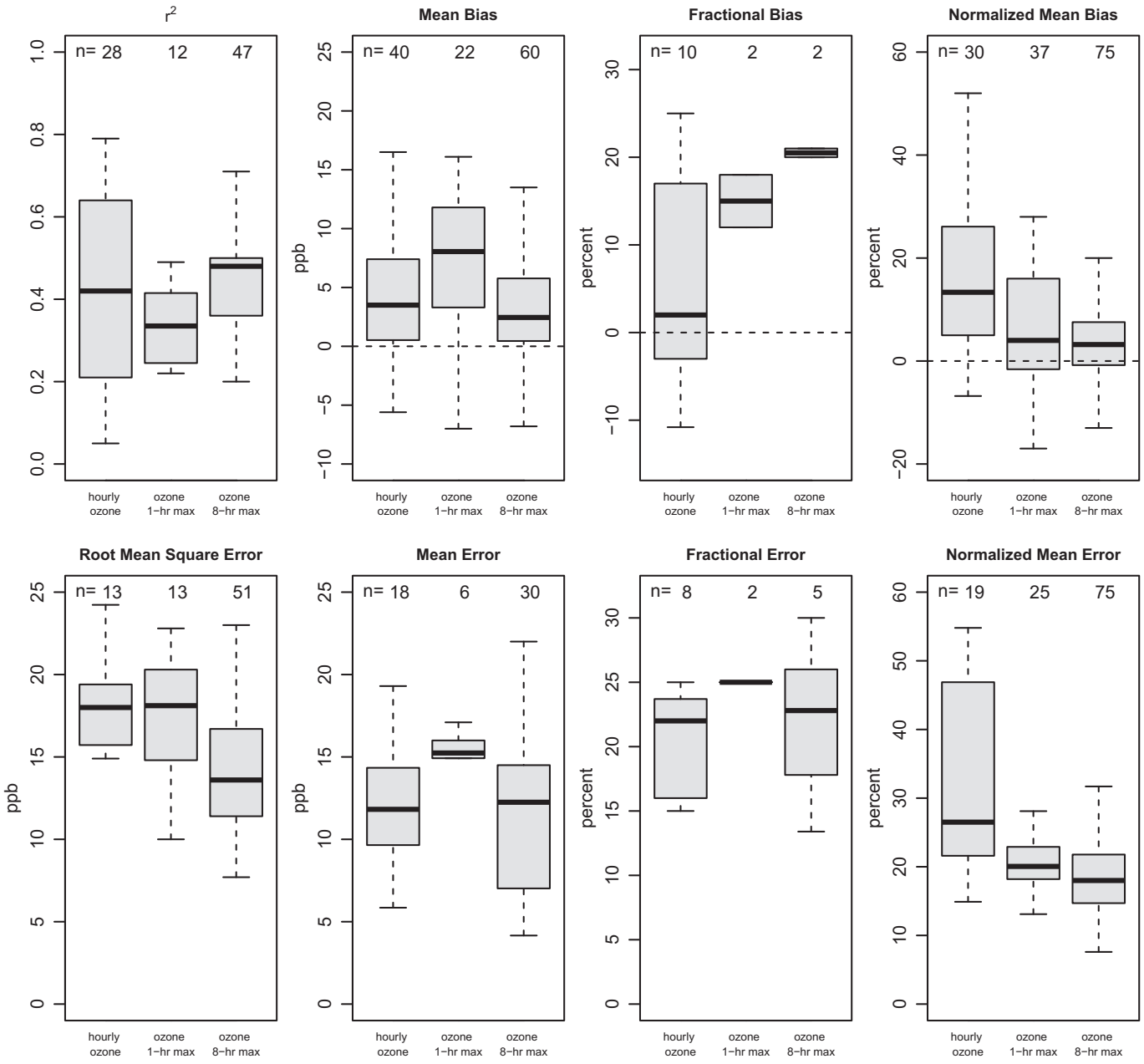


**Fig. 4.** Summary of ozone performance metrics reported in the evaluated modeling studies. Centerlines show median values, boxes outline the 25th and 75th percentile values and whiskers extend to 1.5 times the interquartile range.

systematically improve at higher grid resolutions. Modeling studies which used finer grid resolution (4 km or less) were more likely to report under-estimation of ozone than modeling studies using coarser grid resolution. This trend is likely due to the ways in which models are applied at coarse and fine resolution rather than the resolution itself. For example, model simulations conducted with a resolution at or finer than 4 km typically simulated episodes that ranged in length from 8 to 30 days. The typical episode length increased to 29–120 days for simulations at resolutions coarser than 4 km. In addition, the finer scale studies were much more likely to focus on small areas: 76% were conducted at a local scale (urban to state level) and 24% were conducted at a regional scale (covering several states). Conversely, only 19% of the coarser scale simulations were conducted on a local scale, while 35% were conducted on a regional scale, and 46% were conducted on a superregional scale (consisting of multiple regions and covering at least half of the United States). The local scale model simulations are often performed to capture high pollution episodes, when models tend to be biased low. Conversely, longer simulations covering larger regions tend to include locations and time periods of low and mid-range ozone concentrations. Therefore, the under-estimation of ozone in modeling using fine grid resolutions is likely due to the shorter duration episodes with higher ozone concentrations that are the focus of these studies.

Many studies reported multiple performance metrics, which gives a more complete picture of model evaluation. Scatter plots of $r^2$ values paired with bias (MB, NMB, and FB) are shown in Fig. 6 for 8-h daily maximum ozone and in Fig. 7 for hourly ozone. Each point represents results from a single modeling run. The comparison using 8-h daily maximum ozone in Fig. 6 shows that many low daily

**Table 3**
Performance metric distributions for 8-hr daily maximum ozone based on data from Appel et al. (2011).

| Ozone 8 h daily maximum | | | | Quantile estimate | | | | |
|---|---|---|---|---|---|---|---|---|
| Cut-off | Metric | Units | *n* | 10% | 25% | 50% | 75% | 90% |
| None | $r^2$ | | 20 | 0.56 | 0.59 | 0.62 | 0.64 | 0.73 |
| None | MB | ppb | 20 | 1.2 | 1.9 | 2.9 | 3.8 | 4.4 |
| None | FB | % | 20 | 5.5 | 5.7 | 8.4 | 10.4 | 11.4 |
| None | NMB | % | 20 | 2.4 | 3.9 | 6.1 | 7.8 | 9.6 |
| None | RMSE | ppb | 20 | 9.4 | 9.7 | 10.0 | 10.6 | 10.9 |
| None | ME | ppb | 20 | 7.3 | 7.5 | 7.8 | 8.1 | 8.5 |
| None | FE | % | 20 | 15.7 | 17.2 | 18.1 | 18.7 | 19.4 |
| None | NME | % | 20 | 14.9 | 16.0 | 17.0 | 17.2 | 18.0 |
| 60 ppb | $r^2$ | | 20 | 0.17 | 0.21 | 0.29 | 0.36 | 0.41 |
| 60 ppb | MB | ppb | 20 | −8.61 | −7.47 | −4.95 | −4.47 | −4.03 |
| 60 ppb | FB | % | 20 | −13.90 | −10.88 | −7.60 | −7.12 | −6.67 |
| 60 ppb | NMB | % | 20 | −12.51 | −10.07 | −6.94 | −6.38 | −6.00 |
| 60 ppb | RMSE | ppb | 20 | 9.49 | 9.61 | 10.30 | 12.83 | 13.14 |
| 60 ppb | ME | ppb | 20 | 7.51 | 7.60 | 8.10 | 9.78 | 10.41 |
| 60 ppb | FE | % | 20 | 11.17 | 11.58 | 12.15 | 14.38 | 15.80 |
| 60 ppb | NME | % | 20 | 10.68 | 11.15 | 11.55 | 13.58 | 14.58 |
| 75 ppb | $r^2$ | | 20 | 0.05 | 0.08 | 0.19 | 0.23 | 0.28 |
| 75 ppb | MB | ppb | 20 | −14.5 | −12.1 | −9.0 | −8.3 | −7.8 |
| 75 ppb | FB | % | 20 | −20.0 | −15.8 | −12.3 | −11.5 | −10.4 |
| 75 ppb | NMB | % | 20 | −17.5 | −14.0 | −11.1 | −10.3 | −9.2 |
| 75 ppb | RMSE | ppb | 20 | 12.8 | 13.3 | 13.7 | 16.6 | 19.5 |
| 75 ppb | ME | ppb | 20 | 10.5 | 10.9 | 11.4 | 13.8 | 15.3 |
| 75 ppb | FE | % | 20 | 13.9 | 14.1 | 15.0 | 17.7 | 21.0 |
| 75 ppb | NME | % | 20 | 12.8 | 13.0 | 13.9 | 16.0 | 18.5 |

8-h maximum biases are paired with high $r^2$ values in the studies evaluated. Two outliers shown in Fig. 6 with $r^2$ less than 0.1 come from the Hogrefe et al. (2011) study and represent performance of 8-h max ozone on low ozone days. Even without these two outlier points, a trend of decreasing bias with increasing $r^2$ is discernible although there are some low bias points with $r^2$ values between 0.2 and 0.4. The opposite relationship is shown in Fig. 7, where model simulations with the lowest mean hourly ozone bias also had very low $r^2$ values. This suggests that the low bias in hourly ozone in these studies is a result of averaging over- and under-estimates which does not provide a useful characterization of the ozone over space and time. This illustrates the importance of evaluating temporal subsets of ozone concentrations that are likely to result from similar formation environments and the value of considering multiple metrics to create a comprehensive understanding of model performance.

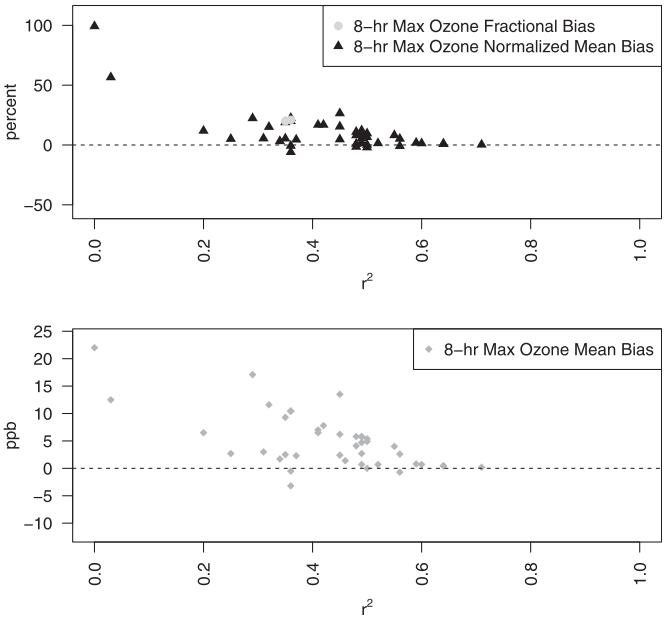


**Fig. 6.** Comparison of paired bias and $r^2$ values in modeled 8-hr maximum ozone estimates.

The range of reported model performance for ozone tended to be similar across many different subcategories including evaluations performed in the Eastern and Western U.S., modeling performed in forecast and retrospective modes, and modeling implemented with a variety of different chemical mechanisms or models. While this study is not designed to isolate causes for specific differences in model performance, it is worth noting that a specific chemical mechanism or modeling system did not show a pronounced improvement in ozone performance. Additional plots showing ozone performance split out by region (Eastern US vs Western US), spatial scale (local, regional, and superregional), and retrospective versus forecast modeling applications are provided in the Supplemental information.

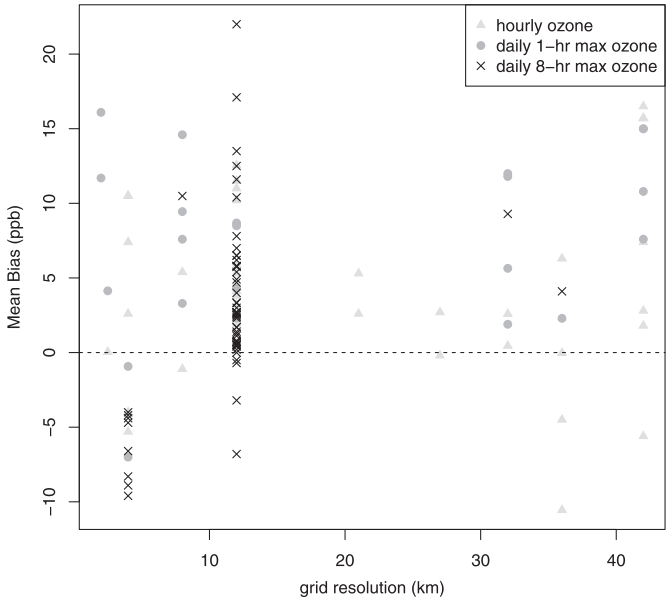


**Fig. 5.** Reported ozone mean bias as a function of grid resolution.

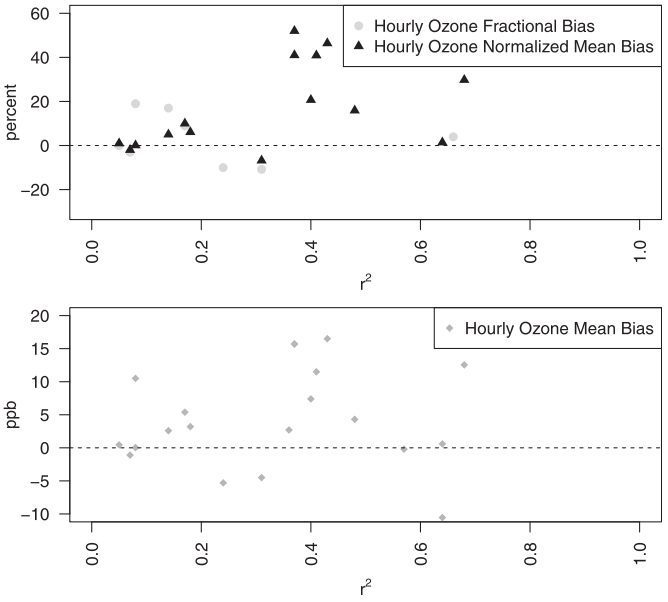


**Fig. 7.** Comparison of paired bias and $r^2$ values in modeled hourly ozone estimates.

### 3.2. Particulate matter

A total of 27 studies in this review reported model performance metrics for total $PM_{2.5}$ mass and 32 studies reported performance metrics for speciated $PM_{2.5}$. These studies compared modeled PM concentrations to measured values from the Chemical Speciation Network (CSN) (http://www.epa.gov/cgi-bin/htmSQL/mxplorer/query_spe.hsql), the Interagency Monitoring of Protected Visual Environments network (IMPROVE) (http://views.cira.colostate.edu/web/), the SouthEastern Aerosol Research and Characterization Study (SEARCH) network, and the Clean Air Status Trends Network (CASTNET). The CSN and IMPROVE networks provide 24-h average speciated $PM_{2.5}$ and total mass measurements every 1, 3, or 6 days while CASTNET measures weekly average concentrations. The CSN sites are generally located in urban areas while IMPROVE and CASTNET sites tend to be located in national parks and rural areas.

Model performance metric distributions for $PM_{2.5}$ and speciated components of $PM_{2.5}$ are shown in Fig. 8 and Table S2. These illustrate how compensating errors in the contributions of chemical components to total mass complicates the interpretation of model performance for total fine PM mass. Although more studies report negative than positive bias for each $PM_{2.5}$ species and total mass, some studies compiled here do report positive bias (MB, FB, or NMB) for every $PM_{2.5}$ component. Because some species may be

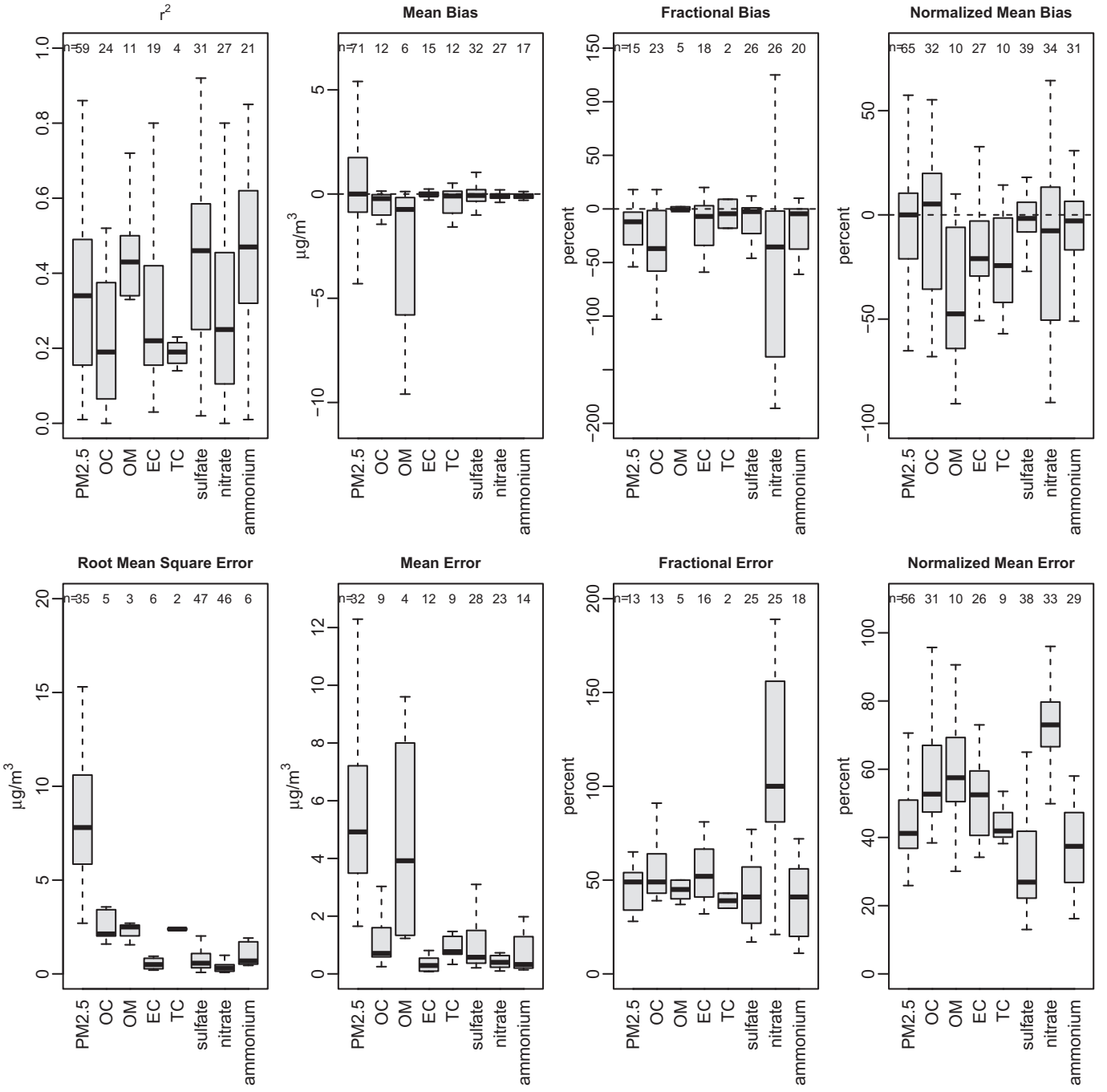


**Fig. 8.** Summary of PM performance metrics reported in the evaluated modeling studies. Centerlines show median values, boxes outline the 25th and 75th percentile values and whiskers extend to 1.5 times the interquartile range.

over-predicted while others are under-predicted, the evaluation of speciated $PM_{2.5}$ in addition to or in place of total mass provides more insight into emissions sources or atmospheric processes that are contributing to model performance.

Inter-comparison of model performance for speciated $PM_{2.5}$ is most challenging due to different pollutant classifications for carbonaceous aerosols and the "other" category. The "other" or "soil/crustal" category is sometimes defined using a linear combination of ions thought to be associated with crustal material with species-specific multipliers assuming the ions are fully oxidized. Alternatively, the "other" category is sometimes calculated as the difference between total measured $PM_{2.5}$ mass and the sum of the major speciated constituents. The "other" category in the model is generally made up of unspeciated primary $PM_{2.5}$ since it is common practice to split primary PM into sulfate, nitrate, OC, EC, and "other". Comparisons to ambient data for this second characterization is challenging to interpret since the type and magnitude of measurement artifacts are different for gravimetric techniques used to quantify total mass versus the chemical analyses used to characterize speciated components (Frank, 2006). Given the difficulty interpreting the "other" or "soil/crustal" category it is preferential to evaluate specific component species such as silicon or titanium when these species are treated explicitly in the simulation.

Carbonaceous aerosol is evaluated as organic mass (OM), organic carbon (OC), elemental carbon (EC), and/or total carbon (TC) (OC plus EC) in various articles. These disparate pollutant classifications limit our ability to intercompare studies. It is important to understand how each of these components is defined in both the model and measurements. Most monitoring networks report values of OC while many models estimate OM. To compare the model to measurements, an OM/OC ratio can be used either to convert measured carbon to total mass or modeled mass to carbon. When the former conversion is performed, it is common practice to use a single OM/OC ratio for all measurements (often 1.4 or 1.8) (Pun et al., 2006; Sakulyanontvittaya et al., 2008; Spak and Holloway, 2009; Stroud et al., 2011) even though many studies have shown that this ratio can vary substantially throughout the country (El-Zanan et al., 2005; Frank, 2006; Malm and Hand, 2007; Simon et al., 2011; Turpin and Lim, 2001). However, the latter conversion can be done more accurately in models which track precursor-specific secondary organic aerosol (SOA) species and whether OM is aged or fresh. For example, the CMAQ model tracks 19 different SOA species each with a species-specific OM/OC ratio (Carlton et al., 2010).

The evaluation of carbonaceous aerosols is further complicated by the fact that OC and EC are operationally defined. There are two common measurement techniques to quantify ambient OC and EC: thermal optical reflectance (TOR) and thermal optical transmittance (TOT). OC values determined by these two techniques are similar in magnitude (Chow et al., 2001), but EC values determined by TOT have been reported to average about 60% less than EC values determined by TOR (Chow et al., 2001). Ideally, the method used to determine OC and EC emissions splits (and thus model splits for primary TC) would be the same method used to determine OC and EC concentrations in the ambient aerosol. Historically, the CSN used TOT to calculate OC/EC splits, but has recently switched (between 2007 and 2009) to TOR (IMPROVE has always used TOR). One option to provide consistency is to evaluate TC which is comparable between the two methods. However, this approach will mask valuable information about model performance related to physical processes and emissions sources. Evaluating OC and EC should be done with knowledge of the techniques being used to split OC and EC for ambient measurements and for the emissions inventory used in the model simulation.

Model performance for many PM species has some seasonally consistent features across the studies included in this analysis. The vast majority of studies evaluated PM species during summer months, but some report metrics for other seasons. Fig. 9 shows reported NMB values for $PM_{2.5}$, sulfate, nitrate, and OC split out by season. A common trend among all studies is that $PM_{2.5}$ is overestimated during the winter and underestimated during the summer. OC and nitrate overestimates contribute to the wintertime overestimate of $PM_{2.5}$ total mass. Appel et al. (2008) and Foley et al. (2010) also report that unspeciated PM, mostly consisting of crustal material, can be substantially overestimated during the winter. Sulfate, nitrate, and OC are all reported as having negative bias

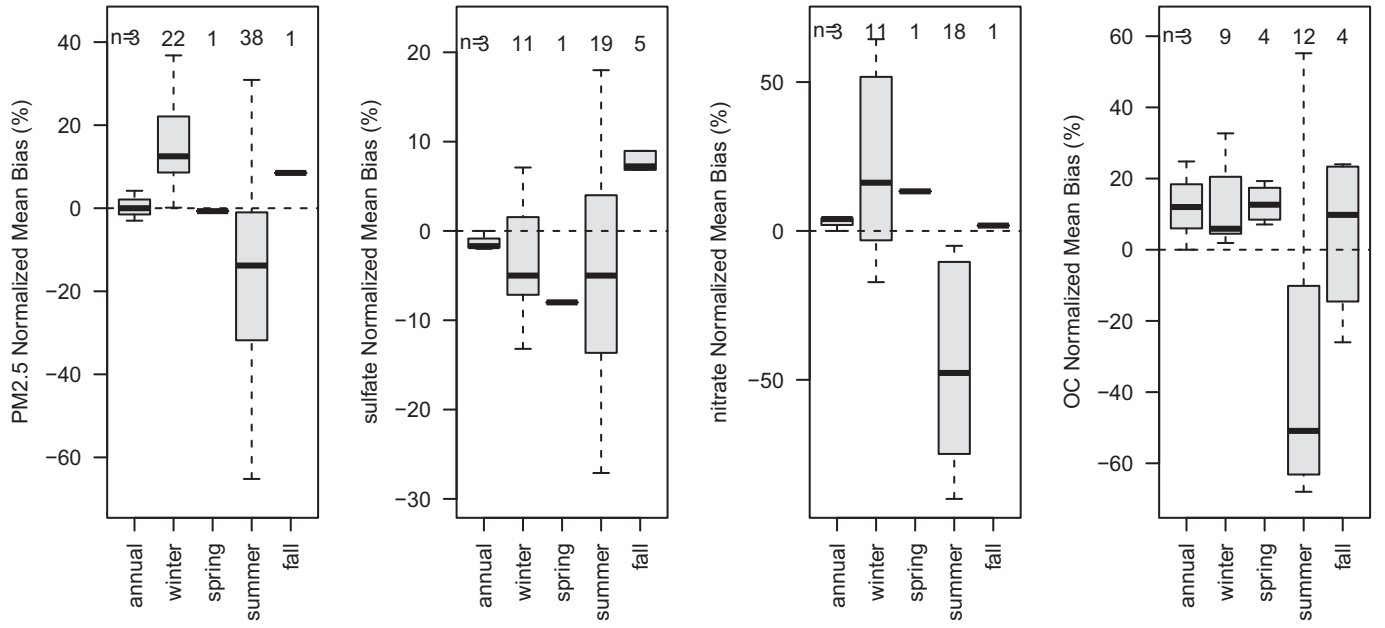


**Fig. 9.** PM NMB split out by season. Centerlines show median values, boxes outline the 25th and 75th percentile values and whiskers extend to 1.5 times the interquartile range.

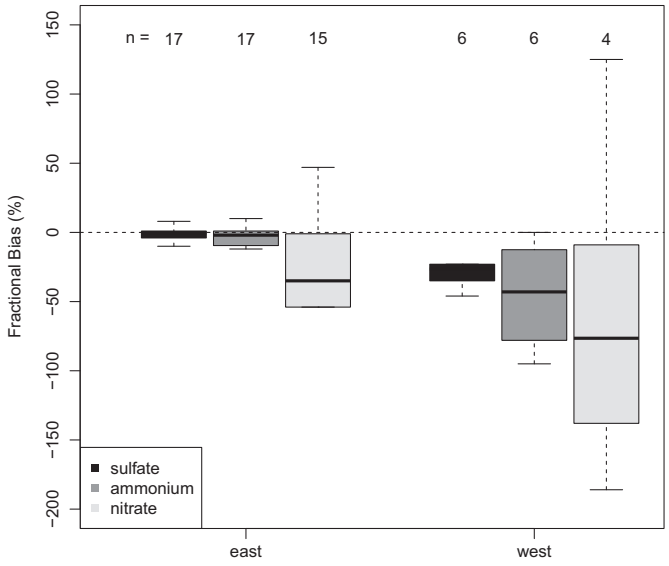


**Fig. 10.** PM FB in Eastern versus Western North America. Centerlines show median values, boxes outline the 25th and 75th percentile values and whiskers extend to 1.5 times the interquartile range.

during the summer. Since nitrate concentrations tend to be low in the summer, sulfate and OC contribute most of the summertime $PM_{2.5}$ underestimate reported in the literature. Model underestimates of SOA, which is a substantial component of OC during the summer (except during fire events), have been documented by a large number of studies (Carlton et al., 2010; de Gouw and Jimenez, 2009; Goldstein and Galbally, 2007; Simpson et al., 2007; Volkamer et al., 2006).

Model performance for $PM_{2.5}$ has been predominantly evaluated in the Eastern U.S. Fig. 10 shows a comparison of reported FB values for sulfate, nitrate, and ammonium in the East versus the West, with the Rocky Mountains as the dividing line. Other metrics and PM species are not shown due to lack of sufficient number of studies in the West. Reported model performance appears to be substantially better in the East versus the West. This result may be due to the predominant focus of model evaluations in the Eastern half of the US which presumably have been used to improve model inputs and processes in that region. In addition, measured and predicted sulfate concentrations are much lower in the West which may lead to higher relative biases. Also, high nitrate (and organic carbon) episodes which are related to meteorology that is specific to certain airsheds located in relatively complex terrain in the Western US may be difficult to fully capture in current models (Baker et al., 2011).

Additional plots showing speciated $PM_{2.5}$ performance for 6 metrics split out by grid resolution, region (Eastern US vs Western

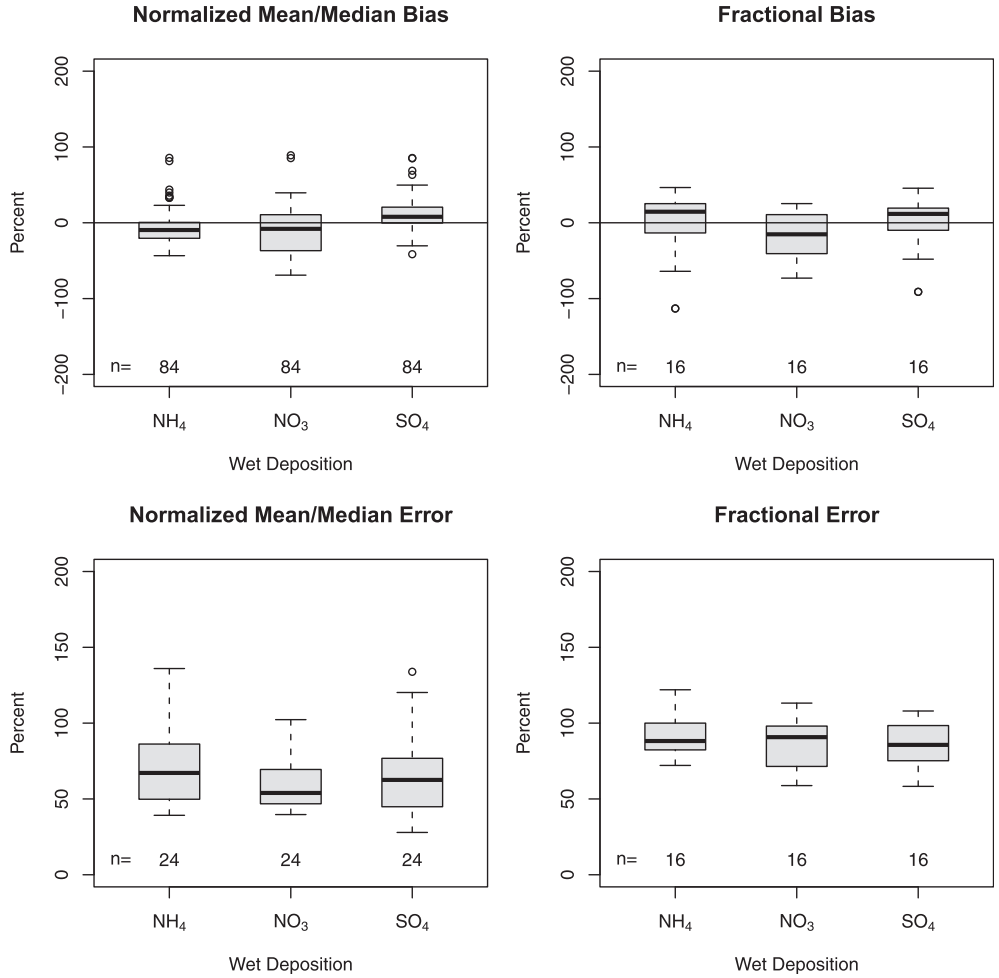


**Fig. 11.** Distribution of model performance metrics NMB, NME, FB, FE, NMdnB, and NMdnE for total wet deposition of sulfate, nitrate, and ammonium. Centerlines show median values, boxes outline the 25th and 75th percentile values and whiskers extend to 1.5 times the interquartile range.

US), spatial scale (local, regional, and superregional), and retrospective versus forecast modeling applications are provided in the Supplemental information.

### 3.3. Wet deposition: sulfate, nitrate, and ammonium

The evaluation of deposition is an important compliment to the evaluation of ambient estimates. Since deposition is a direct function of ambient concentrations the evaluation of deposited species can help partially explain over or under-estimates of ambient species. Ideally, total deposition through dry and wet processes would be evaluated for all species but only wet deposition for certain species are routinely measured in the United States. Total wet deposition of sulfate, nitrate, and ammonium are measured as weekly totals at National Air Deposition Program (NADP) (http://nadp.sws.uiuc.edu/NTN/) monitor locations (Bigelow, 1991). These observations are paired with model estimates in 8 separate studies, 7 of which apply CMAQ and one CAMx. The version of CMAQ ranges from 4.3 to the more recent 4.7 (Table 1). Spatial scale ranges from a single State (North Carolina) at 4 km grid resolution to regional at 12 km grid resolution and continental at 36 km grid resolution. The number of metrics extracted from a study ranges from 12 to 180.

Reported metrics in these studies include FB, FE, NMB, NME, MdnB, MdnE, NMdnB, and NMdnE. The distribution of reported NMB, NME, FB, FE, NMB, and NME metrics are shown in Fig. 11. The quantiles of the distribution for each performance metric are shown in Table S3 for quantitative comparison. In the studies examined, sulfate wet deposition is slightly overestimated while nitrate wet deposition is slightly underestimated. Ammonium wet deposition appears to be slightly underestimated when examining normalized mean bias and normalized median bias but slightly overestimated using fractional bias. It is not clear that photochemical modeling systems generally demonstrate more skill in estimating total wet deposition of one pollutant compared to another.

The studies included in this analysis include performance metrics estimated on an annual basis and by season emphasizing summer and winter months. Fig. 12 shows NMB and NME by season for total wet deposition of sulfate, ammonium, and nitrate. Error is typically highest during the warmer months when increased rainfall is more frequent. Total sulfate wet deposition tends to be overestimated during most of the year while total nitrate wet deposition tends to be underestimated in the summer and overestimated in the winter. The outliers reflecting poorer performance on these Figures are generally from applications with smaller grid resolution and domain size.

### 3.4. Wet deposition: mercury

Weekly total mercury wet deposition measurements are taken at sites that are part of the Mercury Deposition Network (http://nadp.sws.uiuc.edu/MDN/), which operate as part of the NADP (Vermette et al., 1995). Modeled estimates from 8 studies are compared to these observations and include 6 different photochemical models: CMAQ, TEAM, CAMx, REMSAD, CMAQ-MADRID, and CMAQ-AMSTERDAM. The spatial scales generally cover the continental United States or just the eastern United States with a horizontal grid resolution of 36 km. Grid resolutions less than or equal to 20 km are also included in this analysis but comprise less than half of the compiled metrics. Most of the studies are annual model simulations, approximately half of which are for the year 2001 or earlier (1995, 1996, and 1998). Few MDN sites were operational in the western U.S. before the early 2000s, meaning the

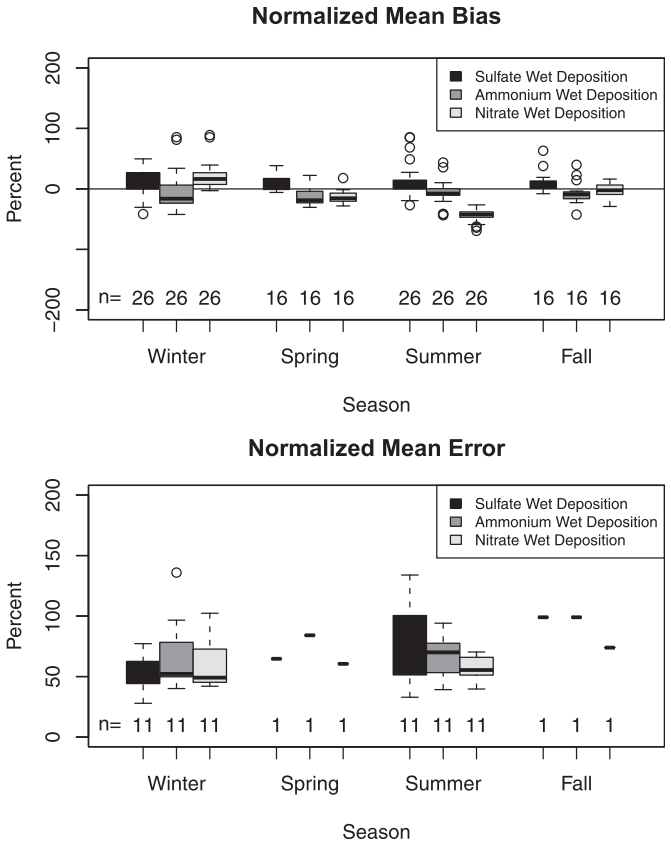


**Fig. 12.** Seasonal distribution of model performance metrics NMB and NME for total wet deposition of sulfate, nitrate, and ammonium. Lower and upper edges of the box represent the 25th and 75th percentiles.

performance presented in this analysis strongly focuses on the eastern United States.

The distribution of total mercury wet deposition NMB, NME, MB, ME, FE, and $r^2$ are shown in Fig. 13. The distribution quantiles for each performance metric are shown in Table S3 for quantitative comparison. Comparability between studies for reported performance metrics is challenging due to differences in averaging approaches. Some studies average all modeled and observed values at a particular monitor location then estimate performance metrics while others match observations and model estimates in space and time (weekly) then estimate performance metrics. Model performance for total mercury wet deposition is quite variable. Reported performance metrics indicate a general tendency of modeling systems to overestimate total mercury wet deposition. However, there is a clear need for more evaluation of total mercury wet deposition in the western United States using more recent modeling periods to take advantage of newer sites operating in that region. In addition, few studies evaluated mercury wet deposition at grid resolutions finer than 36 km. The outlier point for NMB in Fig. 13 represents a single simulation that was part of a larger study looking at performance of multiple mercury models with multiple meteorological input data (Bullock et al., 2009). The worst performing study for mean and fractional error is from the only 12 km application that estimated metrics specifically for the western United States (Baker and Bash, 2012). The outlier showing the best performance for normalized mean error reflects a modeling scenario that looks only at part of the eastern United States and that averaged the modeled and observed values before pairing to estimate the performance metric (Seigneur et al., 2006).

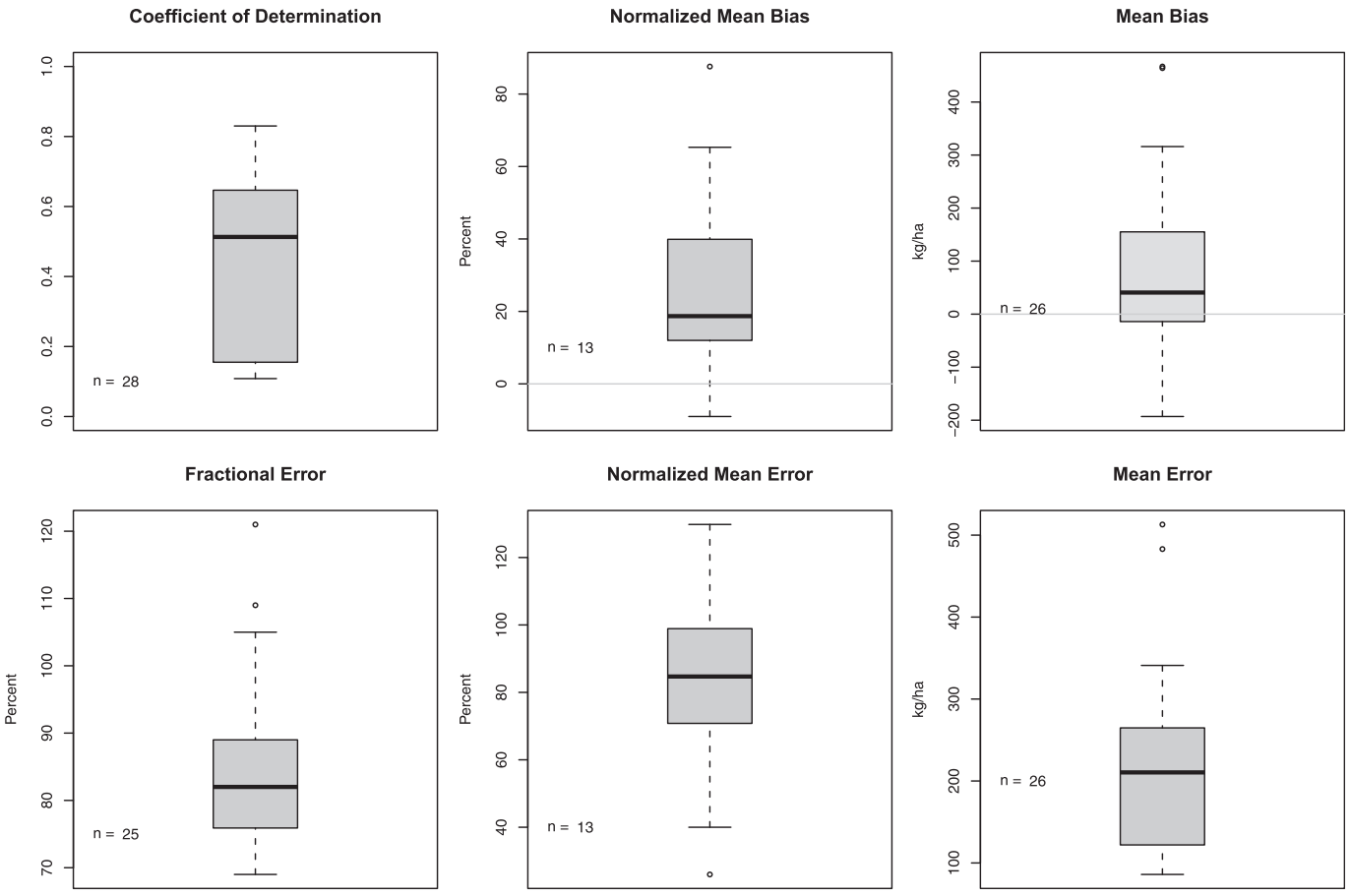


**Fig. 13.** Distribution of model performance metrics NMB, NME, MB, ME, FE, and $r^2$ for total mercury wet deposition. Lower and upper edges of the box represent the 25th and 75th percentiles.

## 4. Recommendations for regulatory model evaluations

Studies that present results intended to be relevant for regulatory modeling applications should at a minimum report mean observation, mean prediction, MB, ME (or RMSE), and a normalized bias and error (NMB/NME or FB/FE). In addition, the coefficient of determination provides useful information about the model's ability to capture observed variability. Reporting multiple performance metrics gives a more complete picture of the model's ability to capture magnitude of and variation in pollutant concentrations/deposition.

One problem associated with pairing the bias/error with the observed value is that data taken at low ambient concentrations can have very large percentages of bias/error. Consequently MNB and MNE values tend to be skewed by data points taken at very low concentrations and the bias tends to be skewed towards positive numbers. This metric can often result in counterintuitive results, for instance having a negative MB but a positive (and large) MNB. Given the propensity for misinterpretation and lack of symmetry around zero, the use of MNB and MNE metrics is not encouraged.

The literature review of operational model performance evaluations revealed large differences in the reporting of important details underlying estimation of aggregated metrics. Few of the studies explicitly state how observations and modeled estimates were paired in space and time before metrics were calculated. Many studies did not mention whether any spatial or temporal averaging was performed prior to the metric calculation. Evaluating data by pairing measurements with observations on the highest temporal resolution available and for important regulatory averaging times (e.g. 8-h daily maximum ozone) will give the most meaningful results. It is important to report data processing steps for pairing model predictions and observation data in time and space and whether data are spatially/temporally averaged before or after statistics are calculated. The most appropriate approach is to match observed estimates with the modeled estimate from the grid cell where the monitor is located. For any evaluation it is important to not solely focus on grid cells with monitors but examine spatial plots of model estimates for reasonableness. Depending on the nature of the problem or pollutant you may want to consider if an important feature was captured by the model but was spatially displaced. At grid resolutions less than 12 km, it may be useful to use a bilinear interpolation of modeled values in the grid cells nearest the monitor to better characterize model skill.

Metrics should be calculated for subsets of pollutants that are most relevant for understanding the processes leading to high pollution episodes. For ozone, different processes tend to dominate the formation, destruction, and transport during times of high and low ambient concentrations. Consequently, model performance at higher observed concentrations (above 60 ppb) provides insight into how well the model replicates ozone pollution episodes. For $PM_{2.5}$, different chemical species generally have different sources and formation pathways. Therefore, it is especially important to evaluate each $PM_{2.5}$ component separately in order to understand whether the model is properly replicating the causes of high $PM_{2.5}$ episodes. For similar reasons it is desirable to evaluate model predictions subsetted by season and region, as different processes

may dominate pollution emissions and formation at different times and locations. Evaluations of performance over long time periods and large areas provide a robust characterization of overall performance but may fail to distinguish specific episodes for which the model either performs especially well or especially poorly. Depending on the nature and intent of the particular study, finer temporal and spatial aggregations may be appropriate if there are sufficient data to permit the calculation of meaningful statistics. For instance, metrics aggregated by day and monitor may be necessary to understand specific pollution events.

Although ozone and $PM_{2.5}$ are the most commonly evaluated species in photochemical models, these models also output estimates of other criteria pollutants: carbon monoxide (CO), nitrogen dioxide ($NO_2$) and sulfur dioxide ($SO_2$). Proper treatment of $NO_2$ and $SO_2$ in photochemical models is important since they are precursors to particulate matter (and ozone in the case of $NO_2$). In addition, CO is often evaluated as a tracer for vehicle emissions or other combustion sources and can help in the interpretation of model performance for other pollutants originating from these sources (de Gouw and Jimenez, 2009; Docherty et al., 2008; Slemr et al., 2002).[1] Only 14 of the 69 studies evaluated photochemical model performance for one or more of these species (CO, $NO_2$, or $SO_2$) (Fig. S-1). In addition to criteria pollutants, other ozone and PM precursors have been evaluated in a limited number of studies (Fig. S-2). Carlton and Baker (2011) evaluated formaldehyde and isoprene concentrations in the Ozark Mountains, Jin et al. (2010) report performance for non-methane hydrocarbons, Yu et al. (2006) report performance for peroxy acetyl nitrate (PAN), and seven studies report performance for various classifications of nitrogen species (NO, $NO_x$, and $NO_y$). More routine evaluation of ozone and $PM_{2.5}$ precursors and chemical intermediates are desirable but may be limited by available measurements.

In addition to quantitative performance statistics that were the focus of this paper, other qualitative analyses may be useful for understanding model performance. Visualizing data through time series plots of modeled and observed ozone at one or more monitors and maps of mean bias/error statistics can help modelers identify times and locations of especially good or poor model performance. These analyses along with the statistics discussed at length may be used to improve model formulations or model inputs in order to achieve more accurate model simulations.

A range of model performance for ozone, $PM_{2.5}$, and wet deposition of various species are presented in this paper. This allows future model application projects to provide context for operational performance metrics. Future work should focus on metrics that are robust when aggregated, including mean bias, mean error, normalized mean bias, normalized mean error, fractional bias, and fractional error and include a description of how observations and predictions are paired in time and space before averaging. Ozone performance should be presented in bins of observed ozone to provide insight into the different physical and chemical processes that may influence ozone formation. When elevated levels of ozone are the focus, a minimum threshold of 60 ppb may be applied to remove prediction–observation pairs that are less relevant to the level of the ozone NAAQS. Finally, it is necessary to understand measurement artifacts and measurement uncertainty in order to make a meaningful interpretation of comparisons of modeled data.

[1] It is noteworthy that point measurements of CO (which can have steep gradients near sources) may not be commensurate with model predictions that are averaged over the area of a grid cell.

## Disclaimer

Although this work was reviewed by EPA and approved for publication, it may not necessarily reflect official Agency policy.

## Acknowledgments

The authors would like to recognize the contribution of Barron Henderson, Neal Fann, Wyat Appel, and Norm Possiel.

## Appendix A. Supplementary information

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.atmosenv.2012.07.012.

## References

Appel, K.W., Bhave, P.V., Gilliland, A.B., Sarwar, G., Roselle, S.J., 2008. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: sensitivities impacting model performance; part II – particulate matter. Atmospheric Environment 42, 6057–6066.

Appel, K.W., Foley, K.M., Bash, J.O., Pinder, R.W., Dennis, R.L., Allen, D.J., Pickering, K., 2011. A multi-resolution assessment of the Community Multiscale Air Quality (CMAQ) model v4.7 wet deposition estimates for 2002–2006. Geoscientific Model Development 4, 357–371.

Appel, K.W., Gilliland, A.B., Sarwar, G., Gilliam, R.C., 2007. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: sensitivities impacting model performance. Part I – ozone. Atmospheric Environment 41, 9603–9615.

Appel, K.W., Roselle, S.J., Gilliam, R.C., Pleim, J.E., 2010. Sensitivity of the Community Multiscale Air Quality (CMAQ) model v4.7 results for the eastern United States to MM5 and WRF meteorological drivers. Geoscientific Model Development 3, 169–188.

Arnold, J.R., Dennis, R.L., 2006. Testing CMAQ chemistry sensitivities in base case and emissions control runs at SEARCH and SOS99 surface sites in the southeastern US. Atmospheric Environment 40, 5027–5040.

Baker, K., Scheff, P., 2007. Photochemical model performance for $PM_{2.5}$ sulfate, nitrate, ammonium, and precursor species $SO_2$, $HNO_3$, and $NH_3$ at background monitor locations in the central and eastern United States. Atmospheric Environment 41, 6185–6195.

Baker, K., Scheff, P., 2008. Assessing meteorological variable and process relationships to modeled $PM_{2.5}$ ammonium nitrate and ammonium sulfate in the central United States. Journal of Applied Meteorology and Climatology 47, 2395–2404.

Baker, K.R., Bash, J.O., 2012. Regional scale photochemical model evaluation of total mercury wet deposition and speciated ambient mercury. Atmospheric Environment 49, 151–162.

Baker, K.R., Simon, H., Kelly, J.T., 2011. Challenges to modeling "Cold Pool" meteorology associated with high pollution episodes. Environmental Science & Technology 45, 7118–7119.

Bigelow, D.S., 1991. Comparability of wet-only precipitation chemistry measurements from the United States National Atmospheric Deposition Program (NADP) to those of the Canadian network for sampling acid precipitation (CANSAP). Environmental Science & Technology 25, 1867–1875.

Boylan, J.W., Russell, A.G., 2006. PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. Atmospheric Environment 40, 4946–4959.

Bullock, O.R., Atkinson, D., Braverman, T., Civerolo, K., Dastoor, A., Davignon, D., Ku, J.Y., Lohman, K., Myers, T.C., Park, R.J., Seigneur, C., Selin, N.E., Sistla, G., Vijayaraghavan, K., 2009. An analysis of simulated wet deposition of mercury from the North American Mercury Model Intercomparison Study. Journal of Geophysical Research-Atmospheres 114, D08301.

Byun, D.W., Kim, S.-T., Kim, S.-B., 2007. Evaluation of air quality models for the simulation of a high ozone episode in the Houston metropolitan area. Atmospheric Environment 41, 837–853.

Carlton, A.G., Baker, K.R., 2011. Photochemical modeling of the Ozark isoprene volcano: MEGAN, BEIS, and their impacts on air quality predictions. Environmental Science & Technology 45, 4438–4445.

Carlton, A.G., Bhave, P.V., Napelenok, S.L., Edney, E.D., Sarwar, G., Pinder, R.W., Pouliot, G.A., Houyoux, M., 2010. Model representation of secondary organic aerosol in CMAQv4.7. Environmental Science & Technology 44, 8553–8560.

Chen, J., Vaughan, J., Avise, J., O'Neill, S., Lamb, B., 2008. Enhancement and evaluation of the AIRPACT ozone and $PM_{2.5}$ forecast system for the Pacific Northwest. Journal of Geophysical Research-Atmospheres 113, D14305.

Cho, S., Makar, P.A., Lee, W.S., Herage, T., Liggio, J., Li, S.M., Wiens, B., Graham, L., 2009. Evaluation of a unified regional air-quality modeling system (AURAMS) using PrAIRie2005 field study data: the effects of emissions data accuracy on particle sulphate predictions. Atmospheric Environment 43, 1864–1877.

Chow, J.C., Watson, J.G., Crow, D., Lowenthal, D.H., Merrifield, T., 2001. Comparison of IMPROVE and NIOSH carbon measurements. Aerosol Science and Technology 34, 23–34.

de Gouw, J., Jimenez, J.L., 2009. Organic aerosols in the Earth's atmosphere. Environmental Science & Technology 43, 7614–7618.

Docherty, K.S., Stone, E.A., Ulbrich, I.M., DeCarlo, P.F., Snyder, D.C., Schauer, J.J., Peltier, R.E., Weber, R.J., Murphy, S.M., Seinfeld, J.H., Grover, B.D., Eatough, D.J., Jimenez, J.L., 2008. Apportionment of primary and secondary organic aerosols in southern California during the 2005 Study of Organic Aerosols in Riverside (SOAR-1). Environmental Science & Technology 42, 7655–7662.

Dunlea, E.J., Herndon, S.C., Nelson, D.D., Volkamer, R.M., San Martini, F., Sheehy, P.M., Zahniser, M.S., Shorter, J.H., Wormhoudt, J.C., Lamb, B.K., Allwine, E.J., Gaffney, J.S., Marley, N.A., Grutter, M., Marquez, C., Blanco, S., Cardenas, B., Retama, A., Villegas, C.R.R., Kolb, C.E., Molina, L.T., Molina, M.J., 2007. Evaluation of nitrogen dioxide chemiluminescence monitors in a polluted urban environment. Atmospheric Chemistry and Physics 7, 2691–2704.

Eder, B., Kang, D., Mathur, R., Pleim, J., Yu, S., Otte, T., Pouliot, G., 2009. A performance evaluation of the National Air Quality Forecast Capability for the summer of 2007. Atmospheric Environment 43, 2312–2320.

Eder, B., Kang, D., Mathur, R., Yu, S., Schere, K., 2006. An operational evaluation of the Eta-CMAQ air quality forecast model. Atmospheric Environment 40, 4894–4905.

Eder, B., Yu, S., 2006. A performance evaluation of the 2004 release of Models-3 CMAQ. Atmospheric Environment 40, 4811–4824.

El-Zanan, H.S., Lowenthal, D.H., Zielinska, B., Chow, J.C., Kumar, N., 2005. Determination of the organic aerosol mass to organic carbon ratio in IMPROVE samples. Chemosphere 60, 485–496.

ENVIRON, 2010. User's Guide Comprehensive Air Quality Model with Extensions. ENVIRON International Corporation, Novato, California.

Foley, K.M., Roselle, S.J., Appel, K.W., Bhave, P.V., Pleim, J.E., Otte, T.L., Mathur, R., Sarwar, G., Young, J.O., Gilliam, R.C., Nolte, C.G., Kelly, J.T., Gilliland, A.B., Bash, J.O., 2010. Incremental testing of the Community Multiscale Air Quality (CMAQ) modeling system version 4.7. Geoscientific Model Development 3, 205–226.

Frank, N.H., 2006. Retained nitrate, hydrated sulfates, and carbonaceous mass in Federal Reference Method fine particulate matter for six eastern US cities. Journal of the Air & Waste Management Association 56, 500–511.

Gaydos, T.M., Pinder, R., Koo, B., Fahey, K.M., Yarwood, G., Pandis, S.N., 2007. Development and application of a three-dimensional aerosol chemical transport model, PMCAMx. Atmospheric Environment 41, 2594–2611.

Gego, E., Porter, P.S., Hogrefe, C., Irwin, J.S., 2006. An objective comparison of CMAQ and REMSAD performances. Atmospheric Environment 40, 4920–4934.

Goldstein, A.H., Galbally, I.E., 2007. Known and unexplored organic constituents in the earth's atmosphere. Environmental Science & Technology 41, 1514–1521.

Gong, W., Dastoor, A.P., Bouchet, V.S., Gong, S., Makar, P.A., Moran, M.D., Pabla, B., Menard, S., Crevier, L.-P., Cousineau, S., Venkatesh, S., 2006. Cloud processing of gases and aerosols in a regional air quality model (AURAMS). Atmospheric Research 82, 248–275.

Gorline, J.L., Lee, P., 2009. Performance evaluation of NOAA-EPA developmental aerosol forecasts. Environmental Fluid Mechanics 9, 109–120.

Grell, G.A., Emeis, S., Stockwell, W.R., Schoenemeyer, T., Forkel, R., Michalakes, J., Knoche, R., Seidl, W., 2000. Application of a multiscale, coupled MM5/chemistry model to the complex terrain of the VOTALP valley campaign. Atmospheric Environment 34, 1435–1453.

Grell, G.A., Peckham, S.E., Schmitz, R., McKeen, S.A., Frost, G., Skamarock, W.C., Eder, B., 2005. Fully coupled "online" chemistry within the WRF model. Atmospheric Environment 39, 6957–6975.

Hogrefe, C., Civerolo, K.L., Hao, W., Ku, J.-Y., Zalewsky, E.E., Sistla, G., 2008. Rethinking the assessment of photochemical modeling systems in air quality planning applications. Journal of the Air & Waste Management Association 58, 1086–1099.

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

Hogrefe, C., Hao, W., Zalewsky, E.E., Ku, J.Y., Lynn, B., Rosenzweig, C., Schultz, M.G., Rast, S., Newchurch, M.J., Wang, L., Kinney, P.L., Sistla, G., 2011. An analysis of long-term regional-scale ozone simulations over the Northeastern United States: variability and trends. Atmospheric Chemistry and Physics 11, 567–582.

Jin, L., Brown, N.J., Harley, R.A., Bao, J.-W., Michelson, S.A., Wilczak, J.M., 2010. Seasonal versus episodic performance evaluation for an Eulerian photochemical air quality model. Journal of Geophysical Research-Atmospheres 115, D09302.

Kang, D., Mathur, R., Rao, S.T., 2010. Real-time bias-adjusted $O_3$ and $PM_{2.5}$ air quality index forecasts and their performance evaluations over the continental United States. Atmospheric Environment 44, 2203–2212.

Karamchandani, P., Vijayaraghavan, K., Chen, S.-Y., Seigneur, C., Edgerton, E.S., 2006. Plume-in-grid modeling for particulate matter. Atmospheric Environment 40, 7280–7297.

Kim, S.W., Heckel, A., Frost, G.J., Richter, A., Gleason, J., Burrows, J.P., McKeen, S., Hsie, E.Y., Granier, C., Trainer, M., 2009. $NO_2$ columns in the western United States observed from space and simulated by a regional chemistry model and their implications for $NO_x$ emissions. Journal of Geophysical Research-Atmospheres 114, D11301.

Kim, Y., Fu, J.S., Miller, T.L., 2010. Improving ozone modeling in complex terrain at a fine grid resolution – part II: influence of schemes in MM5 on daily maximum 8-h ozone concentrations and RRFs (Relative Reduction Factors) for SIPs in the non-attainment areas. Atmospheric Environment 44, 2116–2124.

Kleeman, M.J., Cass, G.R., 2001. A 3D Eulerian source-oriented model for an externally mixed aerosol. Environmental Science & Technology 35, 4834–4848.

Lee, D., Byun, D.W., Kim, H., Ngan, F., Kim, S., Lee, C., Cho, C., 2011. Improved CMAQ predictions of particulate matter utilizing the satellite-derived aerosol optical depth. Atmospheric Environment 45, 3730–3741.

Lin, C.J., Pongprueks, P., Rusell Bulock, O., Lindberg, S.E., Pehkonen, S.O., Jang, C., Braverman, T., Ho, T.C., 2007. Scientific uncertainties in atmospheric mercury models II: sensitivity analysis in the CONUS domain. Atmospheric Environment 41, 6544–6560.

Lin, C.J., Shetty, S.K., Pan, L., Pongprueksa, P., Jang, C., Chu, H., 2012. Source attribution for mercury deposition in the contiguous United States: regional difference and seasonal variation. Journal of the Air & Waste Management Association 62, 52–63.

Liu, X.-H., Zhang, Y., Olsen, K.M., Wang, W.-X., Do, B.A., Bridgers, G.M., 2010. Responses of future air quality to emission controls over North Carolina, part I: model evaluation for current-year simulations. Atmospheric Environment 44, 2443–2456.

Makar, P.A., Gong, W., Mooney, C., Zhang, J., Davignon, D., Samaali, M., Moran, M.D., He, H., Tarasick, D.W., Sills, D., Chen, J., 2010. Dynamic adjustment of climatological ozone boundary conditions for air-quality forecasts. Atmospheric Chemistry and Physics 10, 8997–9015.

Malm, W.C., Hand, J.L., 2007. An examination of the physical and optical properties of aerosols collected in the IMPROVE program. Atmospheric Environment 41, 3407–3427.

Marmur, A., Liu, W., Wang, Y., Russell, A.G., Edgerton, E.S., 2009. Evaluation of model simulated atmospheric constituents with observations in the factor projected space: CMAQ simulations of SEARCH measurements. Atmospheric Environment 43, 1839–1849.

Misenis, C., Zhang, Y., 2010. An examination of sensitivity of WRF/Chem predictions to physical parameterizations, horizontal grid spacing, and nesting options. Atmospheric Research 97, 315–334.

Molders, N., Porter, S.E., Cahill, C.F., Grell, G.A., 2010. Influence of ship emissions on air quality and input of contaminants in southern Alaska National Parks and Wilderness Areas during the 2006 tourist season. Atmospheric Environment 44, 1400–1413.

Morris, R.E., Koo, B., Guenther, A., Yarwood, G., McNally, D., Tesche, T.W., Tonnesen, G., Boylan, J., Brewer, P., 2006. Model sensitivity evaluation for organic carbon using two multi-pollutant air quality models that simulate regional haze in the southeastern United States. Atmospheric Environment 40, 4960–4972.

Napelenok, S.L., Foley, K.M., Kang, D., Mathur, R., Pierce, T., Rao, S.T., 2011. Dynamic evaluation of regional air quality model's response to emission reductions in the presence of uncertain emission inventories. Atmospheric Environment 45, 4091–4098.

Otte, T.L., 2008. The impact of nudging in the meteorological model for retrospective air quality simulations. Part I: evaluation against national observation networks. Journal of Applied Meteorology and Climatology 47, 1853–1867.

Park, S.H., Gong, S.L., Gong, W., Makar, P.A., Moran, M.D., Zhang, J., Stroud, C.A., 2010. Relative impact of windblown dust versus anthropogenic fugitive dust in $PM_{2.5}$ on air quality in North America. Journal of Geophysical Research-Atmospheres 115, D16210.

Pudykiewicz, J.A., Koziol, A.S., 2001. The application of Eulerian models for air quality prediction and the evaluation of emission control strategies in Canada. International Journal of Environment and Pollution 16, 425–438.

Pun, B.K., Seigneur, C., Vijayaraghavan, K., Wu, S.Y., Chen, S.Y., Knipping, E.M., Kumar, N., 2006. Modeling regional haze in the BRAVO study using CMAQ-MADRID: 1. Model evaluation. Journal of Geophysical Research-Atmospheres 111, D06302.

Queen, A., Zhang, Y., 2008a. Examining the sensitivity of MM5-CMAQ predictions to explicit microphysics schemes and horizontal grid resolutions, part II – PM concentrations and wet deposition predictions. Atmospheric Environment 42, 3856–3868.

Queen, A., Zhang, Y., 2008b. Examining the sensitivity of MM5-CMAQ predictions to explicit microphysics schemes and horizontal grid resolutions, part III – the impact of horizontal grid resolution. Atmospheric Environment 42, 3869–3881.

Rodriguez, M.A., Barna, M.G., Gebhart, K.A., Hand, J.L., Adelman, Z.E., Schichtel, B.A., Collett Jr., J.L., Malm, W.C., 2011. Modeling the fate of atmospheric reduced nitrogen during the Rocky Mountain Atmospheric Nitrogen and Sulfur Study (RoMANS): performance evaluation and diagnosis using integrated processes rate analysis. Atmospheric Environment 45, 223–234.

Rodriguez, M.A., Barna, M.G., Moore, T., 2009. Regional impacts of oil and gas development on ozone formation in the western United States. Journal of the Air & Waste Management Association 59, 1111–1118.

Roy, B., Mathur, R., Gilliland, A.B., Howard, S.C., 2007. A comparison of CMAQ-based aerosol properties with IMPROVE, MODIS, and AERONET data. Journal of Geophysical Research-Atmospheres 112, D14301.

SAI, 2002. Regional Modeling System for Aerosols and Deposition (REMSAD). Systems Applications International, San Rafael, CA.

Sakulyanontvittaya, T., Guenther, A., Helmig, D., Milford, J., Wiedinmyer, C., 2008. Secondary organic aerosol from sesquiterpene and monoterpene emissions in the United States. Environmental Science & Technology 42, 8784–8790.

Seigneur, C., Lohman, K., Vijayaraghavan, K., Jansen, J., Levin, L., 2006. Modeling atmospheric mercury deposition in the vicinity of power plants. Journal of the Air & Waste Management Association 56, 743–751.
Simon, H., Bhave, P.V., Swall, J.L., Frank, N.H., Malm, W.C., 2011. Determining the spatial and seasonal variability in OM/OC ratios across the US using multiple regression. Atmospheric Chemistry and Physics 11, 2933–2949.
Simpson, D., Yttri, K.E., Klimont, Z., Kupiainen, K., Caseiro, A., Gelencser, A., Pio, C., Puxbaum, H., Legrand, M., 2007. Modeling carbonaceous aerosol over Europe: analysis of the CARBOSOL and EMEP EC/OC campaigns. Journal of Geophysical Research-Atmospheres 112, D23s14.
Slemr, F., Baumbach, G., Blank, P., Corsmeier, U., Fiedler, F., Friedrich, R., Habram, M., Kalthoff, N., Klemp, D., Kuhlwein, J., Mannschreck, K., Mollmann-Coers, M., Nester, K., Panitz, H.J., Rabl, P., Slemr, J., Vogt, U., Wickert, B., 2002. Evaluation of modeled spatially and temporarily highly resolved emission inventories of photosmog precursors for the city of Augsburg: the experiment EVA and its major results. Journal of Atmospheric Chemistry 42, 207–233.
Smyth, S.C., Jiang, W., Roth, H., Moran, M.D., Makar, P.A., Yang, F., Bouchet, V.S., Landry, H., 2009. A comparative performance evaluation of the AURAMS and CMAQ air-quality modelling systems. Atmospheric Environment 43, 1059–1070.
Smyth, S.C., Jiang, W.M., Yin, D.Z., Roth, H., Giroux, T., 2006. Evaluation of CMAQ O3 and $PM_{2.5}$ performance using Pacific 2001 measurement data. Atmospheric Environment 40, 2735–2749.
Spak, S.N., Holloway, T., 2009. Seasonality of speciated aerosol transport over the Great Lakes region. Journal of Geophysical Research-Atmospheres 114, D08302.
Stroud, C.A., Makar, P.A., Moran, M.D., Gong, W., Gong, S., Zhang, J., Hayden, K., Mihele, C., Brook, J.R., Abbatt, J.P.D., Slowik, J.G., 2011. Impact of model grid spacing on regional- and urban- scale air quality predictions of organic aerosol. Atmospheric Chemistry and Physics 11, 3107–3118.
Tang, W., Cohan, D.S., Morris, G.A., Byun, D.W., Luke, W.T., 2011. Influence of vertical mixing uncertainties on ozone simulation in CMAQ. Atmospheric Environment 45, 2898–2909.
Tarasick, D.W., Moran, M.D., Thompson, A.M., Carey-Smith, T., Rochon, Y., Bouchet, V.S., Gong, W., Makar, P.A., Stroud, C., Menard, S., Crevier, L.P., Cousineau, S., Pudykiewicz, J.A., Kallaur, A., Moffet, R., Menard, R., Robichaud, A., Cooper, O.R., Oltmans, S.J., Witte, J.C., Forbes, G., Johnson, B.J., Merrill, J., Moody, J.L., Morris, G., Newchurch, M.J., Schmidlin, F.J., Joseph, E., 2007. Comparison of Canadian air quality forecast models with tropospheric ozone profile measurements above midlatitude North America during the IONS/ICARTT campaign: evidence for stratospheric input. Journal of Geophysical Research-Atmospheres 112, D12s22.
Tesche, T.W., 1988. Accuracy of ozone air-quality models. Journal of Environmental Engineering-ASCE 114, 739–752.
Tesche, T.W., Morris, R., Tonnesen, G., McNally, D., Boylan, J., Brewer, P., 2006. CMAQ/CAMx annual 2002 performance evaluation over the eastern US. Atmospheric Environment 40, 4906–4919.
Tong, D.Q., Mauzerall, D.L., 2006. Spatial variability of summertime tropospheric ozone over the continental United States: implications of an evaluation of the CMAQ model. Atmospheric Environment 40, 3041–3056.
Turpin, B.J., Lim, H.J., 2001. Species contributions to $PM_{2.5}$ mass concentrations: revisiting common assumptions for estimating organic mass. Aerosol Science and Technology 35, 602–610.
United States Environmental Protection Agency, 2007. Guidance on the Use of Models and Other Analyses for Demonstrating Attainment of Air Quality Goals for Ozone, $PM_{2.5}$, and Regional Haze, Research Triangle Park. EPA-454/B-07–002.
Vermette, S., Lindberg, S., Bloom, N., 1995. Field-tests for a regional mercury deposition network – sampling design and preliminary test-results. Atmospheric Environment 29, 1247–1251.
Vijayaraghavan, K., Karamchandani, P., Seigneur, C., Balmori, R., Chen, S.-Y., 2008. Plume-in-grid modeling of atmospheric mercury. Journal of Geophysical Research-Atmospheres 113, D24305.
Vijayaraghavan, K., Seigneur, C., Karamchandani, P., Chen, S.Y., 2007. Development and application of a multipollutant model for atmospheric mercury deposition. Journal of Applied Meteorology and Climatology 46, 1341–1353.
Volkamer, R., Jimenez, J.L., San Martini, F., Dzepina, K., Zhang, Q., Salcedo, D., Molina, L.T., Worsnop, D.R., Molina, M.J., 2006. Secondary organic aerosol formation from anthropogenic air pollution: rapid and higher than expected. Geophysical Research Letters 33, L17811.
Wu, S.-Y., Krishnan, S., Zhang, Y., Aneja, V., 2008. Modeling atmospheric transport and fate of ammonia in North Carolina – part I: evaluation of meteorological and chemical predictions. Atmospheric Environment 42, 3419–3436.
Ying, Q., Lu, J., Allen, P., Livingstone, P., Kaduwela, A., Kleeman, M., 2008. 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. Atmospheric Environment 42, 8954–8966.
Yu, S., Mathur, R., Kang, D., Schere, K., Eder, B., Pleim, J., 2006. Performance and diagnostic evaluation of ozone predictions by the Eta-community multiscale air quality forecast system during the 2002 New England Air Quality Study. Journal of the Air & Waste Management Association 56, 1459–1471.
Yu, S., Mathur, R., Schere, K., Kang, D., Pleim, J., Otte, T.L., 2007. A detailed evaluation of the Eta-CMAQ forecast model performance for $O_3$, its related precursors, and meteorological parameters during the 2004 ICARTT study. Journal of Geophysical Research-Atmospheres 112, D12s14.
Yu, S., Mathur, R., Schere, K., Kang, D., Pleim, J., Young, J., Tong, D., Pouliot, G., McKeen, S.A., Rao, S.T., 2008. Evaluation of real-time $PM_{2.5}$ forecasts and process analysis for $PM_{2.5}$ formation over the eastern United States using the Eta-CMAQ forecast model during the 2004 ICARTT study. Journal of Geophysical Research-Atmospheres 113, D06204.
Zhang, L.M., Moran, M.D., Makar, P.A., Brook, J.R., Gong, S.L., 2002. Modelling gaseous dry deposition in AURAMS: a unified regional air-quality modelling system. Atmospheric Environment 36, 537–560.
Zhang, Y., Huang, J.-P., Henze, D.K., Seinfeld, J.H., 2007. Role of isoprene in secondary organic aerosol formation on a regional scale. Journal of Geophysical Research-Atmospheres 112, D20207.
Zhang, Y., Liu, P., Queen, A., Misenis, C., Pun, B., Seigneur, C., Wu, S.-Y., 2006. A comprehensive performance evaluation of MM5-CMAQ for the Summer 1999 Southern Oxidants Study episode – part II: gas and aerosol predictions. Atmospheric Environment 40, 4839–4855.
Zhang, Y., Pan, Y., Wang, K., Fast, J.D., Grell, G.A., 2010. WRF/Chem-MADRID: incorporation of an aerosol module into WRF/Chem and its initial application to the TexAQS2000 episode. Journal of Geophysical Research-Atmospheres 115, D18202.
Zhang, Y., Vijayaraghavan, K., Wen, X.-Y., Snell, H.E., Jacobson, M.Z., 2009. Probing into regional ozone and particulate matter pollution in the United States: 1. A 1 year CMAQ simulation and evaluation using surface and satellite data. Journal of Geophysical Research-Atmospheres 114, D22304.