Reviewer 1 comment	Response
Overall I found the manuscript to be generally well written and organized.	We thank the reviewer for this comment.
As is typically the case when attempting to evaluate model simulations that span large spatial domains and time periods, the difficulty becomes in summarizing the results in a meaningful way that does not overwhelm the reader with statistics and numbers. Here, the authors present annual mean performance metrics for the entire domain, along with regional/seasonal statistics. I generally don't find the annual statistics to be helpful in any way, other than perhaps initially to make sure there isn't some huge gross error in the model results. Otherwise, bulk annual/domain-wide statistics are typically difficult to interpret due to often compensating seasonal biases (e.g. particulate nitrate is often underestimated in the summer and overestimated in the winter). To their credit, the authors do acknowledge this issue with the bulk statistics. I'm wondering if the manuscript would benefit from dropping the annual domain-wide statistics and just focus on presenting the seasonal and regional statistics. I will leave this decision to the authors, but just note that I think most readers would find a much value in the annual/domain-wide stats and would immediately focus on the seasonal/regional stats.	We agree with the review that the seasonal and regional statistics are important, but we agree with Reviewer 2 that the annual/domain-wide statistics are also useful: they show an overall summary of model performance. For this reason we would like to keep the annual/domain-wide statistics in the main manuscript.
It might be nice to move some of the seasonal/regional plots for the speciated PM2.5 components from the supplemental material to the main text.	This is a good idea. We have moved the seasonal/regional scatterplots for PM2.5 subspecies to the main text.
Finally, the authors need to support some of their statements with references, specifically regarding difference in sampling protocols and/or analysis techniques between the different networks.	In response to this comment, we have added a citation of the variability in OC analysis methods. We also fixed two inconsistencies in our processing of the data, the result of which that some of this text was no longer relevant and so was removed.
Abstract: Perhaps mention the modeling year earlier in the abstract.	We added "year 2005" to the first sentence of the abstract.
Provide some examples of "contemporary models".	We thank the reviewer for this suggestion. We provide examples of contemporary models and their performance in the main text. (Adding examples in the abstract (and defining the lengthy acronyms that make up their names) would add to the word count and distract from the main messages of the abstract.) In the abstract, we changed "contemporary models" to "contemporary modeling efforts" to better reflect the comparisons

	that we do in the manuscript.
Again, bulk annual average statistics are not all the useful. Maybe replace these with more meaningful seasonal/regional metrics.	We respect this viewpoint. As mentioned above, we feel that both types of statistics are useful.
It's a little strange to look at 24h average ozone, given the large biases that typically can occur with ozone overnight. It might be better to present a different, more meaningful metric for ozone here (e.g. daily 8hr average maximum).	In response to this comment, we clarified the abstract text to state that average daytime and daily peak concentrations are more relevant for health effects and regulatory analysis, and the model performance is better for those metrics.
Page 8435, lines 13-15: It might be a little disingenuous to refer to 12-km as "finescale". Understanding that scale is relative thing (15 years ago, 12-km was "fine-scale"), 12-km is probably better referred to as regional-scale at this point in time, considering that more and more modeling is taking place at 4-km and below.	In response to this comment, we changed the text "fine-scale (12 km or better)" to "12 km or finer scale".
Page 8436, line 21: 28 layers seems like it's on the low-end of layer structures these days. Were the computer limitations the deciding factor in going with 28 layers instead of something closer to say 40 or even 50? Do the authors feel that increasing the number of vertical layers (and in particular using the smaller first layer) would significantly impact the results?	We thank the reviewer for this comment. In response, we added the text "Previous studies (e.g., Appel et al., 2012; Yahya et al., 2014) have used 34 vertical layers; our choice of 28 vertical layers represents a tradeoff between vertical grid resolution and computational expense". We have not investigated the question of how the results would be impacted by increasing or decreasing the number of vertical layers; that issue is important and worthy of further consideration but for the present article is outside the scope of our investigation.
Page 8437, line 13: What exactly constitutes "miscellaneous PM2.5"?	In response to this comment, we changed "miscellaneous PM2.5" to "unclassified PM2.5".
Page 8438, lines 7-9: The 2008 NEI has been available for quite some time now (and 2011 NEI is now available too). It seems like 2005 is a fairly old year to simulate at this point. When the authors say that the 2005 NEI was most recent available it makes it seem like this work started a long time ago. Has it just taken that long from start to finish for this modeling exercise?	We thank the reviewer for this question. This manuscript is part of a larger modeling exercise, which has taken a number of years to complete. The other part of this study was recently published here: http://www.pnas.org/content/111/52/18490.abstract
Page 8439, Line 23: A 50-60 meter first layer height seems quite large, especially since nighttime boundary layers can often reach 50m or below. What impact do the authors feel there is from having such a deep first layer?	Testing the impact of the number of layers on model performance is outside of the scope of this study. We note in the Discussion that the investigation of model parameters on performance is an important area for future research.
Page 8445, Lines 9-10: Exactly what differences are there between the network measurement techniques and why would they result in such larger differences? IMPROVE sites are rural, so perhaps background SO2/SO4 is greatly overestimated.	In response to this comment, we revisited the measurement data documentation and found that the IMPROVE network reports elemental sulfur concentrations instead of SO_4 concentrations. Adjusting our calculations to account for this decreases the differences between measurement networks for SO_4.

Page 8445: First, the authors state a MFB = -110%. What is does that statistic represent, since later in the paragraph the authors state a contiguous US MFB = -120%?	We thank the reviewer for calling this to our attention. -110% refers to the bias in annual average predictions, whereas -120% refers to wintertime predictions. Since the 10% difference is probably not large enough to warrant discussing both statistics separately, we have removed the mention of the wintertime statistics and clarified that the -110% is for the annual average.
The nitrate biases reported are really large. Do the authors have any explanation as to why nitrate is underpredicted by so much (especially in the west where nitrate makes up a greater percentage of the total PM2.5 than in the east)?	Particulate nitrate formation is strongly temperature dependent, and as we discuss in the article, many model performance evaluations only cover the summer months. We state in the article that nitrate predictive performance is better in the summer than in the winter. In Table A2 we compare our results to another full-year, contiguous U.S. modeling simulation. Predictive performance for nitrate in that study is similar to our results.
Page 8446: The OC underestimation at CSN sites is really large too. How is it that the differences don't appear to be rural vs. urban, since the urban CSN sites have an OC MFB = -113%, but the IMPROVE sites have an MFB = 15%)? That seems indicative of an urban emissions problem (or possibly meteorological, or both). I'd really like to know how those large differences are the result of simply sampling or analysis. References are needed if the authors are going to make statements like that.	In response to this comment we reviewed our calculations and found and fixed a configuration error which was partially responsible for the difference between networks. As noted in the text, figure A12 shows that the difference between networks in similar when considering only urban vs. only rural locations. We have added a reference that discusses the variability in measured values of OC when using different analysis methods, which can be up a factor of 5.
Page 8446, Line 26: Change "lower" to "worse".	We thank the reviewer for this comment, but we think that he or she may have misinterpreted Table A2. We state that for most pollutants and networks, NME is lower in our study than in Yahya et al. The numbers in Table A2 support this statement. Since lower error is generally considered to be better than higher error, we don't feel that it would be appropriate to change "lower" to "worse". To clarify this, we changed the text to "lower (i.e., better)".
Table A2: Are these annual values being reported?	We edited the title of Table A2 to clarify that the values are for annual average performance.
Reviewer 2 comment	**Response**
The manuscript is well written and exhaustive, and will provide an excellent reference for future studies using WRF-Chem at 12 km resolution.	We thank the reviewer for this comment.
I do not have any major comment on the manuscript and I think that it could be published as is.	We thank the reviewer for this comment.

[I] agree with reviewer number 1 that seasonal statistics are more useful than annual statistics, but I do find annual statistics interesting to get a first idea of the model performances.	We agree that both sets of statistics are useful.
I suggest to add some maps of the different species making up for PM2.5 to the main text.	As mentioned above, we have added the figures for PM2.5 subspecies to the main text.
Some species are more sensitive to emission errors, other to scavenging efficiency, others result from chemistry, so comparing these measurements can give an idea of what is causing the bias.	We agree that these comparisons are useful, and we have tried to suggest possible reasons for the model errors we observer wherever possible.

Manuscript prepared for Geosci. Model Dev. Discuss.
with version 2014/09/16 7.15 Copernicus papers of the LaTeX class copernicus.cls.
Date: 23 February 2015

Twelve-month, 12 km resolution North American WRF-Chem v3.4 air quality simulation: performance evaluation

C. W. Tessum[1], J. D. Hill[2], and J. D. Marshall[1]

[1]Department of Civil, Environmental, and Geo- Engineering, University of Minnesota, Minneapolis, Minnesota, USA
[2]Department of Bioproducts and Biosystems Engineering, University of Minnesota, St. Paul, Minnesota, USA

Correspondence to: J. D. Marshall (julian@umn.edu)

Abstract

We present results from and evaluate the performance of a 12 month, 12 km horizontal resolution year 2005 air pollution simulation for the contiguous United States using the WRF-Chem (Weather Research and Forecasting with Chemistry) meteorology and chemical transport model (CTM). We employ the 2005 US National Emissions Inventory, the Regional Atmospheric Chemistry Mechanism (RACM), and the Modal Aerosol Dynamics Model for Europe (MADE) with a Volatility Basis Set (VBS) secondary aerosol module. Overall, model performance is comparable to contemporary ~~models~~ modelling efforts used for regulatory and health-effects analysis, with an annual average daytime ozone (O_3) mean fractional bias (MFB) of 12 % and an annual average fine particulate matter ($PM_{2.5}$) MFB of −1 %. WRF-Chem, as configured here, tends to overpredict total $PM_{2.5}$ at some high concentration locations, and generally overpredicts average 24 h O_3 concentrations~~, with better performance at predicting average daytime~~ . Performance is better at predicting daytime-average and daily peak O_3 concentrations, which are more relevant for regulatory and health effects analyses relative to annual average values. Predictive performance for $PM_{2.5}$ subspecies is mixed: the model overpredicts particulate sulfate (MFB = ~~65~~ = 36 %), underpredicts particulate nitrate (MFB = −110 %) and organic carbon (MFB = −~~65~~29 %), and relatively accurately predicts particulate ammonium (MFB = 3 %) and elemental carbon (MFB = 3 %), so that the accuracy in total $PM_{2.5}$ predictions is to some extent a function of offsetting over- and underpredictions of $PM_{2.5}$ subspecies. Model predictive performance for $PM_{2.5}$ and its subspecies is in general worse in winter and in the western US than in other seasons and regions, suggesting spatial and temporal opportunities for future WRF-Chem model development and evaluation.

1 Introduction

Epidemiological studies have established the importance of health effects from acute and chronic exposure to fine particulate matter ($PM_{2.5}$) and ground-level ozone (O_3) (Jerrett

et al., 2009; Krewski et al., 2009; Pope and Dockery, 2006). The accuracy of health-impact predictions for future air pollutant emissions (e.g., Tessum et al., 2012; Tessum et al., 2014) depends in part on the performance of air quality models over long time scales and in all seasons. Accurate health-impact predictions often depend on model simulations that cover large geographic areas such as the contiguous US, so as to capture the full impacts of the long-range transport of pollutants (Levy et al., 2003). Whereas chemical transport model (CTM) simulations for a full year for the contiguous US often use 36 km horizontal grids (e.g., Tesche et al., 2006; Yahya et al., 2014), increasing horizontal grid resolution to 12 km can result in the more accurate prediction of pollutant concentrations (Fountoukis, 2013) and population exposure. However, increasing horizontal resolution from 36 to 12 km in a CTM typically results in a $\sim 27\times$ increase in computational intensity (number of grid cells increases nine-fold; number of time steps increases three-fold).

Although recent CTM evaluation efforts have focused on 12 month and contiguous US model evaluations (Galmarini et al., 2012), CTM model performance for ~~fine-scale horizontal grid size (~~12 km or ~~better)~~ finer horizontal grid size for an entire year for the entire contiguous US is largely unexplored in the peer-reviewed literature. We know of only one such study: Appel et al. (2012) evaluated the performance of the Community Multiscale Air Quality (CMAQ) model (Foley et al., 2010) in reproducing year 2006 concentrations of $PM_{2.5}$ and O_3 for the contiguous US. In a second study (not peer reviewed), the US EPA (2012) describes model evaluation for $PM_{2.5}$ concentrations for year 2007, also for the contiguous US and using CMAQ. Our study contributes to this literature by evaluating a different model with different parameterizations over a different time period. We also provide greater investigation regarding how model performance varies in space, in time, and by chemical species.

We employ and evaluate the performance of WRF-Chem (the Weather Research and Forecasting model with Chemistry) (Grell et al., 2005) for year 2005 for a North American domain. WRF-Chem is functionally similar to CMAQ, but differs from the version used by Appel et al. (2012) in that WRF-Chem predicts meteorological quantities and air pollution concentrations simultaneously, allowing meteorology quantities to be updated more

frequently as the model is running and allowing representation of interactions between meteorology and air pollution. WRF-Chem users can follow a simplified modeling workflow that does not require running a separate meteorological model. This aspect can be beneficial for the modeler, not necessarily for the model's computation demands. For the domain and settings used here, meteorological modeling accounts for only $\sim 10\,\%$ of the total computational expense.

Table A1 summarizes spatial and temporal aspects of recent chemical transport model evaluation efforts, with a focus on WRF-Chem evaluations in the US. WRF-Chem performance in predicting air quality observations has been extensively quantified for simulations of individual regions of the US, with simulation periods of several weeks or months (Ahmadov et al., 2012; Chuang et al., 2011; Fast et al., 2005; Grell et al., 2005; McKeen et al., 2007; Misenis and Zhang, 2010; Zhang et al., 2010, 2012). One study evaluated WRF-Chem performance for a full year for the contiguous US with a 36 km grid (Yahya et al., 2014). We present here WRF-Chem results from a full year, 12 km resolution simulation for the contiguous US, evaluate the performance of the model compared to ambient measurements, and compare WRF-Chem performance to published goals and criteria (Boylan and Russell, 2006) and to recent CMAQ results for a similar simulation (Appel et al., 2012).

2 Methods

2.1 Model setup

We run the WRF-Chem model version 3.4 using a 12 km resolution grid with 444 rows, 336 columns, and 28 vertical layers. The modeling domain (see Fig. 1) covers the contiguous US, southern Canada, and northern Mexico. Previous studies (e.g., Appel et al., 2012; Yahya et al., 2014) have used 34 vertical layers; our choice of 28 vertical layers represents a tradeoff between vertical grid resolution and computational expense.

Within WRF-Chem, we use the Regional Atmospheric Chemistry Mechanism (RACM) (Stockwell et al., 1997) for gas-phase reactions and the Modal Aerosol Dynamics for Eu-

rope (MADE) (Ackermann et al., 1998) module for aerosol chemistry and physics. RACM and MADE were selected because of their relatively modest computational expense; at the time of this study, alternatives to RACM/MADE are impractical for large-scale simulations such as ours. We use the Volatility Basis Set (VBS) (Ahmadov et al., 2012) to simulate formation and evaporation of secondary organic aerosol (SOA). The VBS approach differs from other SOA parameterizations in that it assumes that primary organic aerosol (POA) is semi-volatile. Meteorology options are set as recommended by the WRF user manual (Wang et al., 2012) and the WRF-Chem user manual (Peckham et al., 2012) for situations similar to those studied here. Table 1 summarizes the model options and inputs used. See supporting information for additional details.

We use results from the MOZART global chemical transport model (Emmons et al., 2010) as processed by the MOZBC file format converter (available: http://web3.acd.ucar.edu/wrf-chem) to provide initial and boundary conditions for chemical species. Because the MOZBC boundary conditions for ~~miscellaneous~~ unclassified $PM_{2.5}$ are unrealistic for the southeastern edges of the modeling domain – their use results in substantial $PM_{2.5}$ over-predictions in the southeastern US – we set all initial and boundary concentrations to zero for ~~miscellaneous~~ unclassified $PM_{2.5}$. As in Ahmadov et al. (2012), owing to uncertainty in secondary organic aerosol (SOA) concentrations over the open ocean, we assume that initial and boundary concentrations of SOA are zero. Data from the National Centers for Environmental Prediction (NCEP) Eta model (UCAR, 2005) provide meteorological inputs; boundary conditions; and, for the Four Dimensional Data Assimilation (FDDA) employed here, observational "nudging" values.

We use the 2005 National Emissions Inventory (NEI) (US EPA, 2009) to estimate pollutant emissions. The NEI includes emissions from area, point, and mobile sources for year 2005 in the US, year 2006 in Canada, and year 1999 in Mexico. We use the model evaluation version of the NEI, which also includes hourly Continuous Emission Monitoring System (CEMS) data for electricity generating units, hourly wildfire data, and biogenic emissions from the BEIS model (Schwede et al., 2005), version 3.14.

We prepare pollutant emissions at 12 km spatial resolution using the Sparse Matrix Operating Kernel Emissions (SMOKE) program (Houyoux, 1999), version 2.6, as bundled with the NEI data (available from http://www.epa.gov/ttn/chief/emch/index.html), then we convert the emissions files output by SMOKE to WRF-Chem format and apply a plume rise algorithm (ASME, 1973, as cited in Seinfeld and Pandis, 2006) to estimate the mixing height of elevated emissions sources and wildfires. Source code for the file format conversion and plume-rise program is available at https://bitbucket.org/ctessum/emcnv.

We simulate atmospheric pollutant concentrations for the period from 1 January through 31 December 2005. We choose the year 2005 because at the time this study was performed it was the most recent year for which emissions data were available. For logistical expediency, we separate the year into eight independent model runs, each approximately 1.5 months in length plus a discarded 5 day model spin-up period. We run the simulations on a high-performance computing system consisting of 2.8 GHz Intel Xeon X5560 "Nehalem EP" processors with a 40 Gbit QDR InfiniBand (IB) interconnect and a Lustre parallel file system. Using 768 processors, each 1.5 month model run takes ~ 19 h to complete (~ 13 processor-years for each annual model run).

2.2 Comparison with observations

We compare WRF-Chem wind speed, air temperature, relative humidity, and precipitation predictions to data from the US Environmental Protection Agency (EPA) Clean Air Status and Trends Network (CASTNET) observations. We compare modeled ground-level concentrations of total $PM_{2.5}$ to EPA Air Quality System (AQS) observations (US EPA, 2005) using 24 h average data (EPA parameter code 88101) and using the less extensive hourly measurement network (EPA parameter code 88502), which allows us to compare modeled vs. measured diurnal profiles. We compare WRF-Chem predictions of O_3 to measurements from the AQS (EPA parameter code 44201) and CASTNET networks. We compare the predictions of $PM_{2.5}$ subspecies to observation data from the EPA's Chemical Speciation Network (CSN) (US EPA, 2005) (formally called Speciation Trends Network (STN)) for organic carbon (OC, parameter code 88305), elemental carbon (EC, code 88307), particulate

sulfate (SO_4, code 88403), particulate nitrate (NO_3, code 88306), and particulate ammonium (NH_4, code 88301). We additionally compare predictions to data from the Interagency Monitoring of Protected Visual Environments (IMPROVE) network (University of California Davis, 1995) for particulate OC (code 88320), EC (code 88321), sulfur (code 88169), and NO_3 (code 88306); and to CASTNET observations for particulate SO_4, NH_4, and NO_3. WRF-Chem outputs organic aerosol (OA) concentrations, but methods for measuring organic aerosol only quantify organic carbon (OC). OC comprises a variable fraction of OA, but it is common to assume an OA : OC ratio of 1.4 (Aiken et al., 2008). Therefore, we divide WRF-Chem OA predictions by a factor of 1.4 for comparison with OC measurements. Finally, we compare WRF-Chem predictions of gas-phase sulfur dioxide (SO_2) and nitrogen dioxide (NO_2) to AQS observations. We remove from consideration those stations with $\geq 25\,\%$ missing data relative to the number of scheduled measurements during the simulation period. The fractions of excluded data for each type of comparison are in the Supplement.

WRF-Chem, as configured here, outputs instantaneous concentrations at the start of each hour, whereas the observation data are reported as hourly or daily averages. WRF-Chem calculates grid-cell-average concentrations, whereas observations generally represent concentrations at specific locations.

We compare measured and modeled values pair-wise at each time of measurement in the grid cell containing each measurement station. Twenty-four hour average measurements are compared to the average of the modeled (hourly instantaneous) values within the same period. Comparisons are only made with observations that occur within the first (nearest to ground) model layer (height: $\sim$ 50–60 m). Source code for the program used to extract and pair model and measurement data is available at https://bitbucket.org/ctessum/aqmcompare.

2.3 Aggregation of results

In addition to reporting annual average model performance for the entire model domain, we also disaggregate results spatially and temporally. We evaluate performance using two

spatial approaches. First, we use four regional subdomains: Midwest, Northeast, South, and West (basis: US Census regions (US Census Bureau, 2013); see Fig. 2). Second, we evaluate urban vs. rural (i.e., not urban) locations, also as defined by the US Census (US Census Bureau, 2014). CSN monitors tend to be placed in urban areas (85 % of 186 monitors are urban), whereas IMPROVE monitors tend to be placed in protected rural areas (10 % of 122 monitors are urban). All 67 monitors in the CASTNET network are in rural locations. We also split the analysis into four seasons: winter (January through March), spring (April through June), summer (July through September), and fall (October through December). Employing these time-periods allows us to compare against previously published results (Appel et al., 2012).

2.4 Performance metrics

After matching all measured values with their corresponding modeled values, and averaging modeled and measured values across the appropriate time period, we calculate metrics shown in Eqs. (1)–(8):

$$\text{MB} = \frac{1}{n}\sum_{i=1}^{n}(M_i - O_i) \tag{1}$$

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}|M_i - O_i| \tag{2}$$

$$\text{NMB} = \frac{\sum_{i=1}^{n}(M_i - O_i)}{\sum_{i=1}^{n} O_i} \times 100\,\% \tag{3}$$

$$\text{NME} = \frac{\sum_{i=1}^{n}|M_i - O_i|}{\sum_{i=1}^{n} O_i} \times 100\,\% \tag{4}$$

$$\text{MFB} = \frac{1}{n}\sum_{i=1}^{n}\frac{2(M_i - O_i)}{M_i + O_i} \times 100\,\% \tag{5}$$

$$\text{MFE} = \frac{1}{n}\sum_{i=1}^{n}\frac{2|M_i - O_i|}{M_i + O_i} \times 100\,\% \tag{6}$$

$$\text{MR} = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{O_i} \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(M_i - O_i)^2}{n}} \tag{8}$$

where i corresponds to one of n measurement locations, M and O are time-averaged modeled and observed values, respectively, MB is mean bias, ME is mean error, NMB is normalized mean bias, NME is normalized mean error, MFB is mean fractional bias, MFE is mean fractional error, MR is model ratio, and RMSE is root-mean-square error. We additionally calculate the slope (S), intercept (I), and squared Pearson correlation coefficient (R^2) of a linear regression between modeled and measured values.

Each metric provides a useful and distinct evaluation of model performance. In general, metrics with "bias" in the name evaluate the accuracy of the model, whereas metrics with "error" in the name incorporate both precision and accuracy. Metrics that are in normalized or fractional form tend to emphasize errors where measured and observed values are relatively small, whereas non-normalized metrics tend to emphasize errors where measured and observed values are relatively large. We mainly focus here on MFB and R^2 to evaluate performance as they facilitate direct comparisons among pollutants. Results for all combinations of time periods, measurement networks, spatial subdomains, and metrics are in the Supplement.

For O_3, we calculate model performance via three model-measurement comparisons: (1) annual averages, (2) daytime-only (8 a.m. to 8 p.m.) annual averages, as in Appel et al. (2012), and (3) annual-averages of daily peak concentrations, to match the epidemiological findings in Jerrett et al. (2009).

Model performance goals and criteria have been published for $PM_{2.5}$ (Boylan and Russell, 2006). Goals reflect performance that models should strive to achieve; criteria reflect

performance that models should achieve to be used for regulatory purposes. The goals and criteria suggested by Boylan and Russell (2006) vary with concentration: they are MFB less than ± 30 and ± 60 % and MFE less than 50 and 75 %, respectively, for most concentrations, but increase exponentially as concentration decreases below $\sim 3\,\mu g\,m^{-3}$. To incorporate this aspect of performance evaluation, we calculate the fraction of observation stations for which our $PM_{2.5}$ model results meet both the MFB and MFE performance goals (fG) and criteria (fC).

3 Results

Figure 1 shows modeled annual average concentrations of $PM_{2.5}$ and O_3, where the edges of the maps represent the edges of the modeling domain. An animated version of Fig. 1 showing pollutant concentration as a function of time is available in the Supplement. Maps of additional pollutants, as well as monthly, weekly, and diurnal maps and profiles of population-weighted average concentrations, are also available in the Supplement. Modeled O_3 concentrations over water in the Gulf of Mexico and along the Atlantic coast tend to be higher than concentrations over the adjacent land areas. As only areas over water appear to be affected (as Fig. 2a shows, O_3 overpredictions along the Gulf of Mexico and Atlantic coasts are not greater than overpredictions further inland), this over-water anomaly in the Gulf of Mexico should not adversely impact estimates of population-weighted concentrations.

Figure 2 shows monitor locations for total $PM_{2.5}$ and for O_3, as well annual average fractional bias (MFB) values at each monitor. Results in Fig. 2a ($PM_{2.5}$) display high spatial variability, with no obvious spatial patterns in model performance; large overpredictions are sometimes adjacent to large underpredictions (e.g., in southern Louisiana and Florida). WRF-Chem generally overpredicts daytime O_3 concentrations relative to observations (Fig. 2b). Monitor locations for meteorological variables, $PM_{2.5}$ subspecies, and other gas phase species are in Fig. A1.

3.1 Meteorological performance

Figure 3 contains scatterplots comparing annual average observed and predicted values for meteorological variables and pollutant concentrations. The model tends to overpredict near-ground wind speed (Fig. 3a) and precipitation (Fig. 3d) relative to observations, whereas temperature (Fig. 3b) and relative humidity (Fig. 3c) predictions agree well with observations. Figures A2–A5 in Appendix A disaggregate model performance for meteorological variables by region (region boundaries are shown in Fig. 2) and by season; meteorological performance is relatively consistent among seasons and regions. Model-measurement comparisons provide important evidence on model performance but might overestimate model robustness for meteorological parameters because FDDA "nudges" model meteorological estimates toward observed values.

3.2 $PM_{2.5}$ and O_3 performance

Annual average model-measurement agreement is good for total $PM_{2.5}$ concentration (Fig. 3e, 94 % of measurements meet performance criteria), although the model tends to overpredict $PM_{2.5}$ concentration at relatively high-concentration monitors (Fig. 3e). The model tends to generally overpredict O_3 concentrations, with worse overpredictions for 24 h average concentrations (Fig. 3f) than for daily peak (Fig. 3g) and daytime average (Fig. 3f) concentrations.

Figure 4 shows the median and interquartile range for modeled and measured $PM_{2.5}$ and O_3 concentrations by hour of day (measurements of $PM_{2.5}$ subspecies are only available as 24 h averages). For $PM_{2.5}$, the model generally agrees with measurements, although on average it underpredicts concentrations at night and overpredicts during the day (Fig. 4a). For O_3, on average the model overpredicts for all times-of-day, but with a much lower fractional error during the day than during the night. For both pollutants, the model accurately captures the timing of diurnal trends, including the afternoon peak for O_3 and the morning and evening peaks for $PM_{2.5}$. As a result, when comparing the three averaging-time metrics for O_3, we observe better model performance for the annual-average of daily peak concen-

tration (MFB = 11 %) and of average daytime concentration (MFB = 12 %) than for overall annual average (MFB = 23 %). For O_3, the first two metrics may offer greater relevance than the third. For example, the annual average of daily peak concentrations is more strongly correlated with health effects than are annual average concentrations (Jerrett et al., 2009); and, for comparisons to the 8 h peak concentration National Ambient Air Quality Standard (NAAQS), model performance is more important during daytime than at night.

Figures 5 and 6 disaggregate results by season and by location for total $PM_{2.5}$ and daytime O_3, respectively; analogous results ~~for other pollutants, are in Figs. 7–11 for~~ $PM_{2.5}$ subspecies, in Figs. A2–A5 in Appendix A for meteorological properties, in Figs. A6–A7 for other O_3 temporal summaries, ~~and for meteorological variables are in Appendix A (Figs. A6–A14)~~in Fig. A8 for SO_2, and in Fig. A9 for NO_2. Daytime and peak O_3 predictive performance does not exhibit obvious patterns among seasons or regions; MFB values range from −7 to 48 % (daytime; Fig. 6) and −12 to 29 % (peak; Fig. A7). The overprediction of $PM_{2.5}$ concentrations at high-concentration monitors is more prevalent in the South and in urban areas, and is less prevalent in summer than in other seasons (Fig. 5). Model-measurement correlation for total $PM_{2.5}$ is higher in summer (AQS $R^2 = 0.64$) than in fall and winter (AQS $R^2 = 0.20$ and 0.24, respectively), but overall $PM_{2.5}$ concentrations are not higher in summer. Previous research has suggested that poor PM predictive performance in winter is common among CTMs and may be attributable to difficulty in reproducing the strongly stable meteorological conditions that are responsible for high winter PM concentrations (Solazzo et al., 2012). Annual average $PM_{2.5}$ predictive performance in the West (AQS R^2: 0.45 (summer), 0.13 (winter)) is worse than performance in the Northeast (AQS R^2: 0.70 (summer), 0.37 (winter)). In the Northeast, performance is better in the summer ($R^2 = 0.69$) than in other seasons ($R^2 = 0.30$–0.40). Taken together, these findings suggest that there is an opportunity for future model development for $PM_{2.5}$ to focus on winter or full-year simulations rather than summer-only simulations, and on the western US or the full contiguous US rather than just the Northeast.

3.3 $PM_{2.5}$ subspecies performance

Figure 3i–m illustrates model performance for annual average concentrations of $PM_{2.5}$ component species. In all cases, $>65\,\%$ of locations meet performance criteria for at least one of the three observation networks.

The model overpredicts particulate SO_4 (CSN MFB $= 34\,\%$, IMPROVE MFB ~~$= 126$~~ $= 40\,\%$, CASTNET MFB $= 36\,\%$) (Fig. 3i) and SO_2 (MFB $= 51\,\%$) (Fig. 3n). This finding (overprediction of total sulfur) agrees with prior research for multiple CTMs (McKeen et al., 2007). ~~Performance as compared to the IMPROVE network is worse than performance as compared to the CSN and CASTNET networks, perhaps owing to differences in measurement methods.~~ Particulate SO_4 prediction performance does not vary much by region; as with total $PM_{2.5}$, performance is worse in winter (CSN MFB $= 59\,\%$) than summer (CSN MFB $= 10.\,\%$) (Fig. ~~A8~~7).

WRF-Chem as configured here performs well in predicting observed particulate NH_4 concentrations, with 99 % of locations meeting performance criteria (Fig. 3j). Similar to total $PM_{2.5}$, performance for particulate NH_4 is worst in the urban areas in the West region (Fig. ~~A9~~8), where a number of monitors report relatively high measured concentrations but modeled concentrations are relatively low.

Particulate NO_3 concentrations are consistently underpredicted (annual average MFB $= -110\,\%$) (Fig. 3k). Figure ~~A10~~ 9 shows that these underpredictions are more severe in some seasons and regions than in others. The best predictive performance is for the Midwest in summer (MFB $= -39\,\%$) followed by the Northeast in summer (MFB $= -47\,\%$). NO_3 predictions in the West region are poor for all seasons (MFB $= -148\,\%$)~~, as are wintertime predictions for the contiguous US (MFB $= -120$)~~. As with other $PM_{2.5}$ species, there is an opportunity for future development and evaluation of models for particulate NO_3 prediction to focus on seasons and regions other than summer in the Northeast. Predictions of gas-phase NO_2 (Fig. 3o) agree relatively well with observations (MFB $= 4\,\%$), but, as with other species, the model tends to overpredict NO_2 concentrations in areas where measured con-

centrations are relatively high. This effect is especially prominent in the West and in urban areas (Fig. ~~A14~~A9).

Model-measurement agreement for EC concentrations is relatively good (Fig. 3l), with 96 % of monitor locations meeting performance criteria. As with other comparisons, for EC the model tends to overpredict concentrations for monitors with relatively high concentrations, especially in urban areas (Fig. ~~A11~~10).

Model predictions of OC concentrations (Fig. 3m) are biased low compared to CSN (MFB $= -$~~11~~35 %) but agree relatively well with IMPROVE (MFB $= 15$ %). Mean bias values given here are within the range of values reported by a previous publication using the VBS SOA formation mechanism (Ahmadov et al., 2012). As shown in Fig. ~~A12~~11, the differences ~~between~~ in model-measurement aggreement between the two networks do not appear to be dependent on urban vs. rural monitor location~~; instead~~. Instead, they may reflect between-network differences in sampling or analysis; different analysis techniques are known to produce widely varying OC concentrations (Cavalli et al., 2010).

3.4 Comparison with other studies

Table 2 compares performance of WRF-Chem as configured here to that of the CMAQ model in a similar modeling effort by Appel et al. (2012). In this table, CMAQ as configured by Appel et al. (2012) in most cases predicts O_3 observations with greater accuracy and precision than does WRF-Chem as configured here, while WRF-Chem in most cases does a better job predicting $PM_{2.5}$. However, given the many differences in physical and chemical parameterizations and input data (including a difference in simulation year), the observed differences may or may not be generalizable. Instead, our conclusion from Table 2 is that the models are generally comparable in performance.

Table A2 compares WRF-Chem results from this study to results from Yahya et al. (2014) for a 12 month, contiguous US WRF-Chem simulation with a 36 km horizontal resolution spatial grid. NME results from the simulation performed here are lower (i.e., better) than those reported by Yahya et al. for most pollutants and measurement networks, but NMB results are more mixed. As horizontal grid resolution, input data, and model parameters all

differ between the two studies, we are not able to determine the cause of the differences in results.

4 Discussion

We simulated and evaluated $PM_{2.5}$ and O_3 based on 12 month (year 2005) WRF-Chem modeling for the United States. The spatial and temporal extent investigated, and the horizontal spatial resolution (12 km) employed, are nearly unprecedented; to our knowledge, only one prior peer-reviewed ~~article has investigated CTMs using the same~~ CTM evaluation has used a comparable extent and resolution (Appel et al., 2012). We find that WRF-Chem performance as configured here is generally comparable to other models used in regulatory and health impact assessment situations in that model performance is similar to that reported by Appel et al. (2012) and in most cases meets criteria for air quality model performance suggested by Boylan and Russel (2006).

There is potential for further improvement in model accuracy, especially for these cases: $PM_{2.5}$ concentrations in winter and in the western US, ground-level O_3 at night and in the summer, and particulate nitrate~~and organic carbon~~. The good agreement in total $PM_{2.5}$ predictions and observations in some cases reflects offsetting over- and underpredictions, including by species (~~Figs. A8–A12~~Fig. 3) and time-of-day (Fig. 4a). Performance in predicting concentrations of $PM_{2.5}$ and its subspecies tends to be the worst in winter and in the western US. Overall, WRF-Chem as configured here meets the performance criteria described above for total $PM_{2.5}$ concentrations at 94 % of monitor locations.

The WRF-Chem meteorological and chemical settings employed here are reasonable and justified, but different settings may also be reasonable. Improved understanding of how alternative parameterizations might impact model performance in large-scale applications such as ours is an area for continued research. Another area for future research is identifying opportunities to evaluate model performance in terms of how changes in emissions cause changes in outdoor concentrations.

5 Supporting information

Supplement includes WRF-Chem configuration settings (ascii format); maps showing spatial patterns in pollutant concentrations by annual average, month of year, day of week, and hour of day (pdf format); model-measurement comparison statistics (xlsx format); and monitor-specific paired model and measurement data (json ascii format). A video showing spatially- and temporally-explicit O_3 and $PM_{2.5}$ concentrations is at http://youtu.be/4bpQXBAUVwE.

The Supplement related to this article is available online at doi:10.5194/gmdd-0-1-2015-supplement.

Acknowledgements. We acknowledge the University of Minnesota Institute on the Environment Initiative for Renewable Energy and the Environment Grant No. RI-0026-09 and the US Department of Energy Award No. DE-EE0004397 for funding, the Minnesota Supercomputing Institute and the Department of Energy National Center for Computational Sciences Award No. DD-ATM007 for computational resources, Steven Roste for assistance with model-measurement comparison, and John Michalakes for assistance with WRF-Chem performance tuning.

References

Ackermann, I. J., Hass, H., Memmesheimer, M., Ebel, A., Binkowski, F. S., and Shankar, U.: Modal Aerosol Dynamics Model for Europe: development and first applications, Atmos. Environ., 32, 2981–2999, 1998.

Ahmadov, R., McKeen, S. A., Robinson, A. L., Bahreini, R., Middlebrook, A. M., de Gouw, J. A., Meagher, J., Hsie, E.-Y., Edgerton, E., Shaw, S., and Trainer, M.: A volatility basis set model for summertime secondary organic aerosols over the eastern United States in 2006, J. Geophys. Res., 117, D06301, doi:10.1029/2011JD016831, 2012.

Aiken, A. C., DeCarlo, P. F., Kroll, J. H., Worsnop, D. R., Huffman, J. A., Docherty, K. S., Ulbrich, I. M., Mohr, C., Kimmel, J. R., Sueper, D., Sun, Y., Zhang, Q., Trimborn, A., Northway, M., Ziemann, P. J.,

Canagaratna, M. R., Onasch, T. B., Alfarra, M. R., Prevot, A. S. H., Dommen, J., Duplissy, J., Metzger, A., Baltensperger, U., and Jimenez, J, L.: O/C and OM/OC ratios of primary, secondary, and ambient organic aerosols with high-resolution time-of-flight aerosol mass spectrometry, Environ. Sci. Technol., 42, 4478–4485, 2008.

American Society of Mechanical Engineers (ASME): Recommended Guide for the Prediction of the Dispersion of Airborne Effluents, 2nd edn., ASME, New York, NY, 1973.

Appel, K. W., Chemel, C., Roselle, S. J., Francis, X. V., Hu, R.-M., Sokhi, R. S., Rao, S. T., and Galmarini, S.: Examination of the Community Multiscale Air Quality (CMAQ) model performance over the North American and European domains, Atmos. Environ., 53, 142–155, 2012.

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

Chuang, M.-T., Zhang, Y., and Kang, D.: Application of WRF/Chem-MADRID for real-time air quality forecasting over the southeastern United States, Atmos. Environ., 45, 6241–6250, 2011.

Cavalli, F., Viana, M., Yttri, K. E., Genberg, J., and Putaud, J.-P.: Toward a standardised thermal-optical protocol for measuring atmospheric organic and elemental carbon: the EUSAAR protocol, Atmos. Meas. Tech., 3(1), 79—89, 2010.

Emmons, L. K., Walters, S., Hess, P. G., Lamarque, J.-F., Pfister, G. G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S. L., and Kloster, S.: Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4), Geosci. Model Dev., 3, 43–67, doi:10.5194/gmd-3-43-2010, 2010.

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

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., and Bash, J. O.: Incremental testing of the Community Multiscale Air Quality (CMAQ) modeling system version 4.7, Geosci. Model Dev., 3, 205–226, doi:10.5194/gmd-3-205-2010, 2010.

Fountoukis, C., Koraj, D., Denier van der Gon, H. A. C., Charalampidis, P. E., Pilinis, C., and Pandis, S. N.: Impact of grid resolution on the predicted fine PM by a regional 3-D chemical transport model, Atmos. Environ., 68, 24–32, 2013.

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

Galmarini, S., Rao, S. T., and Steyn, D. G.: AQMEII: an international initiative for the evaluation of regional-scale air quality models – Phase 1 preface, Atmos. Environ., 53, 1–3, 2012.

Houyoux, M. R. and Vukovich, J. M.: Updates to the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system and integration with Models-3, in: Proceedings of the Emission Inventory: Regional Strategies for the Future, Air and Waste Management Association, Raleigh, NC, 26–28 October 1999, 1999.

Jerrett, M., Burnett, R. T., Pope III, C. A., Ito, K., Thurston, G., Krewski, D., Shi, Y., Calle, E., and Thun, M.: Long-term ozone exposure and mortality, New Engl. J. Med., 360, 1085–1095, 2009.

Krewski, D., Jerrett, M., Burnett, R. T., Ma, R., Hughes, E., Shi, Y., Turner, M. C., Pope III, C. A., Thurston, G., Calle, E. E., and Thun, M. J.: Extended Follow-Up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality, Health Effects Institute, Boston, MA, available at: http://www.ncbi.nlm.nih.gov/pubmed/19627030 (last access: 28 November 2014), 2009.

Levy, J. I., Wilson, A. M., Evans, J. S., and Spengler, J. D.: Estimation of primary and secondary particulate matter intake fractions for power plants in Georgia, Environ. Sci. Technol., 37, 5528–5536, 2003.

McKeen, S., Chung, S. H., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Gong, W., Bouchet, V., Moffet, R., Tang, Y., Carmichael, G. R., Mathur, R., and Yu, S.: Evaluation of several $PM_{2.5}$ forecast models using data collected during the ICARTT/NEAQS 2004 field study, J. Geophys. Res., 112, D10S20, doi:10.1029/2006JD007608, 2007.

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

Peckham, S. E., Grell, G. A., McKeen, S. A., Ahmadov, R., Barth, M., Pfister, G., Wiedinmyer, C., Fast, J. D., Gustafson, W. I., Ghan, S. J., Zaveri, R., Easter, R. C., Barnard, J., Chapman, E., Hewson, M., Schmitz, R., Salzmann, M., Beck, V., and Freitas, S. R.: WRF/Chem Version 3.4 User's Guide, available at: http://ruc.noaa.gov/wrf/WG11 (last access: 18 December 2012), 2012.

Pope III, C. A. and Dockery, D. W.: Health effects of fine particulate air pollution: lines that connect, J. Air Waste Manage., 56, 709–742, 2006.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 2nd edn., John Wiley & Sons, Inc., Hoboken, NJ, 2006.

Schwede, D., Pouliot, G., and Pierce, T.: Changes to the Biogenic Emissions Inventory System Version 3 (BEIS3), in: 4th Annual CMAS Model-3 User's Conference, Chapel Hill, NC, 26–28 September 2005, available at: http://cmascenter.org/conference/2005/abstracts/2_7.pdf (last access: 28 November 2014), 2005.

Solazzo, E. Bianconi, R., Pirovano, G., Matthias, V., Vautard, R., Moran, M. D., Appel, K. W., Bessagnet, B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Ferreira, J., Forkel, R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Miranda, A. I., Nopmongcol, U., Prank, M., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S. T., and Galmarini, S.: Operational model evaluation for particulate matter in Europe and North America in the context of AQMEII, Atmos. Environ., 53, 75–92, 2012.

Stockwell, W. R., Kirchner, F., Kuhn, M., and Seefeld, S.: A new mechanism for regional atmospheric chemistry modeling, J. Geophys. Res., 102, 25847–25879, 1997.

Tesche, T. W., Morris, R., Tonnesen, G., McNally, D., Boylan, J., and Brewer, P.: CMAQ/CAMx annual 2002 performance evaluation over the eastern US, Atmos. Environ., 40, 4906–4919, 2006.

Tessum, C. W., Marshall, J. D., and Hill, J. D.: A spatially and temporally explicit life cycle inventory of air pollutants from gasoline and ethanol in the United States, Environ. Sci. Technol., 46(20), 11408—11417, 2012.

Tessum, C. W., Hill, J. D., and Marshall, J. D.: Life cycle air quality impacts of conventional and alternative light-duty transportation in the United States, Proc. Natl. Acad. Sci. U.S.A., 111(52), 18490—18495, 2014.

University Corporation for Atmospheric Research (UCAR): GCIP NCEP Eta model output, available at: http://rda.ucar.edu/datasets/ds609.2/ (last access: 15 January 2012), 2005.

University of California Davis: IMPROVE data guide: a guide to interpret data, Prepared for National Park Service, Air Quality Research Division, Fort Collins, CO, available at: http://vista.cira.colostate.edu/improve/publications/OtherDocs/IMPROVEDataGuide/IMPROVEdataguide.htm (last access: 18 September 2013), 1995.

US Census Bureau: Cartographic Boundary Shapefiles – Regions, available at: https://www.census.gov/geo/maps-data/data/cbf/cbf_region.html (last access: 10 February 2014), 2013.

US Census Bureau: Year-2014 US urban areas and clusters, available at: ftp://ftp2.census.gov/geo/tiger/TIGER2014/UAC/ (last access: 10 February 2014), 2014.

US Environmental Protection Agency (US EPA): Technology Transfer Network (TTN) Air Quality System (AQS), available at: http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm (last access: 6 March 2013), 2005.

US Environmental Protection Agency (US EPA): 2005 National Emissions Inventory (NEI), available at: http://www.epa.gov/ttn/chief/emch/index.html (last access: 7 March 2012), 2009.

US Environmental Protection Agency (US EPA): Air Quality Modeling Technical Support Document for the Regulatory Impact Analysis for the Revisions to the National Ambient Air Quality Standards for Particulate Matter, Research Triangle Park, NC 27711, available at: http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OAR-2010-0955-0017 (last access: 28 November 2014), 2012.

Wang, W., Bruyère, C., Duda, M., Dudhia, J., Gill, D., Kavulich, M., Keene, K., Lin, H.-C., Michalakes, J., Rizvi, S., Zhang, X., Berner, J., and Smith, K.: Weather Research and Forecasting: ARW: Version 3 Modeling System User's Guide, available at: http://www.mmm.ucar.edu/wrf/users/docs/user_guide_V3/contents.html (last access: 28 November 2014), 2012.

Yahya, K., Wang, K., Gudoshava, M., Glotfelty, T., and Zhang, Y.: Application of WRF/Chem over North America under the AQMEII Phase 2: Part I. Comprehensive evaluation of 2006 simulation, Atmos. Environ., doi:10.1016/j.atmosenv.2014.08.063, in press, 2014.

Zhang, Y., Pan, Y., Wang, K., Fast, J. D., and Grell, G. A.: WRF/Chem-MADRID: incorporation of an aerosol module into WRF/Chem and its initial application to the TexAQS2000 episode, J. Geophys. Res., 115, D18202, doi:10.1029/2009JD013443, 2010.

Zhang, Y., Chen, Y., Sarwar, G., and Schere, K.: Impact of gas-phase mechanisms on Weather Research Forecasting Model with Chemistry (WRF/Chem) predictions: mechanism implementation and comparative evaluation, J. Geophys. Res., 117, D01301, doi:10.1029/2011JD015775, 2012.

Table 1. Selected WRF-Chem v3.4 settings and parameters employed in this study.

Category	Option used
Microphysics	WSM 3-class simple ice scheme
Shortwave and longwave radiation	CAM scheme
Land surface	Unified Noah land surface model
Boundary layer physics	YSU scheme
Cumulus physics	New Grell scheme (G3)
FDDA meteorology nudging	Yes (grid-based)
Gas-phase chemistry	NOAA/ESRL RACM
Aerosol chemistry/physics	MADE/VBS
Aerosol feedback	No
Photolysis	Fast-J
Anthropogenic emissions	2005 NEI
Biogenic emissions	BEIS v3.14
Horizontal grid resolution	12 km
Number of vertical layers	28

Table 2. WRF-Chem and CMAQ Seasonal O_3 and $PM_{2.5}$ prediction performance.

	Daytime[a] average O_3 (ppb)		$PM_{2.5}$ ($\mu g\,m^{-3}$)	
	WRF-Chem	CMAQ[b]	WRF-Chem	CMAQ[b]
Winter MB	3.5	−3.5	0.8	3.4
Spring MB	1.5	−1.8	2.0	2.0
Summer MB	9.2	4.4	0.0	−0.6
Fall MB	5.2	2.6	−0.9	4.0
Winter ME	5.5	9.0	3.1	6.0
Spring ME	4.6	9.3	3.3	4.5
Summer ME	10.1	11.0	2.6	4.4
Fall ME	6.2	8.8	2.7	5.6
Winter NMB	12 %	−13 %	6 %	30 %
Spring NMB	3 %	−4 %	17 %	19 %
Summer NMB	21 %	10. %	0 %	−5 %
Fall NMB	19 %	8 %	−7 %	36 %
Winter NME	19 %	35 %	25 %	53 %
Spring NME	10 %	29 %	28 %	42 %
Summer NME	23 %	24 %	18 %	31 %
Fall NME	23 %	28 %	23 %	52 %

[a] Daytime is defined as 8 a.m. to 8 p.m. LT.
[b] Adapted from Appel et al. (2012) Tables 1 and 2.

Table A1. Temporal and spatial aspects of recent model evaluations, focusing on WRF-Chem and North America.

Author and year	Model used	Time period	Spatial extent	Horizontal spatial resolution
Ahmadov et al. (2012)	WRF-Chem	Aug–Sep 2006	Contiguous US (evaluation performed for eastern US)	60 and 20 km
Appel et al. (2006)	CMAQ	Full year, 2006	Contiguous US and Europe	12 km
Chuang et al. (2011)	WRF-Chem	May–Sep 2009	Southeastern US	12 km
Fast et al. (2006)	WRF-Chem	Late Aug 2000	City of Houston	1.3 km
Grell et al. (2005)	WRF-Chem	Jul–Aug 2002	Eastern US	27 km
McKeen et al. (2007)	WRF-Chem, CHRONOS, AURAMS, STEM, CMAQ/ETA	Jul–Aug 2004	Northeastern US	12, 21, 27, and 42 km
Misenis and Zhang (2010)	WRF-Chem	Late Aug 2000	Eastern Texas	4 and 12 km
Tesche et al. (2006)	CMAQ, CAMx	Full year, 2002	Contiguous US	12 km Eastern US, 36 km contiguous US
Yahya et al. (2014)	WRF-Chem	Full year, 2006	Contiguous US	36 km
Zhang et al. (2010)	WRF-Chem	Late Aug 2010	Eastern Texas	12 km
Zhang et al. (2012)	WRF-Chem	Jul 2001	Contiguous US	36 km

Table A2. WRF-Chem annual average predictive performance by pollutant in Yahya et al. (2014) and in the current study.

Variable	Network	MB		NMB		NME	
		Yahya et al. (2014)	Current study	Yahya et al. (2014)	Current study	Yahya et al. (2014)	Current study
Daily Peak O_3 (ppb)	CASTNET	−8.6	3.9	−18 %	9 %	24 %	12 %
	AQS	−0.3	5.5	−5 %	13 %	9 %	15 %
Daytime Average O_3 (ppb)	CASTNET	−5.6	3.5	−13 %	9 %	22 %	11 %
	AQS	−1.7	4.9	−4 %	13 %	24 %	16 %
SO_2 (ppb)	AQS	−0.6	5.1	−18 %	130 %	87 %	150 %
NO_2 (ppb)	AQS	1.7	1.6	17 %	12 %	73 %	34 %
Total $PM_{2.5}$ ($\mu g\,m^{-3}$)	CSN	0.0	0.4	0 %	3 %	45 %	18 %
SO_4 $PM_{2.5}$ ($\mu g\,m^{-3}$)	IMPROVE	0.5	0.9	35 %	40 %	66 %	42 %
	CSN	0.9	1.6	32 %	41 %	59 %	42 %
	CASTNET	0.9	1.3	34 %	38 %	55 %	38 %
NH_4 $PM_{2.5}$ ($\mu g\,m^{-3}$)	CSN	0.1	0.0	10. %	−2 %	53 %	16 %
	CASTNET	0.3	0.1	30. %	7 %	50. %	16 %
NO_3 $PM_{2.5}$ ($\mu g\,m^{-3}$)	IMPROVE	−0.1	−0.5	−14 %	−69 %	85 %	69 %
	CSN	−0.6	−1.3	−38 %	−72 %	75 %	72 %
	CASTNET	−0.1	−0.7	−15 %	−65 %	83 %	65 %
EC $PM_{2.5}$ ($\mu g\,m^{-3}$)	IMPROVE	0.0	0.0	15 %	−9 %	67 %	31 %
	CSN	0.4	0.2	54 %	25 %	90. %	43 %
OC $PM_{2.5}$ ($\mu g\,m^{-3}$)	IMPROVE	0.0	0.2	1 %	17 %	59 %	33 %

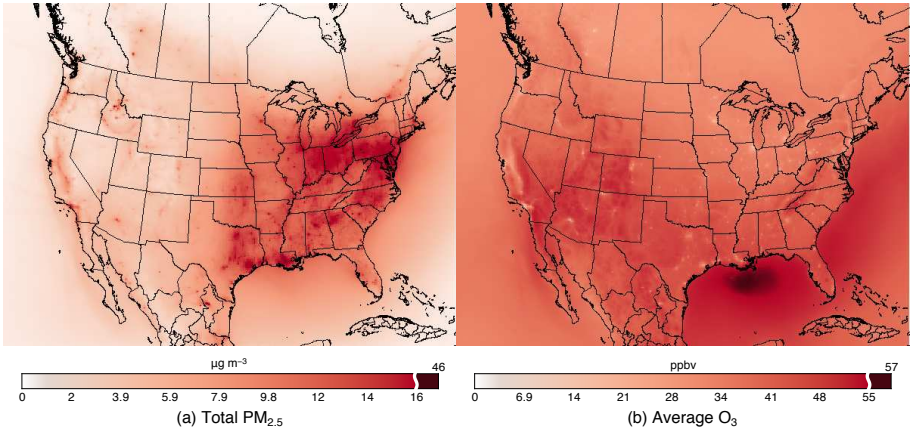

Figure 1. Modeled annual average ground level concentrations of **(a)** $PM_{2.5}$ and **(b)** O_3. For ease of viewing, the color scales contain a break at the 99th percentile of concentrations.

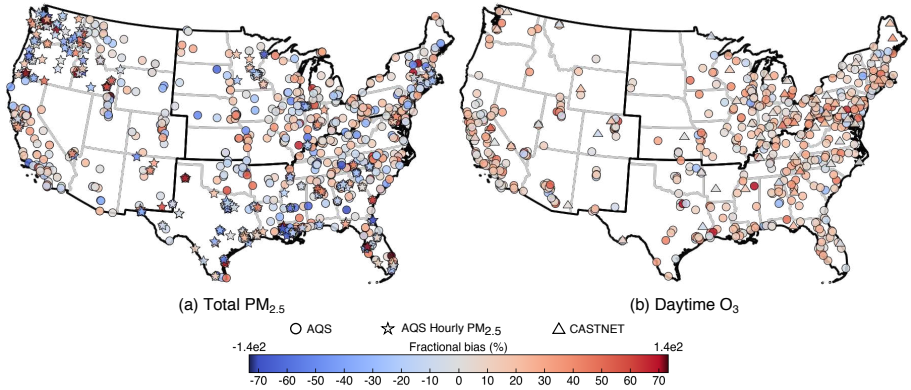

Figure 2. AQS, AQS hourly, and CASTNET monitor locations and annual average fractional bias for total $PM_{2.5}$ **(a)** and daytime average O_3 concentrations **(b)**. Corresponding information for other pollutants and variables is in Fig. A1.

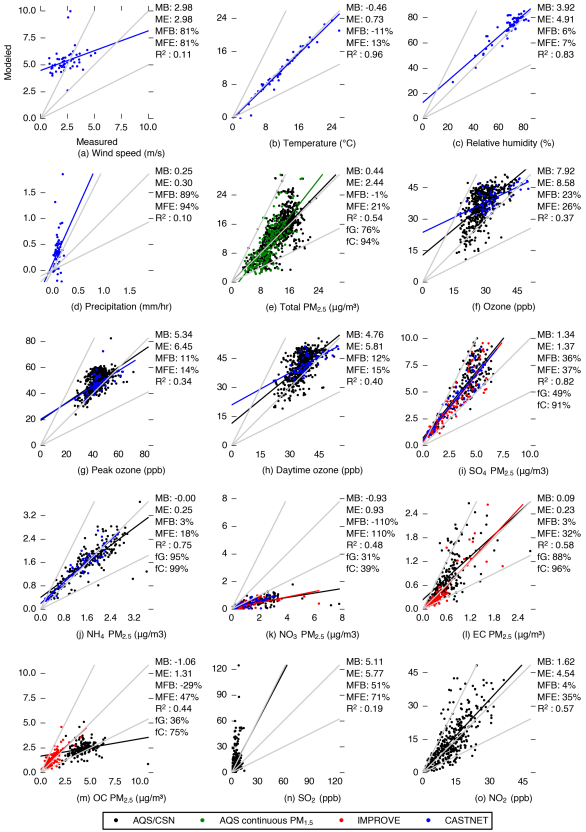

Figure 3. Annual average modeled and measured ground-level meteorological variables **(a–d)** and pollutant concentrations **(e–o)**. Colored lines show linear least-squares fits of the data for the measurement networks with corresponding colors. Grey lines show model to measurement ratios of 2 : 1, 1 : 1, and 1 : 2. Annual average performance statistics are listed to the right of each plot; acronyms are defined in the methods section.

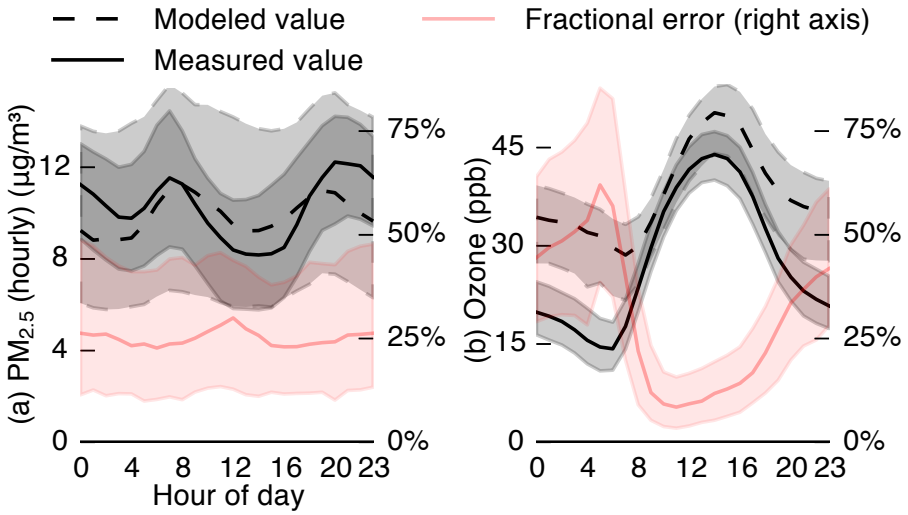

Figure 4. Median values (lines) and interquartile ranges (shaded areas) of annual average modeled values, observed values, and fractional error by hour of day for $PM_{2.5}$ **(a)** and O_3 **(b)**.

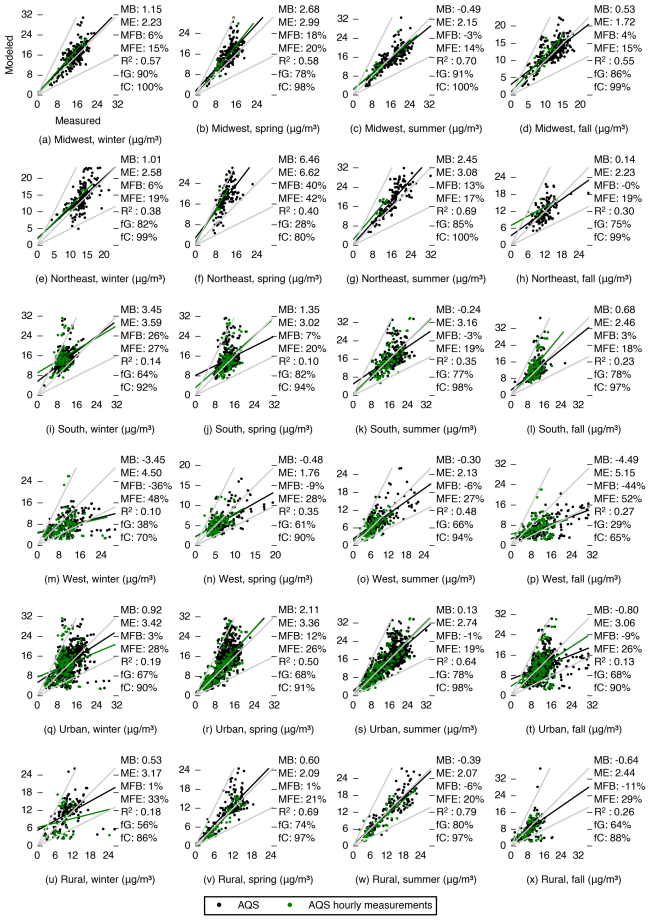

Figure 5. Comparison of measured and modeled $PM_{2.5}$ concentration disaggregated by season and region. Region boundaries are shown in Fig. 2.

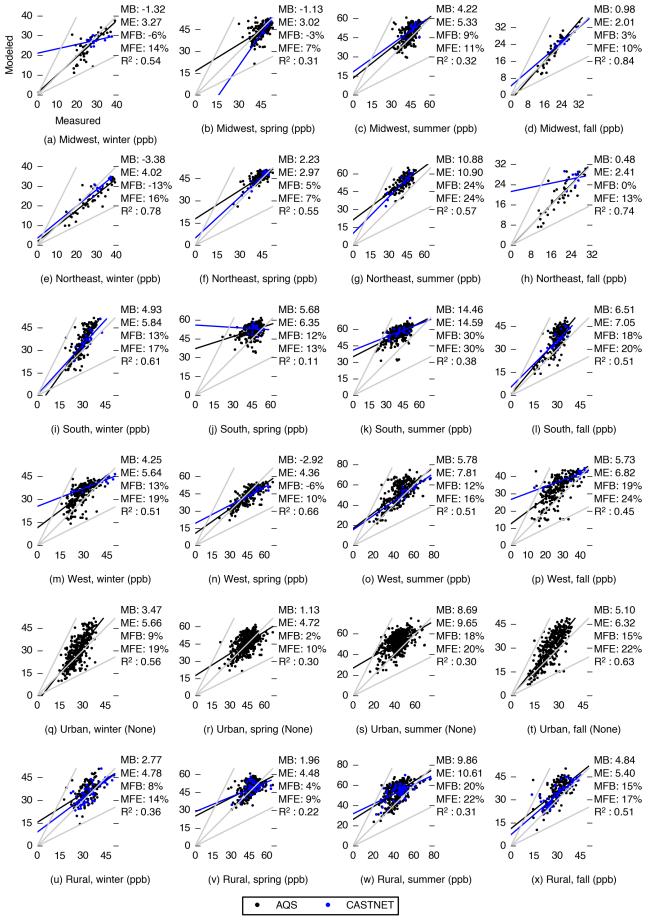

Figure 6. Comparison of measured and modeled annual average of daytime O_3 concentration disaggregated by season and region. Region boundaries are shown in Fig. 2.

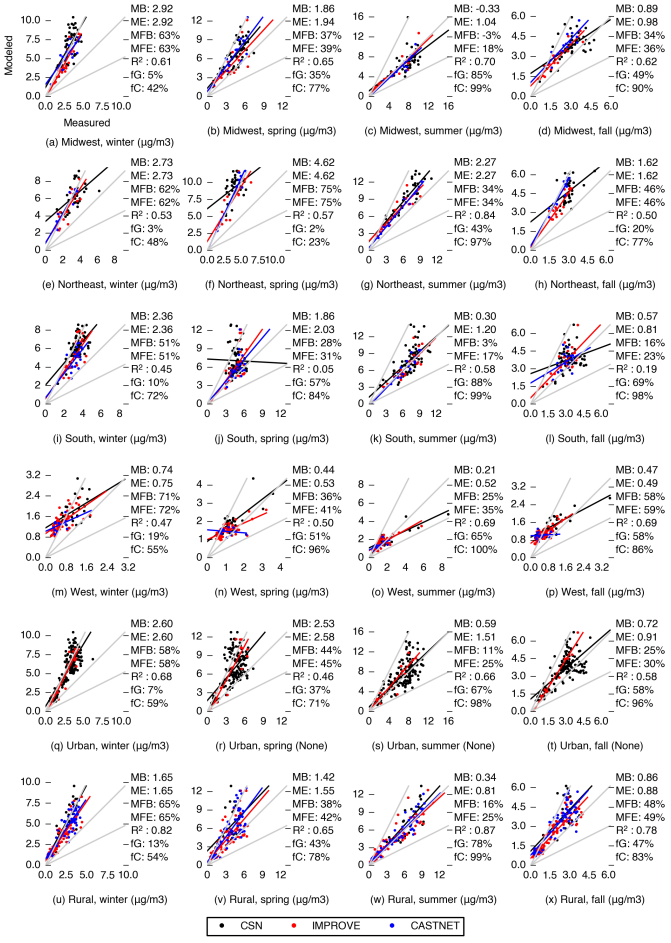

Figure 7. Comparison of modeled and measured *particulate* SO_4 *concentration*, disaggregated by region and season.

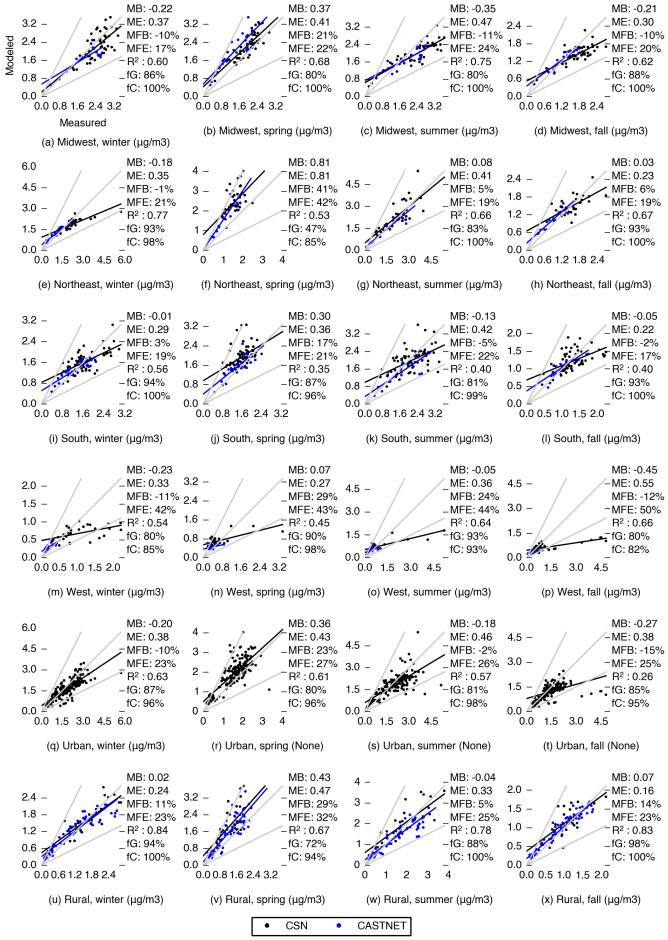

Figure 8. Comparison of modeled and measured *particulate* NH_4 *concentration*, disaggregated by region and season.

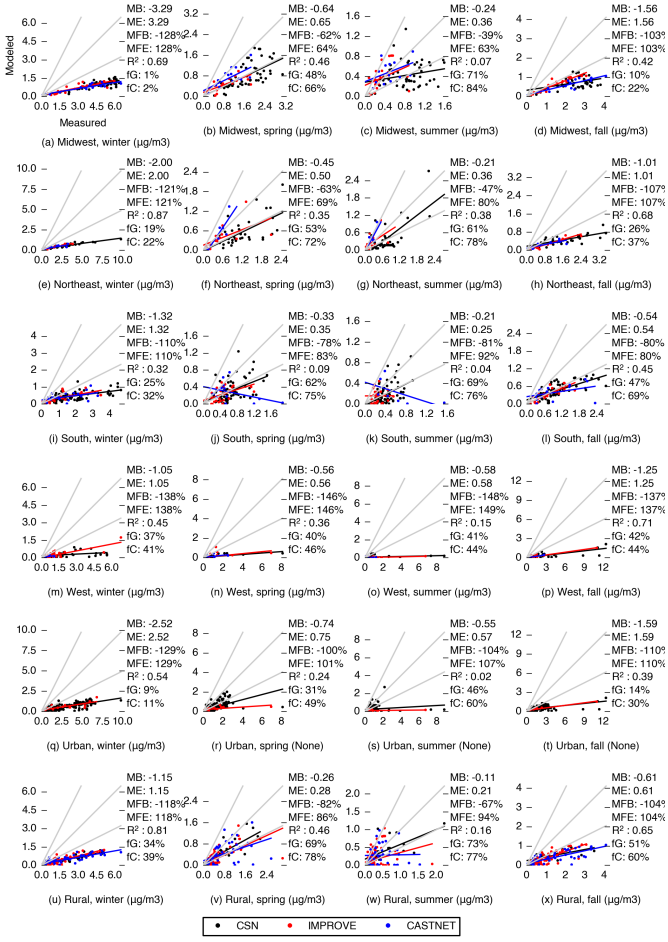

Figure 9. Comparison of modeled and measured *particulate* NO_3 *concentration*, disaggregated by region and season.

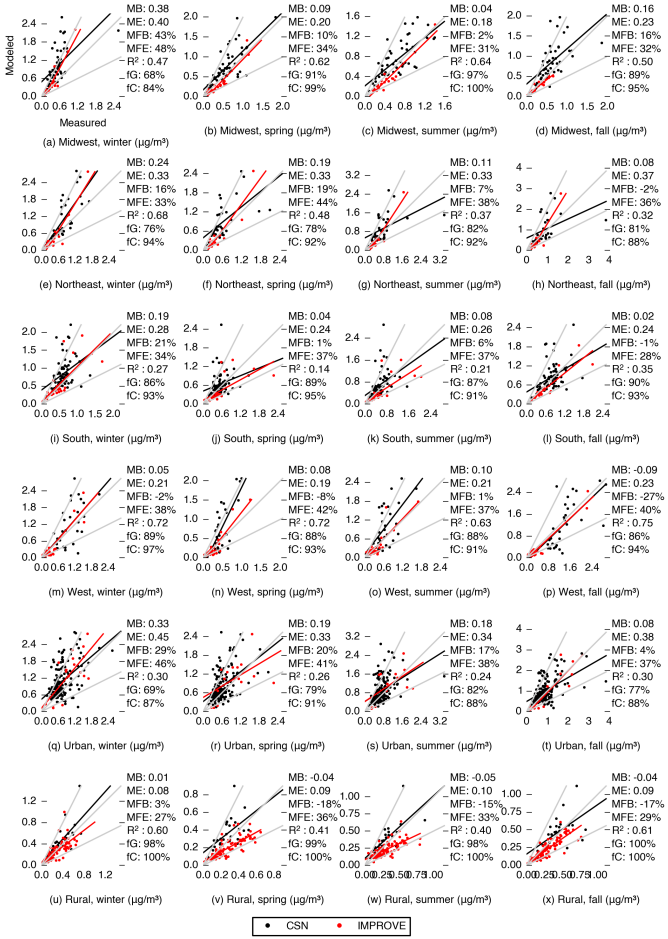

Figure 10. Comparison of modeled and measured *particulate* EC *concentration*, disaggregated by region and season.

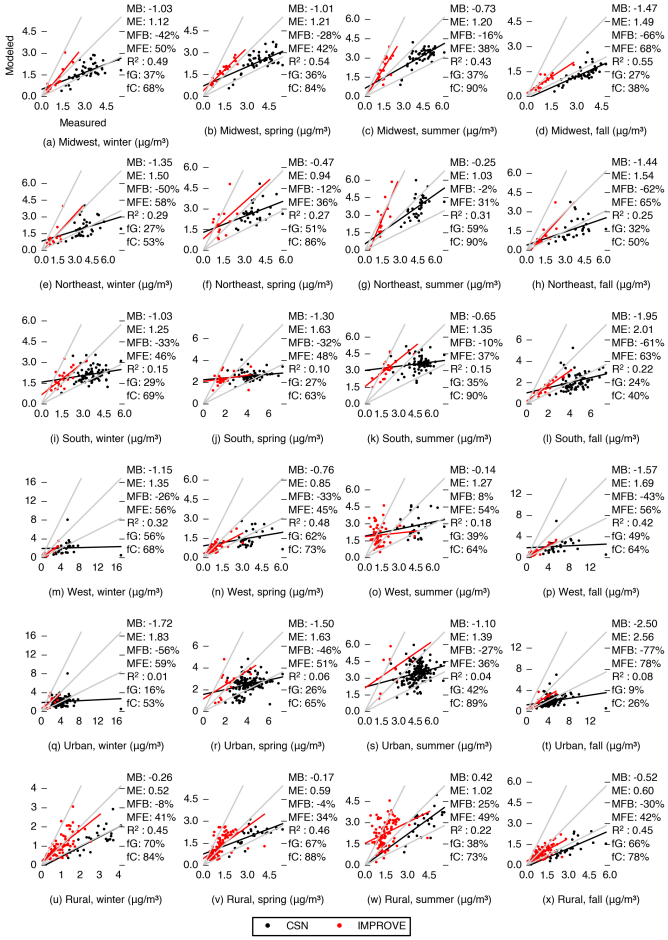

Figure 11. Comparison of modeled and measured *particulate* OC *concentration*, disaggregated by region and season.

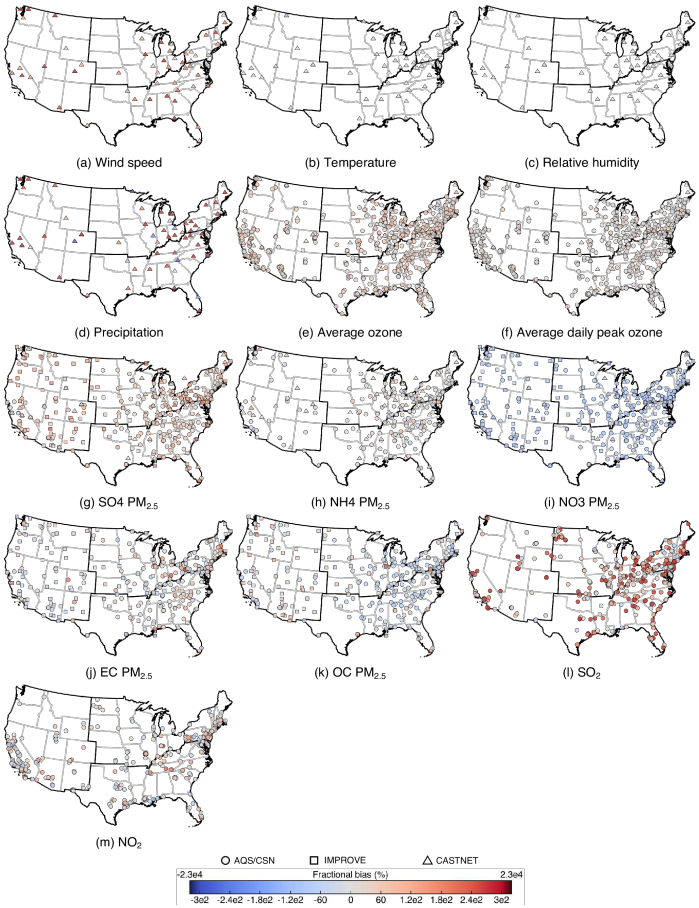

Figure A1. AQS, CSN, IMPROVE AQS and CASTNET monitor locations and annual average fractional bias for total meteorological variables **(a–d)** and pollutant concentrations **(e–m)**.

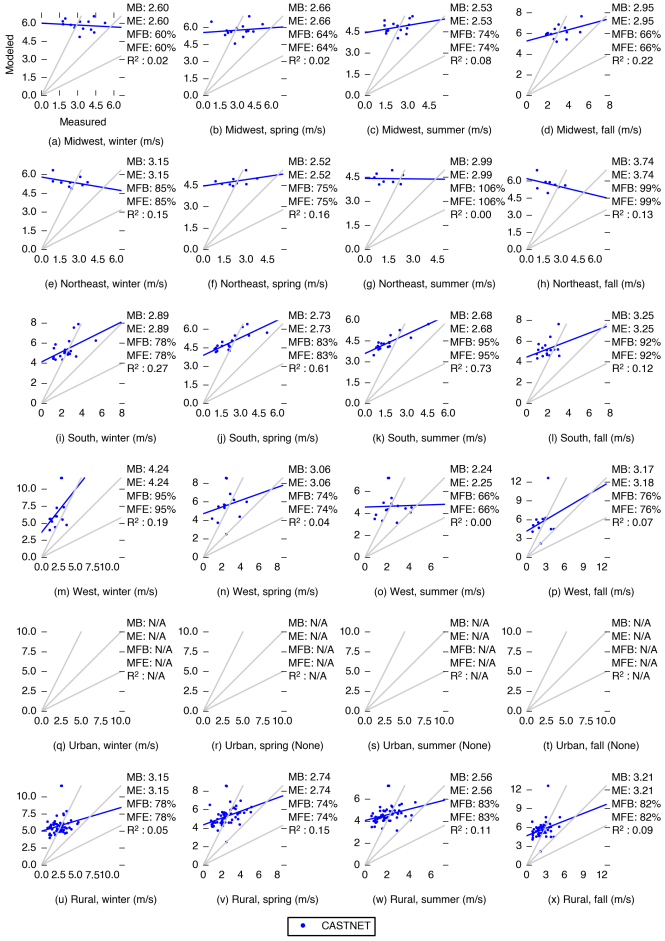

Figure A2. Comparison of modeled and measured *wind speed*, disaggregated by region and season.

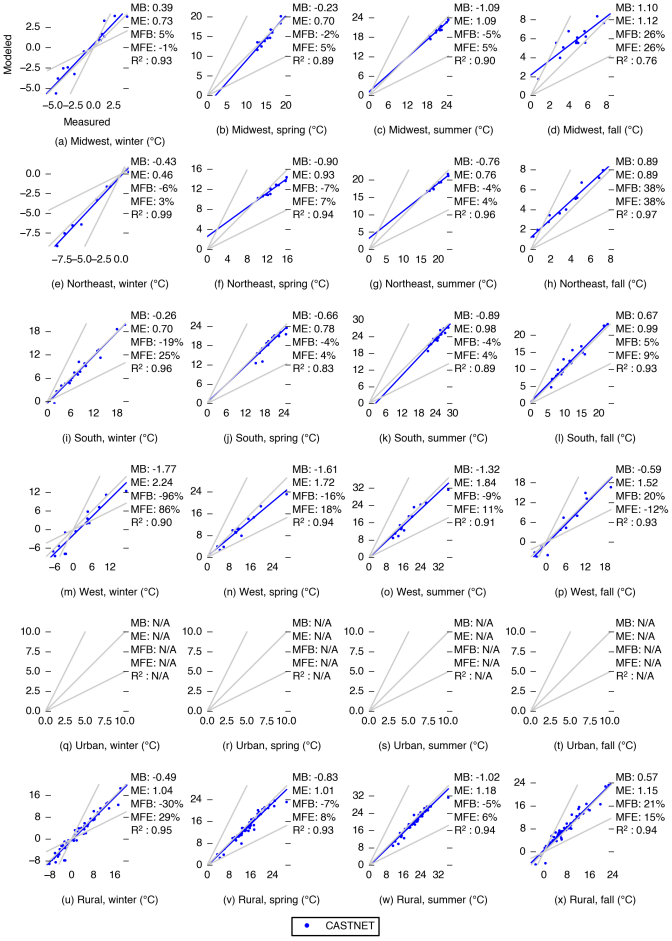

Figure A3. Comparison of modeled and measured *temperature*, disaggregated by region and season.

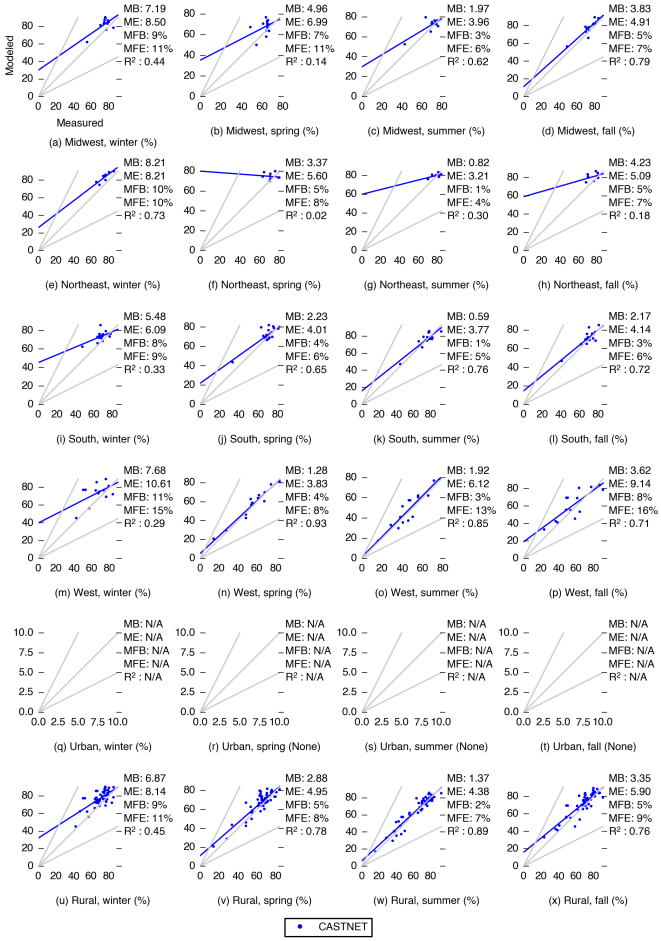

Figure A4. Comparison of modeled and measured *relative humidity*, disaggregated by region and season.

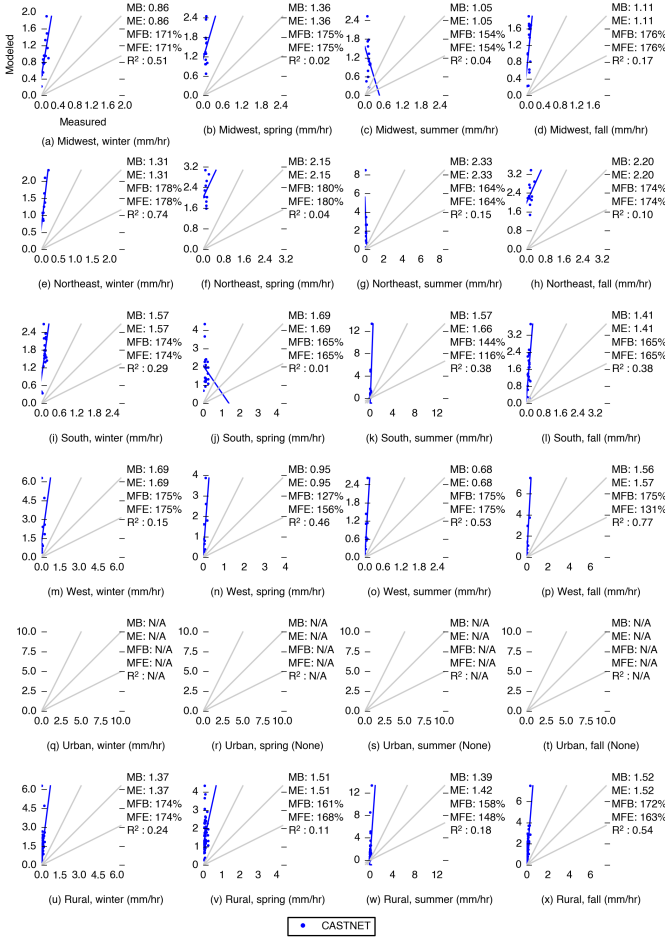

Figure A5. Comparison of modeled and measured *precipitation*, disaggregated by region and season.

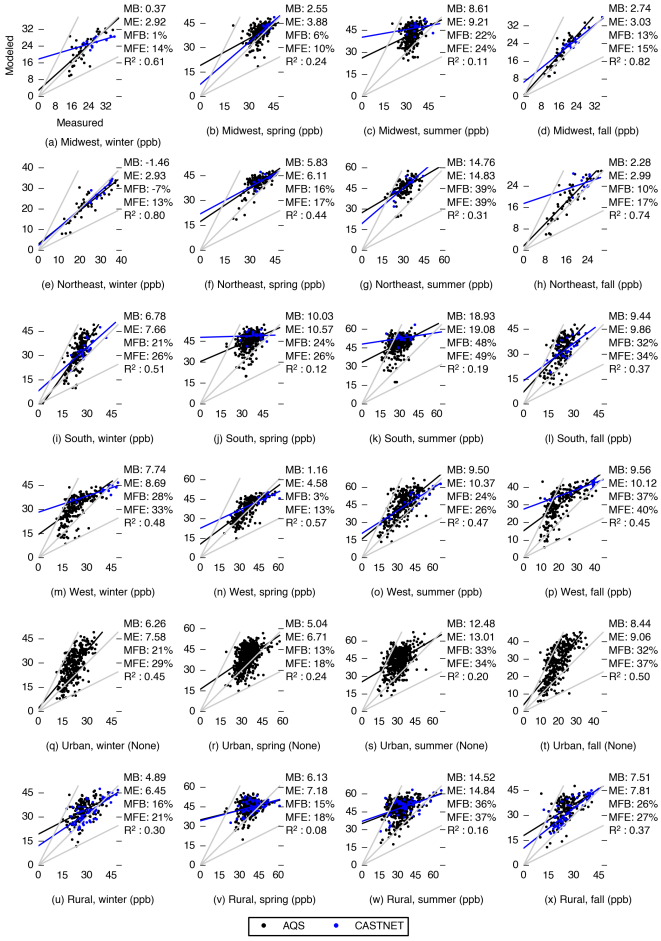

Figure A6. Comparison of modeled and measured *annual-average* O_3 *concentration*, disaggregated by region and season.

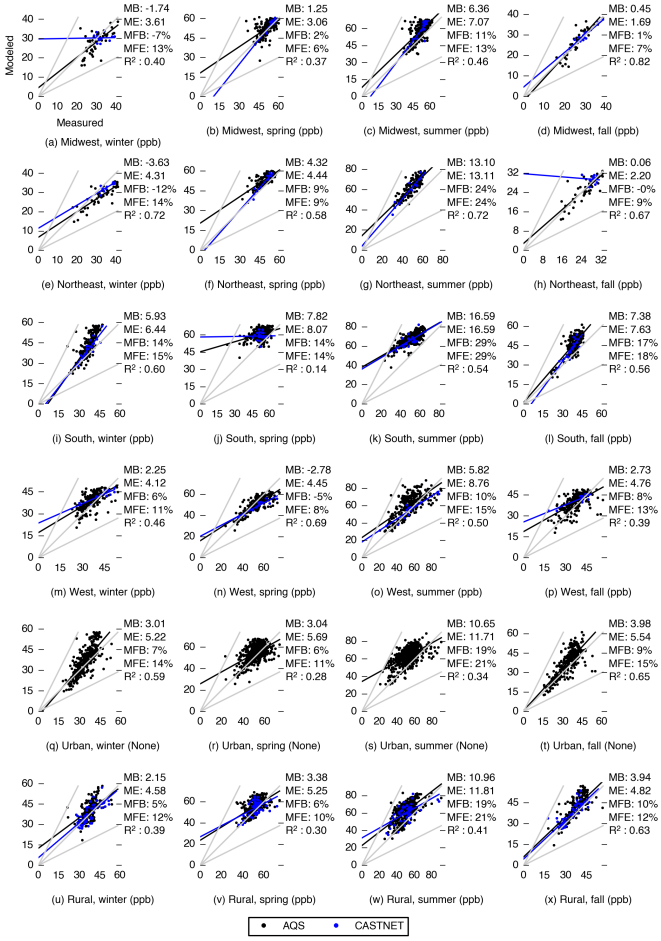

Figure A7. Comparison of modeled and measured average *daily peak* O_3 *concentration*, disaggregated by region and season.

Comparison of modeled and measured *particulate* *concentration*, disaggregated by region and season.

Comparison of modeled and measured *particulate* *concentration*, disaggregated by region and season.

Comparison of modeled and measured *particulate* *concentration*, disaggregated by region and season.

Comparison of modeled and measured *particulate* EC *concentration*, disaggregated by region and season.

Comparison of modeled and measured *particulate* OC *concentration*, disaggregated by region and season.

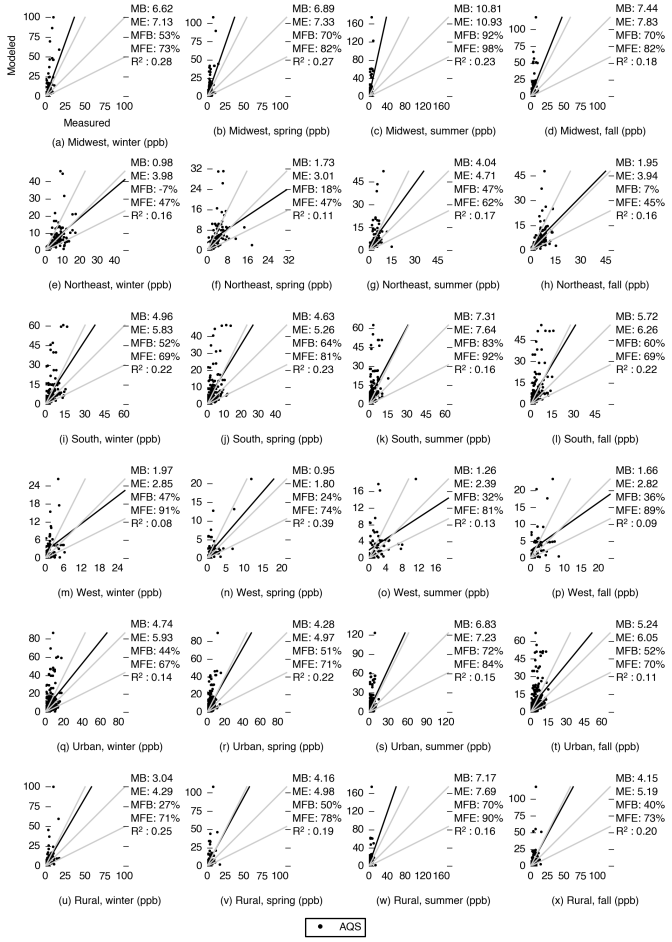

Figure A8. Comparison of modeled and measured SO_2 *concentration*, disaggregated by region and season.

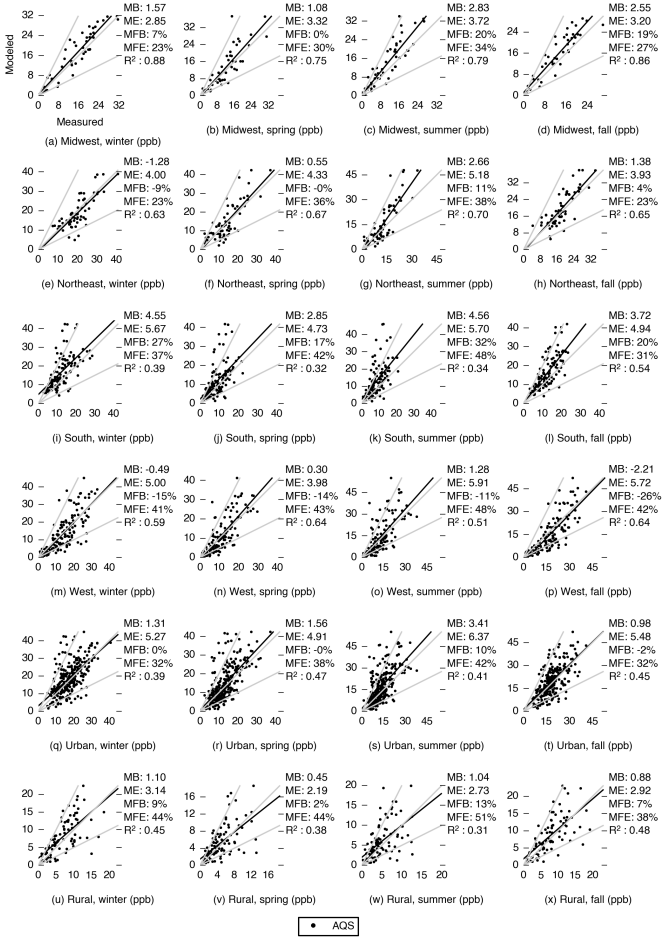

Figure A9. Comparison of modeled and measured NO_2 *concentration*, disaggregated by region and season.