# Comprehensive Evaluation of Multi-Year Real-Time Air Quality Forecasting Using an Online-Coupled Meteorology-Chemistry Model over Southeastern United States

Yang Zhang[1,*], Chaopeng Hong[1,2], Khairunnisa Yahya[1], Qi Li[1], Qiang Zhang[2], and Kebin He[3, 2]
[1]Air Quality Forecasting Laboratory, North Carolina State University, Raleigh, NC, USA
[2]Center for Earth System Science, Tsinghua University, Beijing, P.R. China, 100084
[3]The School of Environment, Tsinghua University, Beijing, P.R. China, 100084

## Abstract

An online-coupled meteorology-chemistry model, WRF/Chem-MADRID, has been deployed for real time air quality forecast (RT-AQF) in southeastern U.S. since 2009. A comprehensive evaluation of multi-year RT-AQF shows overall good performance for temperature and relative humidity at 2-m (T2, RH2), downward surface shortwave radiation (SWDOWN) and longwave radiation (LWDOWN), and cloud fraction (CF), ozone ($O_3$) and fine particles ($PM_{2.5}$) at surface, tropospheric ozone residuals (TOR) in $O_3$ seasons (May-September), and column $NO_2$ in winters (December-February). Moderate-to-large biases exist in wind sped at 10-m (WS10), precipitation (Precip), cloud optical depth (COT), ammonium ($NH_4^+$), sulfate ($SO_4^{2-}$), and nitrate ($NO_3^-$) at the IMPROVE and SEARCH networks, organic carbon (OC) at IMPROVE, and elemental carbon (EC) and OC at SEARCH, aerosol optical depth (AOD) and column carbon monoxide (CO), sulfur dioxide ($SO_2$), and formaldehyde (HCHO) in both $O_3$ and winter seasons, column nitrogen dioxide ($NO_2$) in $O_3$ seasons, and TOR in winter. These biases indicate uncertainties in the boundary layer and cloud process treatments (e.g., surface roughness, microphysics cumulus parameterization), emissions (e.g., $O_3$ and PM precursors, biogenic, mobile, and wildfire emissions), upper boundary conditions for all major gases and $PM_{2.5}$ species, and chemistry and aerosol treatments (e.g., winter photochemistry, aerosol thermodynamics). The model shows overall good skills in reproducing the observed multi-year trends and inter-seasonal variability in meteorological and radiative variables such as T2, WS10, Precip, SWDOWN, and LWDOWN, and relatively well the observed trends in surface $O_3$ and $PM_{2.5}$, but relatively poor for column abundances of CO, $NO_2$, $SO_2$, HCHO, TOR, and AOD. The sensitivity simulation using satellite-constrained boundary conditions for $O_3$ and CO shows substantial improvement for both spatial distribution and domain-mean performance statistics. The model's forecasting skills for air quality can be further enhanced through improving model inputs (e.g., anthropogenic emissions for urban areas and upper boundary conditions of chemical species), meteorological forecasts (e.g., WS10, Precip) and meteorologically-dependent emissions (e.g., biogenic and wildfire emissions), and model physics and chemical treatments (e.g., gas-phase chemistry in winter conditions, cloud processes and its interactions with radiation and aerosol).

**Corresponding author:* Yang Zhang, Department of Marine, Earth, and Atmospheric Sciences, Campus Box 8208, NCSU, Raleigh, NC 27695; e-mail: yang_zhang@ncsu.edu

**Keywords**

WRF/Chem-MADRID, southeastern U.S., discrete evaluation, categorical evaluation, satellite data, multi-year trend analysis

## 1. Introduction

Real-time air-quality forecasting (RT-AQF) of the concentrations of pollutants of special health concerns such as ozone ($O_3$) and fine particulate matter ($PM_{2.5}$) provides a basis for early air quality alerts and preventative actions that reduce air pollution and protect human health. Increasing public awareness of adverse health impacts of ambient air pollution in both developed and developing countries and the availability of complex, deterministic three-dimensional (3-D) numerical models for RT-AQF have provided driving forces for the establishment and advancement of RT-AQF. Despite substantial improvements of ambient air quality in major cities in many countries, the frequent occurrences of severe regional hazes in recent years in a number of countries such as China (e.g., Wang et al., 2014), India, and Singapore necessitate the continuous development and application of techniques for RT-AQF worldwide. A number of 3-D air quality models have been deployed for RT-AQF since the mid-1990s on global (e.g., Takigawa et al., 2007; Mangold et al., 2011) and regional scales (e.g., Carmichael et al., 2003; McHenry et al., 2004; McKeen et al., 2005; 2010; Yu et al., 2007, 2008; Eder et al., 2010). Kukkonen et al. (2011) reviewed 18 regional scale RT-AQF models that are currently used in Europe, among which, 3 out 18 are online-coupled models. Zhang et al. (2012a, b) provided a comprehensive review of history, techniques, current status, and future research needs along with 9 global and 36 regional RT-AQF models that are currently used in Australia, North America, South America, Europe, and Asia, among which, 4 out 9 global models and 5 out of 36 regional models are online-coupled models. Among those models, the 3-D RT-AQF models with coupled meteorology and chemistry such as the online-coupled Weather Research and Forecasting model with Chemistry (WRF/Chem) (Grell et al., 2005) are advanced tools for RT-AQF that can realistically represent the feedback mechanisms between meteorology and chemistry in the atmosphere. They, however, may not always outperform offline RT-AQF models, as there remain larger uncertainties in RT-AQF models than those originating from the feedback mechanisms, and not all RT-AQF models represent all feedback mechanisms that occur in the real atmosphere. The strengths and limitations of online-coupled models have been reviewed in several papers (e.g., Zhang 2008; Baklanov et al., 2014).

Since May 2009, WRF/Chem with the Model of Aerosol Dynamics, Reaction, Ionization, and Dissolution (MADRID) (WRF/Chem-MADRID) (Zhang et al., 2010a, 2012c) has been deployed by the lead author's group for RT-AQF in southeastern U.S. for ozone ($O_3$) season

(May-September) and winter season (December-February) (Chuang et al., 2011; Yahya et al., 2014a). The multi-year RT-AQF enables the assessment of the model's capability and robustness in forecasting major pollutants as well as their inter-annual and inter-season variability, and multi-year trends with the long-term forecasting data. In this work, multi-year forecasts of air quality and meteorology during 2009-2015 using WRF/Chem-MADRID are evaluated against surface and satellite-derived observations. The objectives are to evaluate the model's skill in forecasting the observed air quality and meteorology and their variation trends during 2009-2015 and to identify areas of model improvements for more accurate meteorological and chemical forecasts.

## 2. Model Description and Evaluation Protocol

### 2.1 Model Description

WRF/Chem-MADRID is an online-coupled meteorology and chemistry model. It was developed based on WRF/Chem version 3.0 (Grell et al., 2005) and CMAQ-MADRID (Zhang et al., 2004) with updates in gas-phase chemistry and aerosol treatments by Zhang et al. (2010a, b, 2012c). WRF/Chem-MADRID treats all major aerosol processes such as the thermodynamic equilibrium for both inorganic and organic species, new particle formation, condensation/evaporation, coagulation, gas/particle mass transfer, dry and wet deposition. Unlike offline-coupled air quality models, WRF/Chem-MADRID simulates aerosol direct and semi-direct feedbacks to photolysis, radiation, and planetary boundary layer (PBL) meteorology, as well as aerosol indirect effects on cloud and precipitation formation via many aerosol-cloud interaction processes. The physics and chemistry options used in this study follow those of Chuang et al. (2011) and Yahya et al. (2014a); they are kept the same for all forecasting periods since 2009. The physics options include the cloud microphysics of Lin et al. (1983); the Rapid

Radiative Transfer Model (RRTM) of Mlawer et al. (1997) for longwave radiation; the Goddard scheme of Chou et al. (1998) for shortwave radiation; the Yonsei University (YSU) PBL scheme of (Hong et al. 2006); the National Center for environmental Prediction, Oregon State University, Air Force, Hydrologic Research Lab (NOAH) LSM (Chen and Dudhia, 2001); and the Grell-Devenyi ensemble cumulus parameterization (Grell and Devenyi, 2002). The chemistry and aerosol-related options chosen include the 2005 Carbon Bond gas-phase chemical mechanism (CB05) (Yarwood et al., 2005); the Carnegie-Mellon (CMU) bulk aqueous-phase chemical kinetic mechanism (Fahey and Pandis, 2001), the MADRID1 aerosol module with 8 size sections over the PM aerodynamic diameter range of 0.025-11.630 μm of Zhang et al. (2004, 2010a, b, 2012c), and the aerosol activation of Abdul Razzak and Ghan (2002). A more detailed description of the model can be found in Chuang et al. (2011) and Yahya et al. (2014a).

**2.2 RT-AQF Deployment and Inputs**

The forecasting simulations are performed during the $O_3$ and winter seasons at a horizontal grid resolution of 12 km over an area in southeastern U.S. including the states of Mississippi (MI), Alabama (AL), Georgia (GA), Florida (FL), South Carolina (SC), North Carolina (NC), Tennessee (TN), Kentucky (KY), Virginia (VA), West Virginia (WV), and Delaware (DE), as well as small portions of Louisiana (LA), Arkansas (AR), Missouri (MS), Illinois (IL), Indiana (IN), Ohio (OH), and Maryland (MD). The hourly and daily forecast products are provided at http://www.meas.ncsu.edu/aqforecasting/Real_Time.html. This study analyzes forecast products during six $O_3$ and winter seasons between May 1, 2009 and February 28, 2015. The National Center for Environmental Prediction's (NCEP) meteorological forecast is downloaded at 7 p.m. (Local Standard Time) to initialize a 60-hr forecasting cycle using WRF/Chem-MADRID with 12-hr spin-up and 48-hr forecasting. The anthropogenic emissions are based on the projected 2009 emissions by the Visibility Improvement State and Tribal Association of the Southeast's

(VISTAS) from the 1999 National Emission Inventories (NEI) version 2 based on historical growth factors and assumed control strategies (Barnard and Sabo, 2008). Those emissions vary hourly and account for seasonal variations. For biogenic emissions, offline biogenic emissions available from the VISTAS emissions were originally used for the RT-AQF during May 1, 2009 and February 28, 2011. The online biogenic emissions from the Model for Gases and Aerosols from Nature (MEGAN) version 2 have been used since December 2011. Mineral dust emissions are simulated using online dust emission of Shaw (2008).

The VISTAS 2009 36-km CMAQ simulation results and those from the previous day's simulation are used to provide daily chemical boundary and initial conditions (BCONs and ICONs), respectively. One-week spin up simulation is performed for the first day of the first 60-hr forecasting cycle for each forecasting season.

**2.3 Evaluation Datasets and Protocols**

Zhang et al. (2012a) recommended both discrete and categorical evaluation for RT-AQF models, which are carried out for meteorological and chemical forecasts in this work. The PBL meteorological variables evaluated include temperature at 2-m (T2), relative humidity at 2-m (RH2), wind speed and direction at 10-m (WS10 and WD10), and daily precipitation (Precip). The chemical species evaluated include maximum 1-hr and 8-hr $O_3$, carbon monoxide (CO), sulfur dioxide ($SO_2$), nitric oxide (NO), nitrogen dioxide ($NO_2$), nitric acid ($HNO_3$), 24-hr average $PM_{2.5}$ and $PM_{2.5}$ species such as ammonium ($NH_4^+$), sulfate ($SO_4^{2-}$), nitrate ($NO_3^-$), elemental carbon (EC), organic carbon (OC) and total carbon (TC = EC+OC). Given the low accuracy of anemometers at low wind speed conditions, the observed and simulated data pairs with the observed value below 0.771 m $s^{-1}$ are excluded in the statistical calculation following Olerud et al. (2005). A number of surface networks are used for model evaluation, as summarized in Table S1 in the supplementary material. These include the National Climatic

Data Center (NCDC), the AIRNow database, the Air Quality System (AQS), the Clean Air Status and Trends Network (CASTNET), the Interagency Monitoring of Protected Visual Environments (IMPROVE), the Speciated Trends Network (STN), and the Southern Aerosol Research and Characterization (SEARCH). While AIRNow, AQS, and STN include primarily urban and suburban sites, and NCDC, CASTNET and IMPROVE include mainly rural and remote sites. NCDC and SEARCH includes both urban and rural sites in southeastern U.S.

While 14 statistics defined in Zhang et al. (2006, 2012a) and Yu et al. (2006) are calculated against all surface network datasets in the discrete evaluation, the analysis in this study focuses on several commonly-used metrics including the mean bias (MB), normalized mean bias (NMB), the normalized mean error (NME), mean absolute gross error (MAGE), Root mean square error (RMSE), and correlation coefficient (R). The discrete performance statistical criteria for chemical forecasts are based on Zhang et al. (2006) which recommended the use of NMBs ≤ 15% and NMEs ≤ 30% to indicate a satisfactory performance for $O_3$ and $PM_{2.5}$. For meteorological variables, Tesche and Tremback (2002) suggested a good performance with MB ≤ 0.5 m $s^{-1}$ for WS10, MB ≤ 10 degrees and MAGE ≤ 30 degrees for WD10, and MB ≤ 0.5 K and MAGE ≤ 2 K for T2. However, such criteria were developed for meteorological simulations with data assimilation. Data assimilation is not used in this work because it masks the feedbacks between chemistry and meteorology. The model performance may not be as well as those with data assimilation. Brunner et al. (2014) evaluated meteorological simulations for the year of 2010 from eight simulations of WRF version 3.4 with different combinations of physics options and found that the monthly MBs of T2 are within 2 K and MBs of WS10 are within 1.7 m $s^{-1}$. The reported NMBs of Precip simulated by WRF range from -88% to 66% (e.g., Zhang et al., 2010c; Yahya et al., 2014a, b, 2015a; Penrod et al., 2014). NMBs within ±30% are considered to be acceptable performance for Precip. Categorical statistics are calculated for the maximum 1-hr and 8-hr $O_3$ and 24-hr average $PM_{2.5}$ against near-real time observations from AIRNow in terms of accuracy (A), critical success index (CSI), probability of detection (POD), bias (B), and false alarm ratio (FAR), as defined in Kang et al. (2005) and Zhang et al. (2012a). The threshold values are 80 ppb for the maximum 1-hr $O_3$; 60 ppb for maximum 8-hr $O_3$; and 15 $\mu g\ m^{-3}$ for 24-hr average $PM_{2.5}$ following Chuang et al. (2011). For categorical evaluation, satisfactory performance would yield values close to 1 for A, CSI, and POD and a value close to 0 for FAR. For B, a value of 1 would indicate no bias, and a number greater than 1 means that the model forecasts more exceedances than observed, and vice versa. In addition to domain-mean discrete and categorical statistics, the forecasted meteorological variables and chemical concentrations are evaluated using available observations in terms of domain-mean spatial distributions and site-specific hourly variations. The representative urban and rural sites selected include Atlanta, Georgia; Charlotte and Raleigh in North Carolina; Louisville, Kentucky; Birmingham, Alabama; and Jacksonville, Florida.

In addition to surface evaluation, satellite data are used to assess the model's capability in forecasting column values of meteorological, radiative, and chemical variables, as summarized in Table S1. Such evaluations have not been previously performed for RT-AQF models. These include Precip from the Global Precipitation Climatology Project (GPCP), downward surface shortwave radiation (SWDOWN) and longwave radiation (LWDOWN) from the Cloud's and the Earth's Radiant Energy System (CERES), cloud fraction (CF), aerosol optical depth (AOD), and cloud optical depth (COT) from the Moderate Resolution Imaging Spectroradiometer (MODIS), tropospheric CO column abundances from the Measurements of Pollution in the Troposphere (MOPITT), tropospheric column abundances of $NO_2$, formaldehyde (HCHO), and sulfur dioxide ($SO_2$), as well as tropospheric ozone residuals (TOR) from the Ozone Monitoring Instrument (OMI)/ Microwave Limb Sounder (MLS). All satellite data used are level-3 monthly average (except for column $SO_2$, which is daily average because monthly average is not available) retrieval data that have been validated and quality assured by data providers (Martin, 2008). Following Zhang et al. (2009), the model outputs for all column variables except for TORs are vertically integrated up to the tropopause and averaged at the same satellite crossing time to generate the tropospheric amounts in order to match the satellite data. Column variables are evaluated in terms of domain-mean discrete statistics and spatial distributions.

## 3. **Evaluation of Model Performance**

### 3.1 Evaluation of Meteorological Variables

Meteorological forecasts are evaluated to understand their influence on chemical forecasts. The meteorological performance for three $O_3$ seasons and three winter seasons during May 1, 2009-February 28, 2012 has been evaluated in Yahya et al. (2014a). This study focuses on the evaluation of three $O_3$ seasons and three winter seasons during May 1, 2012-February 28, 2015. Table 1 summarizes domain-mean performance statistics for T2, RH2, WS10, and WD10 against data from CASTNET, NCDC, and SEARCH, Precip against data from CASTNET, NCDC, and GPCP, SWDOWN and LWDOWN against CERES, and CF and COT against MODIS during these three $O_3$ seasons and three winter seasons.

#### 3.1.1 Ozone Seasons

MBs for T2 range from 0.5-2.1 °C, 0.6-1.8 °C, and 0.9-2.6 °C and MAGEs range from 4.0-4.2°C, 3.7-4.0°C, and 3.9-4.7°C against data from CASTNET, NCDC, and SEARCH, respectively. The values of R for T2 range from 0.6-0.7 at CASTNET and NCDC, and 0.3-0.5 at SEARCH. Low R values at the SEARCH sites indicate possible compensation of large positive and negative biases at different sites. While the MBs and MAGEs of T2 are larger than 0.5 K and 2 K, respectively, suggested by Tesche and Tremback (2002), they fall into the typical ranges of MBs (< 2 °C) reported for this and newer versions of WRF and WRF/Chem in the literature (e.g., Brunner et al., 2014). Moderate warm biases in T2 are mainly caused by moderate overpredictions in SWDOWN with NMBs of 14%, 17.6%, and 27.6%, and moderate underpredictions in CF with NMBs of -8.3%, -12.7%, and -14.5% at the CASTNET, NCDC, and SEARCH sites, respectively. In this version of WRF, sub-grid cloud feedbacks to radiation are neglected in the cumulus parameterization, contributing in part to the overpredictions in SWDOWN (Alapaty et al., 2012). Limitations in the surface layer and shortwave radiation schemes also contribute to the overpredictions in SWDOWN. The large underpredictions of COT with NMBs of -65.8% to -60.3% reflect the poor ability of the model in simulating cloud variables, due to the limitations in the parameterizations of cloud dynamics, thermodynamics, and microphysics, and interactions with aerosols (Zhang et al., 2012c, d, 2015). The model simulates LWDOWN well, with NMBs within 2%.

The warm biases in T2 directly affect RH2 forecasts. Moderate underpredictions occur in RH2 with MBs of -16.4% to -9.7%, -14.6% to -6.9%, and -20.5% to -10.4%, at the CASTNET, NCDC, and SEARCH sites, respectively. The values of R of RH2 are lower, ranging from 0.2-0.4 at all sites. The model simulates WS10 at the NCDC sites well with MBs of 0.2-0.4 m $s^{-1}$, MGAEs of 1.8 m $s^{-1}$, and NMBs of 4.8-5.2%. However, the model moderately or significantly overpredicts WS10 at the SEARCH and CASTNET sites with MBs of 0.3-0.9 m $s^{-1}$ and 1.6-1.8 m $s^{-1}$, MAGEs of 1.2-1.3 m $s^{-1}$ and 1.9-2.1 m $s^{-1}$, and NMBs of 15.1-40.9% and 65.6-100.5%, respectively. The MBs at all sites are generally within 1.7 m $s^{-1}$ reported by Brunner et al. (2014) for simulations with WRF version 3.4 and at the NCDC sites they are even smaller than a performance indicator value of 0.5 m $s^{-1}$ suggested by Tesche and Tremback (2002) for simulations with data assimilation. Similar large overpredictions of WS10 by WRF have been reported by a number of studies (e.g., Penrod et al., 2014; Yahya et al., 2014a; Brunner et al., 2014). The WS10 overpredictions are due in part to unresolved surface roughness and topographical features by the surface drag parameterization used in WRF and in part to the use of coarse horizontal and vertical resolutions in the forecasting simulations (Cheng and Steenburgh, 2005; Mass and Ovens, 2011). Comparing to the NCDC sites that were carefully selected for meteorological measurements, the SEARCH and CASTNET sites were selected for air quality measurements, and many sites are difficult to be resolved at a spatial grid resolution of 12-km because of complex topography and surfaces. MBs for WD10 range from 16.3-29.0°, 42.4-47.9°, and 1.7-24.4° and MAGEs range from 79.1-86.2 °C, 85.1-86.1 °C, and 76.2-94.1°C against data from CASTNET, NCDC, and SEARCH, respectively. The values of MBs and MAGEs are much higher than 10 and 30 degrees, respectively, suggested by Tesche and Tremback (2002), indicating a poor performance for WD10 that is partly because the data assimilation is not used and partly because the surface roughness and topographic features cannot be resolved. The values of R for WD10 range from 0.6-0.7 at CASTNET and NCDC, and 0.3-0.6 at SEARCH. These results indicate certain limitations in the YSU PBL and the Monin-Obukhov surface layer schemes used in resolving main features of the PBL meteorology, particularly over complex terrain with uneven surface topography and mountainous regions (e.g., the Appalachian mountains).

Precip is moderately to significantly overpredicted with NMBs of 52.0-56.2%, 34.4-49.7%, and 29.8-54.6% against data from CASTNET, NCDC, and GPCP, respectively, they are mostly beyond the acceptable performance range of ±30%. Similar large overpredictions of Precip by WRF or WRF/Chem have been reported in many studies (e.g., Caldwell et al., 2009; Zhang et al., 2010c, 2012c, d). R values during the $O_3$ seasons are low, ranging from ~0.0 to 0.4. Figure S1 in the supplementary material compares the spatial distributions of forecasted Precip with GPCP Precip in the $O_3$ seasons. The forecasted Precip is largely overpredicted over most areas in the simulation domain. Such large biases and poor correlation can be attributed to three main reasons. First, as reported by Zhang et al. (2010c), the Grell-Devenyi ensemble cumulus parameterization has a tendency to overpredict frequency and the intensity of afternoon convective rainfall. Second, the Purdue Lin microphysics also has a tendency to overpredict cloud ice, graupel, and surface rainfall (Zhang et al., 2012d). Third, as reported in Alapaty et al. (2012), neglecting sub-grid cloud feedbacks to radiation in the cumulus parameterization can overpredict SWDOWN, resulting in unrealistically-large surface forcing for convection thus overpredictions in Precip. Those limitations explain the predicted excessive convection and non-convection rain. While the warm biases in T2 and SWDOWN can lead to higher $O_3$ and $PM_{2.5}$, the positive biases in WS10 and Precip and the negative biases in CF and COT can lead to lower $O_3$ and $PM_{2.5}$. These effects may compensate each other in chemical forecasts.

#### 3.1.2 Winter Seasons

The MBs for T2 in winter are larger than those in the $O_3$ season in 2012 but smaller than those in the $O_3$ seasons in 2013-2014 at the CASTNET and NCDC sites, with a range of 0.7-1.0 °C and 0.8-1.2 °C, respectively. The MB at the SEARCH sites is 1.1 °C during 2014-2015 winter, but -5°C and -3 °C, respectively, in winters during 2012-2013 and 2013-2014. During those winters, heavy snowfall occurred over a large areas in southeastern U.S., particularly during the record-cold winter in Jan-Feb., 2014. The cold biases at the SEARCH sites indicate that the model tends to underestimate the snow melting rates in southeastern U.S. and the effects of urban heat island during winters. RH2 are better forecasted in winters than in $O_3$ season at all sites except for SEARCH during 2012-2013 and 2013-2014 winters during which large cold biases in T2 occur. Similar to the $O_3$ season, WS10 in winters is simulated well at the NCDC sites with MBs of 0.2-0.8 m $s^{-1}$ but largely overpredicted at the CASTNET and SEARCH sites with MBs of 1.6-2.5 and 0.2-1.5 m $s^{-1}$ because of the model's limited capability in resolving surface roughness and topographical features. WD10 forecasts are similar to those in the $O_3$ season at the CASTNET and NCDC sites but worse at the SEARCH sites with MBs of 15.9-30.8°, 38.7-46.7°, and 22.2-46.6° and MAGEs of 89.8-98.3 °C, 92.8-97.5 °C, and 86.7-97.6°C against data from CASTNET, NCDC, and SEARCH, respectively. Comparing to the $O_3$ seasons, MBs of Precip during winters are smaller at the NCDC sites but similar or slightly worse at other sites, with NMBs of 56.1-60.1%, 19.4-42.2%, 36.9-67.3% against data from CASTNET, NCDC, and GPCP, respectively. As shown in Figure S1, the forecasted Precip is overpredicted in winters over most areas in the simulation domain. The spatial distributions of forecasted Precip with GPCP Precip correlate each other better in winter than in the $O_3$ seasons, with higher R values of 0.2-0.7. Similar to the $O_3$ seasons, the model simulates well LWDOWN but moderately overpredicts SWDOWN in winters. Relatively larger underpredictions occur in CF, with NMBs of -23.9% to -18%, leading to slightly larger underpredictions in COT than those during the $O_3$ seasons. Comparing to the $O_3$ seasons, the R values are generally higher for all meteorological variables except for CF and COT during winters, indicating that the model can better simulate the spatial/temporal variations of most meteorological variables during winters than in warm seasons.

### 3.2 Discrete, Spatial, and Temporal Evaluation of Surface Chemical Forecasts

The chemical performance during May 1, 2009-February 28, 2012 has been evaluated in Yahya et al. (2014a). Table 2 summarizes domain-mean performance statistics for chemical species at surface and chemical column abundances during three $O_3$ seasons and three winter seasons during May 1, 2012-February 28, 2015.

#### 3.2.1 Ozone Seasons

During the three $O_3$ seasons in 2012-2014, as shown in Table 2a, the maximum 1-hr $O_3$ mixing ratios are well forecasted with NMBs within ±15% against data at AIRNow, AQS, CASTNET, and SEARCH (except for SEARCH in 2013 where the NMB is 17%). Maximum 8-hr $O_3$ mixing ratios are also well forecasted with NMBs within ±15% in 2012 and 2014 but slightly higher NMBs (15-22%) in 2013 at all sites. Larger overpredictions in maximum 1-hr and 8-hr $O_3$ mixing ratios in 2013 comparing to 2012 and 2014 may be caused by higher warm biases in T2 and greater overpredictions in $NO_x$ (indicated by NMBs of 36% for NO and 56% for $NO_x$ at the SEARCH sites). Higher T2 cause higher emissions of biogenic volatile organic compounds (BVOCs), which also contribute to higher $O_3$ formation. The high positive biases in $NO_2$ and other trace gases such as CO and $SO_2$ at the SAERCH sites may be caused by overestimation of their emissions and also the use of 12-km that cannot represent emissions at those sites. Pan et al. (2014) showed that the use of lower $NO_x$ emissions projected for 2012 than those in 2005 can reduce the positive bias in $O_3$ forecast during July 2011. Although $NO_x$ mixing ratios are also significantly overpredicted in 2014, smaller warm biases in T2 in 2014 than in 2013, resulting in lower BVOCs emissions, and thus smaller $O_3$ overpredictions. Although there are no observed BVOCs emissions and mixing ratios for evaluation, the NMBs of OCs are 15% in 2013 but 1% in 2014 and secondary organic aerosol (SOA) dominates OC in southeastern U.S., supporting higher BVOCs emissions and mixing ratios in 2013 than in 2014. Figure 1 (a) compares several discrete statistics of $O_3$ against data from AIRNow for the six $O_3$ seasons during 2009-2015. The MBs range from -2.8 to 6.9 ppb and -1.8 to 6.9 ppb for maximum 1-hr and 8-hr $O_3$ mixing ratios, respectively. The highest and the second highest NMBs for the maximum 1-hr and 8-hr $O_3$ mixing ratios occur in the $O_3$ seasons in 2013 and 2009, respectively, with NMBs of 15% and 17.0% in 2013 and 9.6% and 8.5% in 2009. The model's skill in terms of NMEs, RMSEs, and R values is overall similar among all six $O_3$ seasons. NMEs, RMSEs, and R values are 19.9-26.7%, 13.1-17.0 ppb, 40-60% for maximum 1-hr $O_3$, and 19.6 to 27.5%, 11.4-14.2 ppb, and 37.5-60% for maximum 8-hr $O_3$. At sites from other $O_3$ measurement networks such as CASTNET, AQS, and SEARCH, the performance statistics for $O_3$ in $O_3$ seasons during 2012-2015 in this work are overall similar to those in 2009-2012 shown in Yahya et al. (2014).

Figure 2 shows forecasted maximum 1-hr and 8-hr $O_3$ mixing ratios overlaid with all available observations during the three $O_3$ seasons in 2012-2014. Figure S2 shows the corresponding spatial distributions of MBs. In 2013, the model overpredicts at many sites in NC, GA, KY, and AL when the observed maximum $O_3$ mixing ratios were below 45 ppb, leading to the largest overpredictions among three $O_3$ seasons and relatively low R values of 0.4-0.5. In 2012, the model captures well the high $O_3$ mixing ratios in MD, northern GA, eastern TN, western OH, northwestern WV, and regions along the border of IN and KY, although it tends to overpredict at some sites in NC, GA, and KY and underpredicts at some sites in IL, IN, and OH. The overpredictions and underpredictions of $O_3$ at different sites over different time periods compensate, leading to relatively good R values of 0.5-0.6. In 2014, the observed $O_3$ mixing ratios are slightly lower than 2012 and 2013, partially because only forecasted results from May-July (MJJ) are averaged (Note that the results in August-September were lost due to the failure of backup drives containing such data). The model captures well the high $O_3$ mixing ratios in NC (including the hot dots in western NC), GA, IN, KY, VA, although it underpredicts a few hot spots in MD and the border regions between OH and IN. Similar spatial distributions and correlation are found for maximum 8-hr mixing ratios, despite slightly larger overpredictions at some sites in NC, VA, WV, GA, and AL in 2012, and at most sites in 2013. Figure 3 compares forecasted and observed hourly $O_3$ mixing ratios at the selected six urban sites. The model reproduces well their observed diurnal and daily variations at all six sites in 2012 and MJJ 2014. Larger discrepancies are found at all cities, in particular, Birmingham, Atlanta, and Louisville.

As shown in Table 2a, forecasted $PM_{2.5}$ concentrations agree very well with the observations from AIRNow with NMBs of -4% to 15% and from STN with NMBs of 9-12%, but moderately overpredicted at the IMPROVE and SEARCH sites, with NMBs of 8-25% and 39-53%, respectively. The $PM_{2.5}$ overpredictions are the results of overpredictions of $SO_4^{2-}$ and $NO_3^-$ (no observations of $NH_4^+$ are available) at the IMPROVE sites, and $SO_4^{2-}$, $NO_3^-$, and $NH_4^+$ at the SEARCH sites. The overpredicted inorganic $PM_{2.5}$ may be caused by overestimates in the emissions of precursors such as $SO_2$, $NO_x$, and $NH_3$. As shown in Table 2a, the $SO_2$ and $NO_2$ mixing ratios at the SEARCH sites are significantly overpredicted with NMBs of 99-725% and 49-56%, respectively. The NO mixing ratios are also overpredicted by 36% and 222% in 2013 and 2014, respectively. The large biases in those precursor gases indicate uncertainties in projected 2009 emissions that are used for RT-AQF during 2009-2015, in particular, such emissions do not reflect the continuous reductions in $SO_2$ and $NO_2$ emissions since 2009 as reported in several studies (e.g., Pan et al., 2014). Warm biases in T2 at all sites also contribute to higher inorganic $PM_{2.5}$ because of higher photochemical oxidation rates during the $O_3$ seasons. Despite overpredictions in WS10 and Precip which tend to reduce $PM_{2.5}$ concentrations, the impacts of overestimated precursor emissions and warm biases on $PM_{2.5}$ formation dominate, leading to a net moderate $PM_{2.5}$ overprediction at all sites. Unlike IMPROVE, STN, and SEARCH, inorganic $PM_{2.5}$ concentrations at the CASTNET sites are mostly underpredicted, likely due in part to the underestimates of anthropogenic of $SO_2$, $NO_x$, and $NH_3$ at remote sites and national parks or the impact of their long-range transport from emissions at nearby urban/rural sites, and in part to the larger wet biases in Precip than at other sites, which scavenges more inorganic $PM_{2.5}$ from the atmosphere at the CASTNET sites. While the model simulates well EC, OC, and TC concentrations at the IMPROVE sites, it underpredicts EC, OC, and thus TC at the SEARCH sites. Such differences are related to different site characteristics (rural/remote sites in the IMPROVE network vs. urban/rural sites in southeastern U.S. in the SEARCH network) as well as possible underestimates of EC and OC emissions at the SEARCH sites during $O_3$ seasons.

Figure 1 (a) compares several discrete statistics of $PM_{2.5}$ against data from AIRNow for the six $O_3$ seasons during 2009-2015. The MBs range from -1.3 to 1.4 μg $m^{-3}$ and NMBs range from -10.1% to 14.7%, indicating a very good performance for $PM_{2.5}$ for all six $O_3$ seasons. The ranges of NMEs, RMSEs, and R values are 35.8-40.4%, 5.1-8.7 μg $m^{-3}$, and 0.3-0.4, respectively. The model's skill in terms of NMEs, RMSEs, and R values is overall similar among all six $O_3$ seasons at sites from AirNow, with slightly higher NMEs but lower RMSEs and R values than forecasted $O_3$ during all six $O_3$ seasons. Compared to the performance statistics for $PM_{2.5}$ at sites from IMPROVE, STN, and SEARCH in $O_3$ seasons during 2009-2011 shown in Yahya et al. (2014), those in $O_3$ seasons during 2012-2014 are worse (particularly at SEARCH sites). Several reasons may contribute to the worse performance of $PM_{2.5}$ during 2012-2014 than during 2009-2011. First, primary PM emissions and the emissions of $PM_{2.5}$ precursors used in the simulations may be higher than actual emissions during those years (resulted from the use of the same emissions as 2009-2011). This leads to higher overpredictions for inorganic PM concentrations during 2012-2014 than during 2009-2011. Second, uncertainties may exist in the spatial allocations of these emissions in both seasons, leading to heterogeneity in model performance at sites from different networks. This uncertainty may explain larger biases in EC and OC predictions during 2012-2014 than during 2009-2011 at SEARCH sites in $O_3$ season. Third, T2 predictions show larger warm biases in $O_3$ seasons during 2012-2014 than 2009-2011 at SEARCH sites, which favor the formation of $(NH_4)_2SO_4$ and thus contribute to higher overpredictions in $PM_{2.5}$ concentrations. As shown in Figures 2 and S2, forecasted $PM_{2.5}$ concentrations agree well spatially with observations in all three $O_3$ seasons, indicating that the relatively low R values may be mainly due to mismatching between forecasted and observed hourly $PM_{2.5}$ concentrations. Such mismatching can be illustrated in Figure 4. For example, in 2012, the model overpredicts $PM_{2.5}$ concentrations at Atlanta when observed concentrations were relatively low (e.g., July 9-September 30, 2012), but underpredicts $PM_{2.5}$ concentrations at Louisville when observed concentrations were relatively high (e.g., June 27-July 10, 2012). In 2013, hourly $PM_{2.5}$ concentrations at Birmingham and Atlanta are largely overpredicted, contributing to large overpredictions and low R values of $PM_{2.5}$ against data from SEARCH.

### 3.2.2 Winter Seasons

As shown in Table 2b and Figure 1(b), unlike the $O_3$ seasons during which $O_3$ mixing ratios are overpredicted in some years, the maximum 1-hr and 8-hr $O_3$ mixing ratios are underpredicted in all winters during 2009-2015. The highest and the second highest NMBs for the maximum 1-hr and 8-hr $O_3$ mixing ratio occur in the winter seasons in 2014-2015 and 2010-2011, respectively, with NMBs of -18.1% and -17.7% in 2014-2015 and -11.9% and -13.5% in 2010-2011. The model's skill in terms of NMEs, RMSEs, and R values is overall similar among all six winter seasons at sites from AirNow, with lower NMEs and RMSEs for both maximum 1-hr and 8-hr $O_3$ but lower R values for maximum 8-hr $O_3$ than the $O_3$ seasons. At sites from other $O_3$ measurement networks such as CASTNET, AQS, and SEARCH, the performance statistics for $O_3$ in winter seasons during 2012-2015 in this work are also overall similar to those in 2009-2012 shown in Yahya et al. (2014). Since T2 is moderately overpredicted at most sites during 2009-2015, the $O_3$ underpredictions are caused in part by large $NO_x$ underpredictions (e.g., an NMB of -67.2% for $NO_2$ in 2014-2015). Cai et al. (2008) evaluated the forecasting skills of an RT-AQF model that uses the CB4 gas-phase mechanism (which is an older version of CB05) and reported much significant underpredictions of OH and $HO_2$ radicals at two sites in New York city during January 2004 compared to July 2004. They attributed such underpredictions to greater uncertainties associated with the CB4 mechanism under low light and low temperature conditions. Their analysis of the predicted and observed CO and $NO_x$ regression slopes also showed a much larger discrepancies between the two slopes in winter than in summer, indicating significant uncertainties associated with the 1999 NEI mobile emission inventories during winter time. In this work, the average observed and forecasted ratios of $CO/NO_x$ at the SEARCH sites for $O_3$ seasons during 2012-2014 are 28.1 and 29.3, respectively. Those for winters during 2012-2015 are 17.1 and 25.0, respectively. The larger differences in the observed and forecasted ratios of $CO/NO_x$ indicate possibly larger uncertainties in mobile emissions in wintertime than warm seasons. As an example, Figure S3 shows the correlation plots for forecasted and observed CO and $NO_x$ at the SEARCH sites in the 2012 $O_3$ season. The forecasted ratios of CO and $NO_x$ are higher than their observed ratios in both the 2012 $O_3$ season and the 2012-2013 winter, with slightly larger differences between the two ratios in winter than in the $O_3$ season. Those uncertainties associated with winter gas-phase chemistry of $HO_x$ radicals and emissions may also contribute to moderate underpredictions in $O_3$ at all sites, and large biases in CO, $SO_2$, and $NO_x$ at the SEARCH sites during winters.

As shown in Table 2a, similar to the $O_3$ seasons, forecasted $PM_{2.5}$ concentrations during winters agree very well with the observations from AIRNow with NMBs of 0.8 to 8.3% and from STN with NMBs of 4.9-8.3%, but moderately overpredicted at the IMPROVE and SEARCH sites, with NMBs of 57.4-59.3% and 59.7-68.4%, respectively. Unlike the $O_3$ seasons, the $PM_{2.5}$ overpredictions are the results of overpredictions of OC with NMBs of 80.8-88.7% and EC with NMBs of 24.4-37.3% at the IMPROVE sites, and OC with NMBs of 24-33% and $SO_4^{2-}$ with NMBs of 16.3-24.7% at the SEARCH sites. The concentrations of $NO_3^-$ are also moderately overpredicted with an NMB of 16.4% in 2013 at the SEARCH sites, contributing to $PM_{2.5}$ overpredictions. Overpredictions in both OC and EC lead to large overpredictions in TC concentrations at the IMPROVE sites. Moderate overpredictions in OC dominate over moderate underpredictions in EC, leading to moderate overpredictions in TC at the SEARCH sites. Those results indicate possible overestimates of primary OC emissions at all types of sites and underestimates of EC emissions at urban/rural sites in southeastern U.S. in the SEARCH network during winter seasons.

Figure 1(b) compares several discrete statistics of $PM_{2.5}$ against data from AIRNow for the six winters during 2009-2015. Similar to the $O_3$ season, the model performs very well for $PM_{2.5}$ for all six winter seasons with the NMBs ranging from -10.2% during 2010-2011 winter to 8.3% during the 2012-2013 winter. As discussed in Yahya et al. (2014a), the underpredictions in 2010-2011 winter are the results of underpredictions in inorganic $PM_{2.5}$, due possibly to underestimates in the emissions of precursors such as $SO_2$, $NH_3$, and $NO_x$ during winters. Other possible reasons for underpredictions of $PM_{2.5}$ during 2010-2011 include positive biases in both Precip and WS10. Different from underpredictions in $PM_{2.5}$ during 2009-2011 winter seasons at AirNow shown in Figure 1 (b) and STN shown in Yahya et al. (2014), the model overpredicts $PM_{2.5}$ during 2012-2015 winter seasons at all sites from AirNow, STN, IMPROVE, and SEARCH, with larger absolute biases at IMPROVE and SEARCH sites than those during 2009-2011 winter seasons. As discussed in Section 3.2.1, the inaccurate primary PM emissions and the emissions of $PM_{2.5}$ precursors, as well as uncertainties in the spatial allocations of those emissions used in the simulations contribute to the worse performance of $PM_{2.5}$ during winter seasons during 2012-2015 than during 2009-2011. Comparing to $PM_{2.5}$ forecasts during the $O_3$ seasons, the $PM_{2.5}$ forecasts during winters show slightly higher NMEs and R values and similar RMSEs. Comparing to $O_3$ forecasts during the winter seasons, $PM_{2.5}$ forecasts during winters show higher NMEs and R values but lower RMSEs. As shown in Figure 2, the model captures well the seasonal variations of $PM_{2.5}$, with higher $PM_{2.5}$ concentrations during $O_3$ seasons than during winters. The model shows better spatial correlations with higher R values during winters than $O_3$ seasons. In particular, the model reproduces several observed hot spots in GA, FL, MO, IN, MD, and LA during the 2012-2013, 2013-2014, and 2014-2015 winters (see Figure 2). As shown in Figure 4, the model reproduces well the observed hourly concentrations of $PM_{2.5}$ at all sites except for Birmingham and Atlanta where overpredictions occur during all three winters.

### 3.3 Categorical Evaluation of Surface Chemical Forecasts

Figure 5 shows categorical evaluation of $O_3$ and $PM_{2.5}$ during all six $O_3$ and winter seasons. The accuracy is high for $O_3$ forecasts during all six $O_3$ and winter seasons, with A values of 94-97.7% during $O_3$ seasons and 98.7-100% during winters. High A values indicate higher percentage of forecasts that correctly predict an exceedance or a non-exceedance, with the number of non-exceedance dominating for both maximum 1-hr and 8-hr $O_3$ mixing ratios. Because the observed and forecasted maximum 1-hr and 8-hr $O_3$ mixing ratios during winters are below the threshold values of 80 ppb and 60 ppb, respectively, no values of CSI, POD, B, and FAR can be calculated. During $O_3$ seasons, the ranges of CSI values are 5.2-15.6 and 9.9-25.3 for maximum 1-hr and 8-hr $O_3$ mixing ratios, respectively. The relatively low CSI values are caused by relatively high false alarm forecasts. Higher CSI values for maximum 8-hr $O_3$ than maximum 1-hr $O_3$ indicate a better skill in forecasting medium range of $O_3$ mixing ratios during the daytime than the daily peak $O_3$ mixing ratios. For the same reason, the model gives higher POD values for maximum 8-hr $O_3$ than for maximum 1-hr $O_3$, with a range of 26.6-46.7 and 17-31.3, respectively. The model gives similarly low B values for both maximum 1-hr and 8-hr $O_3$ mixing ratios. The ranges of B values are 0.6-7.9 and 0.6-4.2 for maximum 1-hr and 8-hr $O_3$, respectively; they are greater than 1 in 2009, 2013, and 2014, indicating overpredictions in those years that are consistent with NMBs shown in Figure 1 (a). The FAR values are high, ranging from 67-96.1% and 48.6-88.9% for maximum 1-hr and 8-hr $O_3$ mixing ratios, respectively. High FAR values indicate that a frequent occurrence of forecasted exceedance that did not occur.

Comparing to $O_3$ forecasts, the A values for $PM_{2.5}$ forecasts are lower, ranging from 70.7-83.2% for $O_3$ seasons and 83.5-85.9% for winters, indicating that accurately forecasting $PM_{2.5}$ is more challenging than forecasting $O_3$. The ranges of CSI values are 10.3-27.9% in $O_3$ seasons and 14.8-22.2% in winters, which are slightly higher than those for $O_3$ forecasts during most seasons. The POD values range from 15.3-40.1% in $O_3$ seasons and 28.5-38.3% in winters, which are similar to those for $O_3$ forecasts during $O_3$ seasons. B values for $PM_{2.5}$ forecasts are smaller than those for $O_3$ forecasts, ranging from 0.6-1.3 in $O_3$ seasons and 0.7-1.2 in winters. FAR values for $PM_{2.5}$ forecasts range from 44.6-75.9% in $O_3$ seasons and 61.3-76.6% in winters. They are lower than FAR values of $O_3$ forecasts during $O_3$ seasons.

### 3.4 Comparisons of Surface $O_3$ and $PM_{2.5}$ Forecasting Skill with Other RT-AQF Models

Tables 3 and 4 compare the discrete and categorical performance evaluation for surface $O_3$ and $PM_{2.5}$ forecasting in this work with those reported over U.S. or a region in the U.S. in the literature. Note that those evaluations did not use the same threshold values and observational data for evaluation nor that they were performed over the same domain and forecasting period. The statistics against AIRNow only and against all datasets are provided for Yahya et al. (2014) and this work because all other evaluations were based on AIRNow. The two sets of performance statistics of max 1-hr and 8-hr $O_3$ from WRF/Chem-MADRID in this work are within the range reported, with better performance based on AIRNow than most other models. For example, NMBs and NMEs of max 8-hr $O_3$ from WRF/Chem-MADRID are -17.7% to 17% and 17.8-33.8%, compared to -2.1% to 25.2% and 18.6-30.4%, respectively, reported in the literature. The performance against AIRNow is better than those against all datasets in this work because the model performs worse when the data from the SEARCH sites are included. For 24-h

$PM_{2.5}$ evaluation using all datasets, while the MBs from this work fall into the reported range, the NMBs for the $O_3$ season and the NMEs for the winter seasons slightly exceed the upper range of report values because of inclusion of all datasets in this work rather than AIRNow only as did in most other work. Using AIRNow only for evaluation, the MBs, RMSEs, NMBs, NMEs for 24-h $PM_{2.5}$ simulated by WRF/Chem-MADRID are -0.5 to 1.4, 5.1-5.7, -4 to 15%, and 36-40%, respectively, during the $O_3$ seasons, and 0.2 to 0.8, 5.5-6.1, 0.8 to 8.3%, and 42.6-47.4%, respectively, during the winter seasons, which are smaller than corresponding values from most other models, namely, -3.2 to 6.2, 5.5-15.9, -21 to 32%, and 41.2-80%, respectively. As shown in Table 4, the model's categorical performance for $PM_{2.5}$ forecasts is comparable to or better than those reported in the literature. The FAR values for max 8-h $O_3$ during the $O_3$ season are slightly beyond the reported range, because of a moderate overprediction in the 2013 $O_3$ season.

### 3.5 Discrete and Spatial Evaluation of Column Chemical Forecasts

Table 2 also shows discrete statistics for column mass abundances of CO, $NO_2$, $SO_2$, and HCHO, TOR, and AOD during $O_3$ seasons and winters during 2012-2015. Column CO and $SO_2$ are moderately underpredicted with NMBs of -42.2% to -36.5% and -55.3% to -54.9%, respectively, in $O_3$ seasons during 2012-2014. The underpredictions are even larger in winter for both species, with NMBs of -50.7% to -48.2% and -77.2% to -73.2%, respectively. As shown in Table 2a, the surface CO and $SO_2$ mixing ratios are overpredicted at the SEARCH sites. The overpredictions at surface but underpredictions in their column masses indicate inaccurate vertical profiles used in their boundary conditions. For example, the BCONs of CO used in the forecasts vary from 72.5-96.4 ppb at the surface layer to 50-65 ppb in upper troposphere during July, and from 125-168 ppb at the surface layer to 50-65 ppb in upper troposphere during January. The vertical profiles of CO derived from MOPPIT over the continental U.S. show a value of 105 ppb at surface and 65 ppb at the tropopause during summer and a value of 125 ppb at surface and 68 ppb at the tropopause during winter (Zhang et al., 2009). While the vertical profiles of CO used reflect the observed seasonal variations, the upper CO mixing ratios used are too low comparing to the MOPITT-derived CO levels in both seasons, and the surface CO mixing ratios are also low in $O_3$ seasons, leading to moderate to significant underpredictions in column CO in $O_3$ seasons and winters. The BCONs of $SO_2$ used in the forecasts vary from 0.04-1.35 ppb at the surface layer to 0.01-0.067 ppb in upper troposphere during July, and from 0.103-1.70 ppb at the surface layer to 0.01-0.067 ppb in upper troposphere during January. Those values are also too low to represent BCONs over southeastern U.S. While column $NO_2$ is moderately underpredicted with NMBs of -35.3% to -33.4% in the $O_3$ seasons, NMBs during winters are much smaller, ranging from -7.9% to 26.2%, indicating a more realistic vertical profile used in winters comparing that in $O_3$ seasons. The BCONs of $NO_2$ used in the forecasts vary from 0.082-0.181 ppb at the surface layer to 0 ppb in upper troposphere during July, and from 0.316-4.23 ppb at the surface layer to 0-0.0057 ppb in upper troposphere during January. Figure 6 (a) shows spatial distributions of column $NO_2$, with overall good spatial correlation and R values of 0.7 and 0.9, in the 2012 $O_3$ season and 2012-2013 winter, respectively. In addition to uncertainties in BCONs, inaccurate/missing emissions and inaccurate vertical allocations of emissions may contribute to the moderate to large underpredictions in column CO, $SO_2$, and $NO_2$. For example, while wildfire and lightening $NO_x$ emissions are included, large uncertainties exist in their magnitudes and spatial distributions. Volcanic eruption and/or degassing may make important contribution to column $SO_2$.

Unlike column CO, $SO_2$, and $NO_2$, Column HCHO is moderately overpredicted with NMBs of 13.1-39.9% in $O_3$ seasons but largely underpredicted with NMBs of -59% to -51.5% in winters. The BCONs of HCHO used in the forecasts vary from 0.599-2.47 ppb at the surface layer to 0 ppb in upper troposphere during July, and from 0.292-0.404 ppb at the surface layer to

0 ppb in upper troposphere during January. The performance statistics show that the BCONs of HCHO are too high in $O_3$ seasons but too low in winters. Another possible source of errors in simulated column HCHO may come from inaccurate biogenic emissions of isoprene, which can produce secondary HCHO through its photochemical oxidation reactions.

TOR is slightly-to-moderately underpredicted with NMBs of -15.4 to -4.5% in $O_3$ seasons but moderately overpredicted with NMBs of 29.9-45.1%. The BCONs of $O_3$ used in the forecasts vary from 26.3-44 ppb in July and from 22.8-39.1 ppb in January at the surface layer to 100.5 ppb in upper troposphere during both months. Although $O_3$ can be formed through photochemical oxidations of precursor gases such as $NO_x$, HCHO, and CO above the surface layer, the mixing ratios of those gases are generally low, particularly in mid-to-upper troposphere. Therefore, the column concentrations of $O_3$ are regulated primarily by BCONs. The performance statistics show that the BCONs of $O_3$ are more realistic in $O_3$ seasons than in winters during which the BCON values are too high to represent $O_3$ vertical profile, leading to moderately overpredicted TOR. AOD is moderately overpredicted with NMBs of 14.4% to 47.6% in $O_3$ seasons, and significantly overpredicted with NMBs of 59.4% to 95.7% in winters. The overpredictions of AOD are the results of overpredictions of $PM_{2.5}$ at surface and also possible overpredictions of $PM_{2.5}$ in upper layers, indicating that the BCONs used for $PM_{2.5}$ composition may be too high in both $O_3$ and winter seasons. Figures 6 (a) and (b) show spatial distributions of TOR and AOD. While forecasted TORs correlate well with OMI-derived TORs with an R value of 0.7 during the 2012 $O_3$ season, they do not correlate in the 2012-2013 winter, indicating a need to adjust the vertical profile of $O_3$ in winter. The forecasted and MODIS-derived AOD agree better spatially in the 2012-2013 winter than in the 2012 $O_3$ seasons.

Two sensitivity simulations are performed to further study the importance of BCONs on column forecasts including a sensitive simulation during August 2012 using satellite-constrained

BCONs for CO and a sensitive simulation during December 2012 using satellite-constrained BCONs for $O_3$. Those sensitivity simulations show large improvement in simulated column CO and TOR. Figures 7 and 8 compare the spatial distributions of satellite-derived CO and TOR and the two simulations in August 2012 and December 2012, respectively. The use of satellite-constrained BCONs for CO and TOR improves the simulated CO and TOR substantially. The MB, NMB, and NME of CO from the sensitivity simulation are -0.2, -10.6%, and 18.2%, respectively, comparing to -0.8, -40.6%, and 40.8% from the baseline simulation. The MB, NMB, and NME of TOR from the sensitivity simulation are -0.2, -0.01%, and 0.1%, respectively, comparing to 11.7, 44.8%, and 44.8% from the baseline simulation. Similar improvements are expected for other column variables including column $NO_2$ in O3 season and column $SO_2$ and HCHO in both season.

**3.6 Trend analysis for multiple years**

Given interannual variability in climate and emissions, it is useful to assess the robustness of the model in forecasting the relative changes in terms of magnitudes and signs under different climate conditions, as well as the interannual variability from the year of reference.

**3.6.1 Meteorological Variables**

Figure 9 compares observed and simulated variation trends for T2, Precip, WS10, SWDOWN, LWDOWN, and CF. Note that the trends for SWDOWN, LWDOWN, and CF are only plotted for the 2011-2014 $O_3$ seasons and 2011-2015 winters because the upper layer model outputs during 2009-2010 were not available due to failures of backup drives containing such data. The model forecasts well the observed changes in terms of both magnitudes and signs, as well as the interannual variability of T2 and WS10 in both $O_3$ and winter seasons relative to their values in 2009. It simulates reasonably well for the observed interannual variability of Precip at the CASTNET sites, but not well for the observed changes in magnitudes of Precip. The changes in terms of magnitudes and signs as well as interannual variability relative to their values in 2011-2012 for SWDOWN, LODOWN, and CF are well captured in winters, but in $O_3$ seasons, while the model reproduces well both the magnitudes of the changes and the interannual variability of LWDOWN, and the interannual variability of SWDOWN and CF, it overpredicts the increases in SWDOWN but underpredicts the increases in CF.

**3.6.2 Chemical Variables**

Figure 10 compares observed and simulated variation trends for surface $O_3$ mixing ratios, surface $PM_{2.5}$ concentrations, column mass abundances of CO, $SO_2$, $NO_2$, and HCHO, TOR, and AOD. Note that the trends for column variables are only plotted for the $O_3$ seasons during 2011-2014 and winters during 2011-2015 for the aforementioned reason. Relative to the 2009 $O_3$ season, the observed $O_3$ mixing ratios from AIRNow are higher during $O_3$ seasons in 2010-2012 and 2014 but are lower in $O_3$ season in 2013, this trend is not well captured by the model, as it forecasts slightly lower $O_3$ in 2010 and 2014 $O_3$ seasons, and slightly higher $O_3$ in other $O_3$ seasons. While $PM_{2.5}$ forecasts during $O_3$ seasons generally follow the observed trends, large differences occur in the magnitudes of the changes, with greater increases in 2010-2011 but greater decreases during 2012-2014. Although the differences in the magnitude of the changes for forecasted $O_3$ are smaller in winters than in $O_3$ seasons, the observed and forecasted $O_3$ mixing ratios change in different directions, i.e., the observed $O_3$ mixing ratios either increase or decrease slightly, the forecasted $O_3$ mixing ratios continue to decline during 2010-2014 winters. The large differences in magnitudes and signs remain in forecasted and observed trends of $PM_{2.5}$ concentrations during winters comparing to $O_3$ seasons, however, the forecasted and observed changes of $PM_{2.5}$ concentrations are within ±10%. The forecasted and observed changes in column CO are small, within 3% in $O_3$ seasons and within 1% in winters but they have different signs. The forecasted column $SO_2$ captures well the observed trends in magnitudes and signs for column $SO_2$ in the $O_3$ seasons, larger differences exist in both magnitude and sign in winters. Although the forecasted and observed changes for column $NO_2$ are generally within 10%, larger differences exist in magnitude and sign comparing to those for column CO and $SO_2$. The OMI-derived TORs decrease in $O_3$ seasons during 2012-2014, and increase during winters of 2012-2013, 2013-2014, and 2014-2015. Among all column gases, the forecasted column HCHO shows the largest differences in the $O_3$ seasons in terms of both magnitude and sign. Uncertainties in satellite retrievals of column HCHO may contribute in part to such large discrepancies between forecasts and satellite-derived observations. For example, De Smedt et al. (2008) reported errors in HCHO retrievals of (0.5-2.0)$\times 10^{15}$ molecules $cm^{-2}$, which are on the same order of magnitudes or even larger than the MBs in the forecasted HCHO column for all seasons. The differences in magnitude of the changes in column HCHO are smaller in winters but the signs are opposite in 2012-2013. The forecasted TORs show a slight decrease in 2012 $O_3$ season and slight increases in 2013 and 2014 $O_3$ seasons, and slight decreases in winters of 2012-2013 and 2013-2014 as well as a slight increase in winter of 2014-2015.

The forecasted AOD captures the decreasing trend during $O_3$ seasons of 2012-2014, but with much smaller magnitudes of the changes (by up to 18% versus 40%, respectively). While MODIS-derived AOD shows a large decrease (by up to 22%) from 2011-2012 winter, the forecasted AOD shows a large increase (by up to 23%) in winter. The relatively large discrepancies between satellite-derived and forecasted column variables such as column $NO_2$ and HCHO, TOR, and AOD in both $O_3$ and winter seasons, and column $SO_2$ in winters indicate a need to adjust the vertical profiles of these gases and PM composition in the BCONs.

## Summary and Conclusion

An online-coupled meteorology-chemistry model, WRF/Chem-MADRID, has been deployed for RT-AQF in southeastern U.S. since 2009 for six $O_3$ seasons and six winters. A comprehensive evaluation of meteorological and chemical forecasts is performed using surface and satellite-derived observations in terms of spatial distribution, temporal variation, and discrete and categorical performance statistics. The meteorological evaluation shows moderate to large biases for T2, RH2, WS10, WD10, Precip, SWDOWN, CF, and COT, indicating some limitations in the YSU PBL scheme, the Monin-Obukhov surface layer scheme, the Purdue Lin cloud microphysics module, and the Grell-Devenyi ensemble scheme. In particular, uncertainties exist in the model treatments of PBL processes (e.g., inaccurate representations of surface drag), the dynamics, thermodynamics, and microphysics of clouds, as well as aerosol-radiation-cloud-precipitation interactions (e.g., the missing treatments of sub-grid cloud feedbacks to radiation). Since the forecasts do not use data assimilation, the agreement between meteorological forecasts and observations is not expected to be comparable with the simulations that use data assimilation. The meteorological forecasts for most variables except for WS10, Precip, and COT in this work are therefore deemed to be acceptable. While updating WRF/Chem-MADRID based on the latest WRF/Chem version should help reduce some of those uncertainties with updated schemes and treatments, continuous development and improvement of PBL schemes and cloud parameterizations are important future work to improve meteorological forecasts, which will in turn improve chemical forecasts.

WRF/Chem-MADRID shows consistently good skills for $O_3$ and $PM_{2.5}$ forecast in terms of both categorical and discrete statistics during 2009-2015. It performs well in both $O_3$ and winter seasons with most NMBs within ±15% for $O_3$ forecasts against observations from AIRNow, AQS, CASTNET, and SEARCH. The NMBs for $PM_{2.5}$ forecasts are within ±15% against observations from AIRNow and STN, but larger (up to ±68%) against observations from

IMPROVE and SEARCH. Larger biases are also found for secondary $PM_{2.5}$ against surface observations at IMPROVE and SEARCH, and also for some column variables (e.g., column $NO_2$ in $O_3$ seasons, TOR, column HCHO, and AOD in winters, and COT and column CO and $SO_2$ in both $O_3$ and winter seasons) against satellite data. These biases are due possibly to uncertainties in simulated meteorology (e.g., T2, Precip, and WS10), emissions (e.g., biogenic/wildfire emissions and winter mobile emissions), and BCONs (e.g., inaccurate BCONs for seasonal and inter-annual variations for CO, $NO_2$, $SO_2$, $O_3$, HCHO, and $PM_{2.5}$ composition), as well as limitations in chemical and aerosol treatments (e.g., the production of OH radicals from CB05 in winter, aerosol thermodynamic partitioning, and SOA formation). Comparison of model performance during 2012-2015 with that during 2009-2012 shows that the inaccurate primary PM emissions and the emissions of $PM_{2.5}$ precursors, as well as uncertainties in the spatial allocations of those emissions used in the simulations contribute to the worse performance of $PM_{2.5}$ during both $O_3$ and winter seasons during 2012-2015 than during 2009-2012.

Despite those biases, the model's performance in terms of surface $O_3$ and $PM_{2.5}$ forecasts is overall consistent with or better than the performance of other RT-AQF models reported in the literature for different periods over different domains. Although the model shows overall good skills for meteorological and chemical forecasts at the surface, inaccurate representations of species vertical profiles can potentially affect both meteorological and chemical forecasts at the surface because of turbulent mixing and convective cloud updraft and downdraft movements and because of the feedbacks of radiative species (e.g., $O_3$, $NO_2$, HCHO, and PM composition) to the radiation calculation in the model. The impacts of chemical BCONs on air quality simulations have been shown in several studies (e.g., Giordano et al., 2015; Yahya et al., 2015b) and in this work. Therefore, the vertical profiles of BCONs of those species should be constrained with satellite-derived observations to more realistically represent vertical and seasonal variations.

Forecasted changes in most meteorological variables except for CF generally reproduce well the observed trends in terms of magnitude and sign and interannual variability. While small changes occur in observed seasonal-mean maximum 1-hr and 8-hr $O_3$ concentrations from AIRNow since 2009, those for $PM_{2.5}$ show greater decreases and stronger inter-annual variabilities than $O_3$, reflecting the effects of emission reductions since 2009. Forecasted $O_3$ levels show weaker inter-annual variabilities than observed $O_3$ levels during all $O_3$ and winter seasons. Forecasted $PM_{2.5}$ levels resemble their observed increasing trends from 2009 to 2011 and declining trend from 2011 to 2014 during $O_3$ seasons and remain nearly constant during winter. Such variabilities are mainly attributed to changes in meteorology and meteorology-dependent biogenic and wildfire emissions. Largest discrepancies are found in the forecasted and observed changes in AOD and column gases including CO, $NO_2$, $SO_2$, HCHO, and $O_3$, due mainly to inaccurate representations of the vertical profiles of the BCONs of those gases and PM composition. More accurate meteorological forecasts, anthropogenic emissions, and meteorology-dependent emissions (e.g., biogenic, wildfire, and volcanic), upper BCONs for chemical species, and model treatments of chemical and aerosol processes should improve the model's ability in reproducing not only the observations but also the interannual and inter-seasonal variation trends in terms of magnitude and sign for major chemical species of concerns.

When resources become available, several limitations in this work should be addressed. These may include the code migration of WRF/Chem-MADRID into the latest version of WRF/Chem, the refinement of configurations using available latest physics and chemistry options (e.g., the use of urban canopy model, updated surface roughness treatments, and the multi-scale cumulus parameterization), and updates in emissions and lateral BCONs including using real-time forecasted emissions, and more realistic BCONs derived from satellite retrievals or dynamic BCONs from a validated global RT-AQF model.

## Acknowledgements

This work is not currently funded by any agency. The development and initial deployment of WRF/Chem-MADRID for RT-AQF was supported by the NSF Career Award (#ATM-0348819) (until 2011) and the U.S. Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) (# DW13921548) (until 2008). The continuous application and analysis since 2011 have been performed using students' and faculty member's own time although those students were supported under the EPA-Science to Achieve Results (STAR) program (#R83337601), or the National Science Foundation EaSM program (AGS-1049200), or with the fellowship awards from different sources. The authors acknowledge high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR's CISL, sponsored by the NSF and Information System Library. Thanks are due to some contributions of students at NCSU: Brittney Mitchell, Masilin Gudoshava, and Pornpan Uttamang. AIRNow, AIRS-AQS, CASTNET, and IMPROVE data from the U.S. EPA website; NCDC data from their website; SEARCH data from the website of Atmospheric Research & Analysis, Inc.; STN data for 2012-2015 from the Colorado State University website; STN data for 2010 was provided by Wyat Appel, U.S. EPA; GPCP data from NOAA website; CERES data and MODIS data from NASA website; and OMI data from TEMIS website.

## References

Abdul-Razzak, H., and S. J. Ghan. 2002. A parameterization of aerosol activation, 3. Sectional representation. *J. Geophys. Res.* 107 (D3). doi:10.1029/2001JD000483.

Alapaty, K., J. A. Herwehe, T. L. Otte, C. G. Nolte, O. R. Bullock, M. S. Mallard, J. S. Kain, and J. Dudhia. 2012. Introducing subgrid-scale cloud feedbacks to radiation for regional meteorological and climate modeling. *Geophys. Res. Lett. 39*:L24809. doi:10.1029/2012GL054031.

Baklanov, A., K.H. Schlünzen, P. Suppan, J. Baldasano, D. Brunner, S. Aksoyoglu, G. Carmichael, J. Douros, J. Flemming, R. Forkel, S. Galmarini, M. Gauss, G. Grell, M. Hirtl, S. Joffre, O. Jorba, E. Kaas, M. Kaasik, G. Kallos, X. Kong, U. Korsholm, A. Kurganski, J. Kushta, U. Lohmann, A. Mahura, A. Manders-Groot, A. Maurizi, N. Moussiopoulos, S.T. Rao, N. Savage, C. Seigneur, R.S. Sokhi, E. Solazzo, S. Solomos, B. Sørensen, G. Tsegas, E. Vignati, B. Vogel, and Y. Zhang, 2014. Online Coupled Regional Meteorology-Chemistry Models in Europe: Current Status and Prospects, *Atmospheric Chemistry and Physics*, 14, 317-398, doi:10.5194/acp-14-317-2014.

Barnard, W. R., and E. Sabo. 2008. Documentation of the Base G2 and Best & Final 2002 Base Year, 2009 and 2018 Emission Inventories for VISTAS. Final Report prepared by MACTEC Engineering and Consulting, Inc. for the Visibility Improvement State and Tribal Association of the Southeast (VISTAS), March 14.

Brunner, D., N. Savage, O. Jorba, B. Eder, L. Giordano, P. Makar, A. Badia, A. Balzarini, R. Baró, R. Bianconi, C. Chemel, G. Curci, R. Forkel, P. Jiménez-Guerrero, M. Hirtl, A. Hodzic, L. Honzak, U. Im, C. Knote, A. Manders-Groot, L. Neal, J. L. Pérez, G. Pirovano, R. San Jose, W. Schröder, R. S. Sokhi, D. Syrakov, A. Torian, J. Werhahn, R. Wolke, E. van Meijgaard, K. Yahya, R. Zabkar, Y. Zhang, C. Hogrefe, and S. Galmarini. 2014. Comparative analysis of meteorological performance of coupled chemistry-meteorology models in the context of AQMEII phase 2. *Atmos. Environ*. doi:10.1016/j.atmosenv.2014.12.032.

Cai, C., C. Hogrefe, P. Katsafados, G. Kallos, M. Beauharnois, J. J. Schwab, X. Ren, W. Brune, X. Zhou, Y. He, and K. Demerjian. 2008. Performance evaluation of an air quality forecast modeling system for a summer and winter season – Photochemical oxidants and their precursors. *Atmos. Environ.* 42:8585-8599. doi:10.1026/j.atmosenv.2008.08.029.

Caldwell, P., H. S. Chin, D. C. Bader, and G. Bala. 2009. Evaluation of a WRF dynamical downscaling simulation over California. Clim. Ch. 95:499-521. doi: 10.1007/s10584-009-9583-5.

Carmichael, G. R., Y. Tang, G. Kurata, I. Uno, D. G. Streets, J. H. Woo, H. Huang, J. Yienger, B. Lefer, R. E. Shetter, D. R. Blake, A. Fried, E. Apel, F. Eisele, C. Cantrell, M. A. Avery, J. D. Barrick, G. W. Sachse, W. L. Brune, S. T. Sandholm, Y. Kondo, H. B. Singh, R. W. Talbot, A. Bandy, A. D. Clarke, and B. G. Heikes. 2003. Regional-scale chemical transport modeling in support of intensive field experiments: Overview and analysis of the TRACE-P observations. *J. Geophys. Res*. 108 (D21):8823. doi:10.1029/2002JD003117.

Chai, T., H.-C. Kim, P. Lee, D. Tong, L. Pan, Y. Tang, J. Huang, J. McQueen, M. Tsidulko, and I. Stajner, 2013. Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and NO2 measurements. Geosci. Model Dev., 6, 1831–1850, doi:10.5194/gmd-6-1831-2013.

Chen, F., and J. Dudhia. 2001. Coupling an advanced land surface–hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Weather Rev.* 129 (4):569–585. doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Chen, J., J. Vaughan, J. Avise, S. O'Neill, and B. Lamb, 2008. Enhancement and evaluation of the AIRPACT ozone and $PM_{2.5}$ forecast system for the Pacific Northwest. *J. Geophys. Res, 113,* D14305, doi:10.1029/2007JD009554.

Cheng, W. Y., and W. J. Steenburgh. 2005. Evaluation of Surface Sensible Weather Forecasts by the WRF and Eta Models over the Western United States. *Weather Forecast.* 20 (5):812-821. doi: http://dx.doi.org/10.1175/WAF885.1.

Chou, M. D., M. J. Suarez, C. Ho, M. M. Yan, and K. Lee. 1998. Parameterizations for Cloud Overlapping and Shortwave Single-Scattering Properties for Use in General Circulation and Cloud Ensemble Models. *J. Climate.* 11 (2):202-214. doi: http://dx.doi.org/10.1175/1520-0442(1998)011<0202:PFCOAS>2.0.CO;2.

Chuang, M. T., Y. Zhang, and D. W. Kang. 2011. Application of WRF/Chem-MADRID for Real-Time Air Quality Forecasting over the Southeastern United States. *Atmos. Environ.* 45 (34):6241-6250. doi:10.1016/j.atmosenv.2011.06.071.

De Smedt, I., J.-F. Müller, T. Stavrakou, R. J. van der, A, H. J. Eskes, and M. Van Roozendael. 2008. Twelve years of global observations of formaldehyde in the troposphere using GOME and SCIAMACHY sensors. *Atmos. Chem. Phys.*, *8*(16), 4947-4963.

Djalalova, I., J. Wilczak, S. McKeen, G. Grell, S. Peckham, M. Pagowski, L. DelleMonache, J. McQueen, Y. Tang, P. Lee, J. McHenry, W. Gong, V. Bouchet, and R. Mathur. 2010. Ensemble and bias-correction techniques for air quality model forecasts of surface $O_3$ and $PM_{2.5}$ during the TEXAQS-II experiment of 2006. *Atmos. Environ.* 44:455-467. doi: 10.1016/j.atmosenv.2009.11.007.

Eder, B., D. Kang, R. Mathur, S. Yu, and K. Schere. 2006. An operational evaluation of the Eta-CMAQ air quality forecast model. *Atmos. Environ.* 40:4894-4905. doi:10.1016/j.atmosenv.2005.12.062.

Eder, B., D. Kang, R. Mathur, J. Pleim, S. Yu, T. Otte, and G. Pouliot. 2009. A performance evaluation of the National Air Quality Forecast Capability for the summer of 2007. *Atmos. Environ.* 43:2312-2320. doi: 10.1016/j.atmosenv.2009.01.033.

Emmons, L. K., D. P. Edwards, M. N. Deeter, J. C. Gille, T. Campos, P. Nédélec, P. Novelli, and G. Sachse. 2009. Measurements of Pollution In The Troposphere (MOPITT) validation through 2006. *Atmos. Chem. Phys.* 9 (5):1795–1803. doi:10.5194/acp-9-1795-2009.

Fahey, K.M., and S. N. Pandis. 2001. Optimizing model performance: variable size resolution in cloud chemistry modeling. *Atmos. Environ.* 35:4471-4478. doi:10.1016/S1352-2310(01)00224-2.

Giordano, L., D. Brunner, J. Flemming, U. Im, C. Hogrefe, R. Bianconi, A. Badia, B. Alessandra, R. Baro, C. Chemel, G. Curci, R. Forkel, P. Jimenez-Guerrero, M. Hirtl, A. Hodzic, L. Honzak, O. Jorba, C. Knote, J. J. P. Kuenen, P. A. Makar, A. Manders-Groot, L. Neal, J. Luis Perez, G. Pirovano, G. Pouliot, R. San Jose, N. Savage, W. Schroder, R. S. Sokhi, D. Syrakov, A. Torian, P. Tuccella, J. Werhahn, R. Wolke, K. Yahya, R. Žabkar, Y. Zhang, and S. Galmarini. 2015, Assessment of the MACC/IFS-MOZART model and its influence as chemical boundary conditions in AQMEII phase 2. *Atmos. Environ.* doi:10.1016/j.atmosenv.2015.02.034.

Gong, S., L. A. Barrie, and J. P. Blanchet. 1997. Modeling sea salt aerosols in the atmosphere: 1. Model development. *J. Geophys. Res.* 102:3805–3818. doi:10.1029/96JD02953.

Gong, S. L., L. A. Barrie, and M. Lazare. 2003. Canadian Aerosol Module (CAM): A size-egregated simulation of atmospheric aerosol processes for climate and air quality models 2. Global sea-salt aerosol and its budgets. *J. Geophys. Res.* 107 (D24):4779. doi:10.1029/2001JD002004.

Grell, G. A., and D. Devenyi. 2002. A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Let.* 29 (14): 1693. doi:10.1029/2002GL015311.

Grell G. A., S. E. Peckham, R. Schmitz, S. A. McKeen, G. Frost, W. C. Skamarock, and B. Eder. 2005. Fully coupled 'online' chemistry in the WRF model. *Atmos. Environ.* 39:6957-6976. doi:10.1016/j.atmosenv.2005.04.027.

Hogrefe, C., W. Hao, K. Civerolo, J.-Y. Ku, G. Sistla, R.S. Gaza, L. Sedefian, K. Schere, A. Gilliland, and R. Mathur. 2007. Daily simulation of ozone and fine particulates over New York State: findings and challenges. *J. Appl. Meteor. Climatol*. *46* (7):961-979. doi:10.1175/JAM2520.1.

Hong, S.-Y., Y. Noh, and J. Dudhia. 2006. A New Vertical Diffusion Package with an Explicit Treatment of Entrainment Processes. *Mon. Weather Rev.* 134:2318-2341. doi:10.1175/MWR3199.1.

Kang, D., B. K. Eder, A. F. Stein, G. A. Grell, S. E. Peckham, and J. McHenry. 2005. The New England Air Quality Forecasting Pilot Program: Development of an evaluation protocol and performance benchmark. *J. Air & Waste Manage. Assoc. 55*:1782-1796. doi:10.1080/10473289.2005.10464775.

Kukkonen, J., T. Olsson, D. M. Schultz, A. Baklanov, T. Klein, A. I. Miranda, A. Monteiro, M. Hirtl, V. Tarvainen, M. Boy, V.-H. Peuch, A. Poupkou, I. Kioutsioukis, S. Finardi, M. Sofiev, R. Sokhi, K. E. J. Lehtinen, K. Karatzas, R. San Jose´, M. Astitha, G. Kallos, M. Schaap, E. Reimer, H. Jakobs, and K. Eben. 2012. A review of operational, regional-scale, chemical weather forecasting models in Europe. *Atmos. Chem. Phys.* 12 (1):1-87. doi:10.5194/acp-12-1-2012.

Lee, P., D. Kang, J. McQueen, M. Tsisulko, M. Hart, G. DiMego, N. Seaman, and P. Davidson (2008), Impact of domain size on modeled ozone forecast for the Northeastern United states, *Journal of Applied Meteorology and Climatology*, *47*, 443-461, doi: 10.1175/2007JAMC1408.1.

Lin, Y.-L., R. D. Farley, and H. D. Orville. 1983. Bulk Parameterization of the Snow Field in a Cloud Model. *J. Appl. Met. Clim.* 22:1065-1092. doi:10.1175/1520-0450(1983)022.

Makar, P. A., et al., 2009. Modelling the impacts of ammonia emissions reductions on North American air quality. *Atmos. Chem. Phys.*, *9*, 7183–7212. doi:10.5194/acp-9-7183-2009.

Mangold, A., H. De Backer, B. De Paepe, S. Dewitte, I. Chiapello, Y. Derimian, M. Kacenelenbogen, J.-F. Léon, N. Huneeus, M. Schulz, D. Ceburnis, C. O'Dowd, H. Flentje, S. Kinne, A. Benedetti, J. J. Morcrette, and O. Boucher. 2011. Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: 3. Evaluation by means of case studies. *J. Geophys. Res*. 116:D03302. doi:10.1029/2010JD014864.

Martin, R. V. 2008. Satellite remote sensing of surface air quality. *Atmos. Environ.* 42 (34):7823-7843. doi:10.1016/j.atmosenv.2008.07.018.

Mass, C. F., and D. Ovens. 2011. Fixing WRF's High Speed Wind Bias: A New Subgrid Scale Drag Parameterization and the Role of Detailed Verification. 91st AMS Annual Meeting, Seattle, WA, 23-27 Jan 2011, p. 2011.2019B.2016.

McHenry, J. N., W. F. Ryan, N. L. Seaman, C. J. Coats Jr., J. Pudykiewics, S. Arunachalam, and J. M. Vukovich. 2004. A real-time Eulerian photochemical model forecast system: Overview and initial ozone forecast performance in the Northeast U.S. corridor. *Bull. Amer. Meteor. Soc.* 85:525-548. doi: http://dx.doi.org/10.1175/BAMS-85-4-525.

McKeen, S., J. Wilczak, G. Grell, I. Djalalova, S. Peckham, E.-Y. Hsie, W. Gong, V. Bouchet, S. Menard, R. Moffet, J. McHenry, J. McQueen, Y. Tang, G. R. Carmichael, M. Pagowski, A. Chan, T. Dye, G. Frost, P. Lee, and P. Mathur. 2005. Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *J. Geophys. Res.* 110:D21307. doi:10.1029/2005JD005858.

McKeen, S., G. Grell, S. Peckham, J. Wilczak, I. Djalalova, E.-Y. Hsie, G. Frost, J. Peischl, J. Schwarz, R. Spackman, J. Holloway, J. de Gouw, C. Warneke, W. Gong, V. Bouchet, Gaudreault, J. Racine, J. McHenry, J. McQueen, P. Lee, Y. Tang, G. R. Carmichael, and R. Mathur. 2009. An evaluation of real-time air quality forecasts and their urban emissions over eastern Texas during the summer of 2006 Second Texas Air Quality Study field study. *J. Geophys. Res.* 114:D00F11. doi:10.1029/2008JD011697.

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., Yu, S., 2007. Evaluation of several $PM_{2.5}$ forecast models using data collected during the ICARTT/NEAQS 2004 field study. *J. Geophys. Res.* 112, D10S20. http://dx.doi.org/10.1029/2006JD007608.

Mlawer, E.J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough. 1997. Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.* 102:16663-16682. doi: 10.1029/97JD00237.

Olerud, D. T., A. P. Sims, and M. Abraczinskas. 2005. Annual meteorological modeling in support of visibility improvement in the southeast US. Presented at AMS $85^{th}$ Annual Conference/$7^{th}$ Conference on Atmospheric Chemistry, American Meteorological Society, San Diego, CA.

Pan, Li, Daniel Tong, Pius Lee, H.-C. Kim, Tianfeng Chai, 2014. Assessment of $NO_x$ and $O_3$ forecasting performances in the U.S., National Air Quality Forecasting Capability before and after the 2012 major emissions updates, *Atmos. Environ.* 95, 610-619.

Penrod, A., Y. Zhang, K. Wang, S. Wu, and L. R. Leung. 2014. Impacts of future climate and emission changes on U.S. air quality. *Atmos. Environ.* 89:533-547. doi:10.1016/j.atmosenv.2014.01.001.

Remer, L. A., Y. J. Kaufman, D. Tanré, S. Mattoo, D. A. Chu, J. V. Martins, R.-R. Li, C. Ichoku, R. C. Levy, R. G. Kleidman, T. F. Eck, E. Vermote, and B. N. Holben. 2005. The MODIS aerosol algorithm, products, and validation. *J. Atmos. Sci.* 62:947–973. doi:10.1175/JAS3385.1.

Shaw, P., 2008. Application of aerosol speciation data as an in situ dust proxy for validation of the Dust Regional Atmospheric Model (DREAM). Atmos. Environ. 42, 7304-7309.

Takigawa, M., M. Niwano, H. Akimoto, and M. Takahashi. 2007. Development of a One-way Nested Global-regional Air Quality Forecasting Model. *Scientific Online Letters on the Atmosphere (SOLA)*. 3:81–84. doi:10.2151/sola.2007–021.

Tesche, T. W., and C. Tremback. 2002. Operational Evaluation of the MM5 Meteorological Model over the Continental United States: Protocol for Annual and Episodic Evaluation, Draft Protocol prepared under Task Order 4TCG-68027015 for the Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency.

U.S. EPA. 2001. Guidance for demonstrating attainment of air quality goals for $PM_{2.5}$ and regional haze, draft 2.1, January 2, 2001, the U.S. Environmental Protection Agency, Office of Air and Radiation/Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711.

Wang, L. T., Z. Wei, J. Yang, Y. Zhang, F. F. Zhang, J. Su, C. C. Meng, and Q. Zhang. 2014. The 2013 Severe Haze over the Southern Hebei, China: Model Evaluation, Source apportionment, and Policy implications. *Atmos. Chem. Phys.* 14:3151–3173. doi:10.5194/acp-14-3151-2014.

Wang, K., K. Yahya, Y. Zhang, C. Hogrefe, G. Pouliot, C. Knote, A. Hodzic, R. San Jose and J. L. Perez, P. J. Guerrero, R. Baro, and P. Makar. 2014. A Multi-Model Assessment for the 2006 and 2010 Simulations under the Air Quality Model Evaluation International Initiative

(AQMEII) Phase 2 over North America, Part 2. Evaluation of Column Variable Predictions Using Satellite Data. *Atmos. Environ.* doi:10.1016/j.atmosenv.2014.07.044.

Wielicki, B. A., B. R. Barkstrom, E. F. Harrison, R. B. Lee, G. L. Smith, and J. E. Cooper. 1996. Clouds and the Earth's Radiant Energy System (CERES): An earth observing system experiment. *Bull. Amer. Meteor. Soc.* 77:853–868. doi: http://dx.doi.org/10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2.

Yahya, K., Y. Zhang, and J. M. Vukovich. 2014a. Real-Time Air Quality Forecasting over the Southeastern United States using WRF/Chem-MADRID: Multiple-Year Assessment and Sensitivity Studies. *Atmos. Environ.* 92:318-338. doi:10.1016/j.atmosenv.2014.04.024.

Yahya, K., K. Wang, M. Gudoshava, T. Glotfelty, and Y. Zhang. 2014b. 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.

Yahya, K., K. Wang, Y. Zhang, and T. E. Kleindienst. 2015a. Application of WRF/Chem over North America under the AQMEII Phase 2. Part II. Evaluation of 2010 Simulation and Responses of Air Quality and Meteorology-Chemistry Interactions to Changes in Emissions and Meteorology from 2006 to 2010. *Geosci. Model Dev.* 8:2095-2117. doi:10.5194/gmd-8-2095-2015.

Yahya, K., J. He, and Y. Zhang. 2015b. Multi-Year Applications of WRF/Chem over Continental U.S.: Model Evaluation, Variation Trend, and Impacts of Boundary Conditions. *J. Geophys. Res.* in review.

Yarwood, G., S. Rao, M. Yocke, G. Z.Whitten, and S. Reyes. 2005. Final Report Updates to the Carbon Bond Chemical Mechanism CB05. Report to the U.S. Environmental Protection Agency, December 2005.

Yu, S., B. Eder, R. Dennis, S.-H. Chu, and S. Schwartz. 2006. New unbiased symmetric metrics for evaluation of air quality models. *Atmos. Sci. Lett.* 7:26-34. doi: 10.1002/asl.125.

Yu, S. C., R. Mathur, D. Kang, K. Schere, J. Pleim, and T. L. Otte. 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. *J. Geophys. Res. 112*:D12S14. doi:10.1029/2006JD007715.

Yu, S., Mathur, K. Schere, D. Kang, J. Pleim, J. Young, D. Tong, G. Pouliot, S. A. McKeen, and S. T. Rao. 2008. Evaluation of real-time $PM_{2.5}$ forecasts and process analysis for $PM_{2.5}$ formation over the eastern U.S. using the Eta-CMAQ forecast model during the 2004 ICARTT Study. *J. Geophys. Res.* 113:D06204. doi:10.1029/2007JD009226.

Yu, S., R. Mathur, G. Sarwar, D. Kang, D. Tong, G. Pouliot, and J. Pleim. 2010. Eta-CMAQ air quality forecasts for $O_3$ and related species using three different photochemical mechanisms (CB4, CB05, SAPRC-99): comparisons with measurements during the 2004 ICARTT study. *Atmos. Chem. Phys.* 10:3001-3025. doi:10.5194/acp-10-3001-2010.

Zhang, Y., 2008. Online Coupled Meteorology and Chemistry models: History, Current Status, and Outlook, *Atmospheric Chemistry and Physics,* 8, 2895-2932

Zhang, Y., B. Pun, K. Vijayaraghavan, S.-Y. Wu, C. Seigneur, S. Pandis, M. Jacobson, A. Nenes, and J. H. Seinfeld. 2004. Development and Application of the Model of Aerosol Dynamics, Reaction, Ionization and Dissolution (MADRID). *J. Geophys. Res.* 109:D01202. doi:10.1029/2003JD003501.

Zhang, Y., P. Liu, B. Pun, and C. Seigneur. 2006. A comprehensive performance evaluation of MM5-CMAQ for the Summer 1999 Southern Oxidants Study episode-Part I: Evaluation protocols, databases, and meteorological predictions. *Atmos. Environ.* 40:4825-4838. doi:10.1016/j.atmosenv.2005.12.043.

Zhang, Y., K. Vijayaraghavan, X.-Y. Wen, H. E. Snell, and M.Z. Jacobson. 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. *J. Geophys. Res.* 114:D22304. doi:10.1029/2009JD011898.

Zhang, Y., Y. Pan, K. Wang, J. D. Fast, and G. A. Grell. 2010a. 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.

Zhang, Y., P. Liu, X.-H. Liu, B. Pun, C. Seigneur, M.Z. Jacobson, and W.-X. Wang. 2010b. Fine Scale Modeling of Wintertime Aerosol Mass, Number, and Size Distributions in Central California. *J. Geophys. Res.* 115:D15207. doi:10.1029/2009JD012950.

Zhang, Y., X.-Y. Wen, and C. J. Jang. 2010c. Simulating Chemistry–Aerosol–Cloud–Radiation–Climate Feedbacks over the Continental U.S. using the Online-Coupled Weather Research Forecasting Model with Chemistry (WRF/Chem). *Atmos. Environ.* 44 (29):3568-3582. doi:10.1016/j.atmosenv.2010.05.056.

Zhang, Y., C. Seigneur, M. Bocquet, V. Mallet, and A. Baklanov. 2012a. Real-time air quality forecasting, Part I: History, Techniques, and Current Status. *Atmos. Environ.* 60:632-655. doi:10.1016/j.atmosenv.2012.06.031.

Zhang, Y., M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov. 2012b. Real-time air quality forecasting, part II: State of the science, current research needs and future prospects. *Atmos. Environ.* 60:656-676. doi:10106/j.atmosenv.2012.02.041.

Zhang, Y., Y. Chen, G. Sarwar, and K. Schere. 2012c. 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.

Zhang, Y., P. Karamchandani, T. Glotfelty, D. G. Streets, G. Grell, A. Nenes, F. Yu, and R. Bennartz. 2012d. Development and initial application of the global-through-urban weather research and forecasting model with chemistry (GU-WRF/Chem). *J. Geophys. Res.* 117:D20206. doi:10.1029/2012JD017966.

Zhang, Y., X. Zhang, K. Wang, J. He, L. R. Leung, J.-W. Fan, and A. Nenes. 2015. Incorporating An Advanced Aerosol Activation Parameterization into WRF-CAM5: Model Evaluation and Parameterization Intercomparison. *J. Geophys. Res.* doi: 10.1002/2014JD023051.

Ziemke, J. R., S. Chandra, B. N. Duncan, L. Froidevaux, P. K. Bhartia, P. F. Levelt, and J. W. Waters. 2006. Tropospheric ozone determined from Aura OMI and MLS: Evaluation of measurements and comparison with the Global Modeling Initiative's Chemical Transport Model. *J. Geophys. Res*. 111:D19303. doi:10.1029/2006JD007089.

| Variable | Network | 3 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2012 | | | | | 2013 | | | | | 2014[1] | | | | |
| | | Mean Obs[2] | Mean Sim | Corr | MB | MAGE | Mean Obs | Mean Sim | Corr | MB | MAGE | Mean Obs | Mean Sim | Corr | MB | MAGE |
| T2 (°C) | CASTNET | 22.2 | 22.8 | 0.6 | 0.5 | 4.2 | 21.2 | 23.3 | 0.7 | 2.1 | 4.0 | 21.2 | 23.3 | 0.7 | 2.1 | 4.0 |
| | NCDC | 23.7 | 24.3 | 0.6 | 0.6 | 4.0 | 23.0 | 24.8 | 0.7 | 1.8 | 3.7 | 22.9 | 24.7 | 0.7 | 1.7 | 3.9 |
| | SEARCH | 24.9 | 25.8 | 0.4 | 0.9 | 4.0 | 24.1 | 26.7 | 0.5 | 2.6 | 3.9 | 23.5 | 25.6 | 0.3 | 2.1 | 4.7 |
| RH2 (%) | CASTNET | 76.3 | 66.5 | 0.3 | -9.8 | 17.5 | 78.2 | 61.8 | 0.3 | -16.4 | 20.3 | 74.3 | 60.5 | 0.4 | -13.8 | 18.6 |
| | NCDC | 74.4 | 67.5 | 0.3 | -6.8 | 17.8 | 78.5 | 64.0 | 0.3 | -14.4 | 19.6 | 74.1 | 60.6 | 0.4 | -13.5 | 20.0 |
| | SEARCH | 74.0 | 63.5 | 0.3 | -10.4 | 18.5 | 80.1 | 59.6 | 0.2 | -20.5 | 23.2 | 74.5 | 54.7 | 0.3 | -19.7 | 23.1 |
| WS10 ($m\ s^{-1}$) | CASTNET | 2.0 | 3.7 | 0.4 | 1.7 | 1.9 | 1.8 | 3.7 | 0.2 | 1.8 | 2.1 | 2.5 | 4.1 | 0.4 | 1.6 | 2.1 |
| | NCDC | 3.6 | 3.8 | 0.2 | 0.2 | 1.8 | 3.6 | 3.9 | 0.2 | 0.4 | 1.8 | 3.7 | 3.9 | 0.2 | 0.2 | 1.8 |
| | SEARCH | 2.2 | 3.1 | 0.3 | 0.9 | 1.3 | 2.1 | 2.7 | 0.3 | 0.6 | 1.2 | 2.2 | 2.5 | 0.2 | 0.3 | 1.2 |
| WDR10 (°) | CASTNET | 201.8 | 224.8 | 0.7 | 23.0 | 79.1 | 207.2 | 223.5 | 0.7 | 16.3 | 86.2 | 195.6 | 224.7 | 0.6 | 29.0 | 85.7 |
| | NCDC | 186.5 | 232.1 | 0.7 | 45.6 | 85.2 | 187.3 | 235.2 | 0.7 | 47.9 | 85.1 | 199.4 | 241.9 | 0.6 | 42.4 | 86.1 |
| | SEARCH | 200.7 | 219.5 | 0.6 | 18.9 | 83.0 | 206.0 | 230.5 | 0.4 | 24.4 | 94.3 | 224.4 | 226.1 | 0.3 | 1.7 | 76.2 |
| Precip ($mm\ hr^{-1}$) | CASTNET | 0.2 | 0.3 | 0.0 | 0.1 | 0.4 | 0.2 | 0.3 | 0.0 | 0.1 | 0.5 | 0.2 | 0.3 | 0.0 | 0.1 | 0.4 |
| | NCDC | 3.2 | 4.7 | 0.0 | 1.5 | 5.3 | 2.8 | 4.1 | 0.0 | 1.4 | 4.6 | 3.2 | 4.3 | 0.1 | 1.1 | 4.9 |
| | GPCP | 0.2 | 0.3 | 0.4 | 0.1 | 0.1 | 0.2 | 0.3 | 0.4 | 0.1 | 0.1 | 0.2 | 0.2 | 0.1 | 0.0 | 0.1 |
| SWDOWN ($W\ m^{-2}$) | CERES | 245.2 | 279.7 | 0.6 | 34.4 | 34.5 | 239.7 | 281.9 | 0.1 | 42.2 | 42.2 | 244.2 | 311.5 | 0.4 | 67.3 | 67.3 |
| LWDOWN ($W\ m^{-2}$) | CERES | 399.9 | 391.0 | 1.0 | -8.9 | 11.9 | 397.8 | 398.5 | 1.0 | 0.7 | 4.4 | 398.6 | 390.5 | 1.0 | -8.1 | 8.1 |
| CF ($W\ m^{-2}$) | MODIS | 57.8 | 53.0 | 0.9 | -4.8 | 6.4 | 64.4 | 56.2 | 0.8 | -8.2 | 9.1 | 66.6 | 57.0 | 0.6 | -9.6 | 10.8 |
| COT | MODIS | 14.2 | 5.4 | 0.3 | -8.8 | 8.8 | 14.9 | 5.9 | -0.1 | -9.0 | 9.0 | 13.5 | 4.6 | 0.1 | -8.9 | 8.9 |

**Table 1b. Discrete statistics of meteorological variables for winter seasons.**

| Variable | Network | Winter Season (December-February) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2012-2013 | | | | | 2013-2014 | | | | | 2014-2015 | | | | |
| | | Mean Obs[2] | Mean Sim | Corr | MB | MAGE | Mean Obs | Mean Sim | Corr | MB | MAGE | Mean Obs | Mean Sim | Corr | MB | MAGE |
| T2 (°C) | CASTNET | 5.7 | 6.7 | 0.9 | 1.0 | 3.5 | 3.3 | 4.3 | 0.9 | 0.9 | 3.6 | 2.8 | 3.5 | 0.9 | 0.7 | 3.7 |
| | NCDC | 8.1 | 9.3 | 0.8 | 1.2 | 3.7 | 6.2 | 7.4 | 0.9 | 1.2 | 3.7 | 5.6 | 6.4 | 0.8 | 0.8 | 3.7 |
| | SEARCH | 10.1 | 5.2 | 0.4 | -5.0 | 8.7 | 7.9 | 5.0 | 0.1 | -3.0 | 7.6 | 10.0 | 11.1 | 0.6 | 1.1 | 4.0 |
| RH2 (%) | CASTNET | 75.0 | 69.0 | 0.4 | -6.0 | 15.9 | 72.0 | 70.1 | 0.4 | -1.9 | 15.9 | 74.2 | 70.2 | 0.4 | -4.1 | 15.1 |
| | NCDC | 76.3 | 70.4 | 0.4 | -6.0 | 16.0 | 74.1 | 69.9 | 0.4 | -4.2 | 16.0 | 74.0 | 69.7 | 0.5 | -4.3 | 15.1 |
| | SEARCH | 73.3 | 59.5 | 0.4 | -13.9 | 21.5 | 73.7 | 61.6 | 0.4 | -12.1 | 21.4 | 79.2 | 72.9 | 0.2 | -6.3 | 16.2 |
| WS10 ($m\ s^{-1}$) | CASTNET | 2.4 | 5.0 | 0.3 | 2.5 | 2.8 | 2.9 | 4.9 | 0.2 | 1.9 | 2.5 | 2.9 | 4.5 | 0.2 | 1.6 | 2.4 |
| | NCDC | 4.3 | 5.1 | 0.3 | 0.8 | 2.4 | 4.2 | 4.8 | 0.2 | 0.6 | 2.3 | 4.2 | 4.4 | 0.2 | 0.2 | 2.2 |
| | SEARCH | 2.5 | 4.1 | 0.6 | 1.5 | 1.7 | 2.4 | 3.9 | 0.5 | 1.4 | 1.7 | 2.3 | 2.5 | 0.3 | 0.2 | 1.1 |
| WDR10 (°) | CASTNET | 201.2 | 217.1 | 0.5 | 15.9 | 91.5 | 208.4 | 237.2 | 0.7 | 28.8 | 89.3 | 212.3 | 243.1 | 0.6 | 30.8 | 99.8 |
| | NCDC | 206.8 | 245.6 | 0.6 | 38.7 | 96.9 | 207.8 | 253.3 | 0.7 | 45.5 | 92.8 | 209.7 | 256.4 | 0.7 | 46.7 | 97.5 |
| | SEARCH | 212.7 | 247.9 | 0.8 | 35.2 | 94.1 | 206.7 | 228.9 | 0.8 | 22.2 | 97.6 | 212.2 | 258.8 | 0.9 | 46.6 | 86.7 |
| Precip ($mm\ hr^{-1}$) | CASTNET | 0.1 | 0.2 | 0.1 | 0.1 | 0.3 | 0.1 | 0.2 | 0.2 | 0.1 | 0.2 | 0.1 | 0.1 | 0.3 | 0.1 | 0.2 |
| | NCDC | 1.9 | 2.7 | 0.1 | 0.8 | 2.7 | 1.9 | 2.5 | 0.1 | 0.6 | 2.4 | 1.8 | 2.2 | 0.1 | 0.4 | 2.2 |
| | GPCP | 0.1 | 0.2 | 0.7 | 0.1 | 0.1 | 0.1 | 0.2 | 0.4 | 0.1 | 0.1 | 0.1 | 0.2 | 0.5 | 0.0 | 0.1 |
| SWDOWN ($W\ m^{-2}$) | CERES | 118.8 | 142.4 | 0.9 | 23.5 | 23.7 | 119.5 | 141.7 | 0.9 | 22.2 | 22.2 | 118.4 | 146.7 | 0.9 | 28.4 | 28.4 |
| LWDOWN ($W\ m^{-2}$) | CERES | 323.4 | 320.3 | 1.0 | -3.2 | 4.1 | 314.2 | 318.3 | 1.0 | 4.1 | 5.5 | 314.3 | 312.1 | 1.0 | -2.1 | 3.8 |
| CF ($W\ m^{-2}$) | MODIS | 64.4 | 50.0 | 0.7 | -14.4 | 14.4 | 64.4 | 52.8 | 0.5 | -11.6 | 11.7 | 65.8 | 50.0 | 0.8 | -15.7 | 15.7 |
| COT | MODIS | 17.8 | 5.4 | 0.2 | -12.3 | 12.3 | 18.8 | 5.6 | 0.2 | -13.2 | 13.2 | 18.2 | 4.9 | 0.1 | -13.3 | 13.3 |

[1] Data pairs only include simulated and observed data during May, June, and July in 2014 because of loss of simulated data in August and September, 2014 due to failure of backup external hard drives containing such data.
[2] Mean Obs: Mean observed data; Mean Sim: Mean simulated data; Corr: Correlation coefficient; MB: Mean bias; MAGE: Mean Absolute Gross Error; N/A: Data not available.

| Variable | Network | [3] | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2012 | | | | | 2013 | | | | | 2014[1] | | | | |
| | | Mean Obs[2] | Mean Sim | Corr | NMB (%) | NME (%) | Mean Obs | Mean Sim | Corr | NMB (%) | NME (%) | Mean Obs | Mean Sim | Corr | NMB (%) | NME (%) |
| CO (ppb) | SEARCH | 161.6 | 267.1 | 0.4 | 65.3 | 97.1 | 177.9 | 264.2 | 0.3 | 48.5 | 75.5 | 178.5 | 322.7 | 0.4 | 80.8 | 99.8 |
| $SO_2$ (ppb) | SEARCH | 1.0 | 1.9 | 0.0 | 99.0 | 195.0 | 0.5 | 1.9 | 0.1 | 308.0 | 362.0 | 0.3 | 2.8 | 0.3 | 725.0 | 736.0 |
| NO (ppb) | SEARCH | 1.9 | 1.4 | 0.1 | -26.0 | 135.0 | 1.2 | 1.6 | 0.2 | 36.0 | 176.0 | 1.5 | 4.9 | 0.1 | 222.0 | 339.0 |
| $NO_2$ (ppb) | SEARCH | 4.6 | 6.8 | 0.4 | 49.0 | 119.0 | 4.1 | 6.4 | 0.4 | 56.0 | 118.0 | 5.1 | 8.0 | 0.5 | 56.0 | 120.0 |
| $HNO_3$ (ppb) | SEARCH | 0.3 | 0.5 | 0.1 | 50.5 | 124.3 | 0.3 | 0.4 | 0.1 | 57.0 | 119.0 | 0.3 | 0.4 | 0.1 | 23.0 | 95.0 |
| Max 1-hr $O_3$ (ppb) | AIRNow | 52.5 | 52.4 | 0.5 | 0.0 | 22.0 | 45.6 | 52.4 | 0.5 | 15.0 | 26.0 | 49.1 | 49.1 | 0.4 | 0.0 | 23.0 |
| | AQS | 52.4 | 52.1 | 0.6 | -1.0 | 22.0 | 45.6 | 51.1 | 0.5 | 12.0 | 26.0 | N/A | N/A | N/A | N/A | N/A |
| | CASTNET | 51.3 | 49.8 | 0.5 | -3.0 | 22.0 | 46.3 | 52.0 | 0.4 | 12.0 | 25.0 | 50.5 | 51.0 | 0.4 | 1.0 | 19.0 |
| | SEARCH | 53.3 | 55.8 | 0.6 | 5.0 | 21.0 | 46.3 | 54.4 | 0.5 | 17.0 | 27.0 | 48.3 | 50.2 | 0.5 | 4.0 | 24.0 |
| Max 8-hr $O_3$ (ppb) | AIRNow | 47.4 | 47.5 | 0.5 | 0.0 | 22.0 | 41.0 | 47.9 | 0.4 | 17.0 | 27.0 | 44.5 | 45.0 | 0.4 | 1.0 | 23.0 |
| | AQS | 46.8 | 46.5 | 0.5 | -1.0 | 22.0 | 40.3 | 46.4 | 0.5 | 15.0 | 27.0 | N/A | N/A | N/A | N/A | N/A |
| | CASTNET | 45.8 | 45.1 | 0.5 | -2.0 | 23.0 | 37.1 | 43.6 | 0.3 | 18.0 | 28.0 | 40.3 | 41.7 | 0.3 | 4.0 | 20.0 |
| | SEARCH | 47.3 | 49.7 | 0.6 | 5.0 | 21.0 | 40.7 | 49.5 | 0.6 | 22.0 | 29.0 | 43.0 | 45.7 | 0.5 | 6.0 | 26.0 |
| 24-hr Avg $PM_{2.5}$ ($\mu g\ m^{-3}$) | AIRNow | 10.8 | 11.0 | 0.3 | 2.0 | 38.0 | 9.8 | 11.2 | 0.4 | 15.0 | 40.0 | 10.2 | 9.7 | 0.3 | -4.0 | 36.0 |
| | IMPROVE | 7.7 | 9.6 | 0.3 | 25.0 | 44.0 | 7.6 | 9.5 | 0.4 | 25.0 | 44.0 | 8.0 | 8.6 | 0.4 | 8.0 | 33.0 |
| | STN | 11.1 | 12.0 | 0.3 | 9.0 | 38.0 | 10.4 | 11.7 | 0.4 | 12.0 | 39.0 | N/A | N/A | N/A | N/A | N/A |
| | SEARCH | 9.8 | 13.6 | 0.2 | 39.0 | 66.0 | 9.1 | 13.9 | 0.3 | 53.0 | 74.0 | N/A | N/A | N/A | N/A | N/A |
| $NH_4^+$ ($\mu g\ m^{-3}$) | CASTNET | 0.8 | 0.8 | 0.6 | -7.0 | 29.0 | 0.8 | 0.8 | 0.8 | -4.0 | 25.0 | 0.7 | 0.6 | 0.5 | -12.0 | 29.0 |
| | IMPROVE | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| | STN | 0.5 | 0.9 | 0.4 | 64.0 | 88.0 | 0.5 | 0.8 | 0.4 | 58.0 | 87.0 | N/A | N/A | N/A | N/A | N/A |
| | SEARCH | 0.8 | 0.9 | 0.2 | 6.0 | 62.0 | 0.7 | 0.8 | 0.3 | 14.0 | 65.0 | N/A | N/A | N/A | N/A | N/A |
| $SO_4^{2-}$ ($\mu g\ m^{-3}$) | CASTNET | 2.6 | 2.5 | 0.4 | -3.0 | 29.0 | 2.4 | 2.6 | 0.5 | 6.0 | 27.0 | 2.3 | 2.1 | 0.2 | -9.0 | 25.0 |
| | IMPROVE | 2.1 | 2.4 | 0.3 | 18.0 | 47.0 | 2.0 | 2.4 | 0.4 | 16.0 | 49.0 | 1.9 | 2.1 | 0.4 | 10.0 | 40.0 |
| | STN | 2.2 | 2.6 | 0.3 | 16.0 | 46.0 | 2.1 | 2.6 | 0.4 | 20.0 | 51.0 | N/A | N/A | N/A | N/A | N/A |
| | SEARCH | 2.3 | 2.9 | 0.2 | 29.0 | 66.0 | 2.1 | 3.0 | 0.3 | 41.0 | 72.0 | N/A | N/A | N/A | N/A | N/A |
| $NO_3^-$ ($\mu g\ m^{-3}$) | CASTNET | 0.4 | 0.3 | 0.5 | -18.0 | 62.0 | 0.3 | 0.2 | 0.6 | -32.0 | 60.0 | 0.4 | 0.2 | 0.6 | -48.0 | 67.0 |
| | IMPROVE | 0.2 | 0.3 | 0.3 | 54.0 | 130.0 | 0.2 | 0.3 | 0.3 | 47.0 | 136.0 | 0.2 | 0.3 | 0.3 | 25.0 | 135.0 |
| | STN | 0.4 | 0.4 | 0.3 | -8.0 | 89.0 | 0.4 | 0.3 | 0.3 | -24.0 | 83.0 | N/A | N/A | N/A | N/A | N/A |
| | SEARCH | 0.2 | 0.3 | 0.0 | 48.0 | 207.0 | 0.2 | 0.3 | 0.0 | 83.0 | 233.0 | N/A | N/A | N/A | N/A | N/A |
| EC ($\mu g\ m^{-3}$) | IMPROVE | 0.2 | 0.2 | 0.3 | 0.0 | 54.0 | 0.2 | 0.2 | 0.5 | 5.0 | 54.0 | 0.2 | 0.2 | 0.2 | 1.0 | 54.0 |
| | SEARCH | 1.5 | 0.5 | -0.1 | -67.0 | 88.0 | 0.9 | 0.5 | 0.0 | -40.0 | 82.0 | N/A | N/A | N/A | N/A | N/A |
| OC ($\mu g\ m^{-3}$) | IMPROVE | 1.3 | 1.9 | 0.3 | 50.0 | 73.0 | 1.2 | 1.3 | 0.1 | 15.0 | 60.0 | 1.2 | 1.2 | 0.1 | 1.0 | 56.0 |
| | SEARCH | 3.0 | 2.9 | 0.2 | -6.0 | 57.0 | 2.2 | 1.6 | 0.2 | -28.0 | 61.0 | N/A | N/A | N/A | N/A | N/A |
| TC ($\mu g\ m^{-3}$) | IMPROVE | 1.5 | 2.2 | 0.3 | 42.0 | 66.0 | 1.4 | 1.6 | 0.2 | 14.0 | 57.0 | 1.4 | 1.5 | 0.1 | 1.0 | 54.0 |
| | STN | 2.8 | 2.7 | 0.5 | -3.0 | 37.0 | 2.7 | 2.0 | 0.1 | -27.0 | 46.0 | N/A | N/A | N/A | N/A | N/A |
| | SEARCH | 2.3 | 3.0 | 0.3 | 30.0 | 89.0 | 3.0 | 2.1 | 0.3 | -29.0 | 53.0 | N/A | N/A | N/A | N/A | N/A |
| Column CO ($10^{18}$ molec. $cm^{-2}$) | MOPITT | 2.1 | 1.2 | 0.5 | -42.2 | 42.2 | 2.0 | 1.3 | 0.5 | -36.5 | 36.5 | 2.0 | 1.3 | 0.3 | -37.2 | 37.2 |
| Column $NO_2$ ($10^{15}$ molec. $cm^{-2}$) | OMI | 1.7 | 1.1 | 0.7 | -35.3 | 45.9 | 1.6 | 1.1 | 0.7 | -33.4 | 42.9 | 1.7 | 1.1 | 0.7 | -33.5 | 40.8 |
| Column $SO_2$ (DU) | OMI | 0.25 | 0.11 | 0.5 | -54.9 | 59.2 | 0.25 | 0.11 | 0.5 | -55.1 | 58.8 | 0.25 | 0.11 | 0.4 | -55.3 | 58.7 |
| Column HCHO ($10^{15}$ molec. $cm^{-2}$) | OMI | 8.6 | 9.8 | 0.8 | 13.1 | 31.8 | 7.6 | 10.7 | 0.8 | 39.9 | 52.4 | 8.2 | 9.2 | 0.7 | 13.2 | 35.3 |
| Column $O_3$ (DU) | OMI | 39.4 | 33.3 | 0.7 | -15.4 | 16.8 | 38.0 | 36.0 | 0.6 | -5.3 | 9.0 | 37.6 | 35.9 | 0.7 | -4.5 | 8.1 |
| AOD | MODIS | 0.2 | 0.2 | 0.0 | 14.4 | 23.6 | 0.1 | 0.2 | -0.4 | 47.6 | 48.4 | 0.1 | 0.2 | -0.2 | 37.2 | 43.8 |

| Variable | Network | 2012-2013 | | | | | 2013-2014 | | | | | 2014-2015 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean Obs[2] | Mean Sim | Corr | NMB (%) | NME (%) | Mean Obs | Mean Sim | Corr | NMB (%) | NME (%) | Mean Obs | Mean Sim | Corr | NMB (%) | NME (%) |
| **CO (ppb)** | **SEARCH** | 200.9 | 279.1 | 0.3 | 38.9 | 72.4 | 203.7 | 345.9 | 0.3 | 69.8 | 89.3 | 244.6 | 514.1 | 0.5 | 110.2 | 122.5 |
| **$SO_2$ (ppb)** | **SEARCH** | 0.8 | 1.9 | 0.1 | 123.3 | 203.1 | 0.8 | 2.6 | 0.1 | 242.2 | 298.3 | 0.5 | 3.6 | 0.3 | 642.6 | 657.9 |
| **NO (ppb)** | **SEARCH** | 4.5 | 2.8 | 0.3 | -37.0 | 112.8 | 5.4 | 4.9 | 0.3 | -8.6 | 129.0 | 8.6 | 25.9 | 0.4 | 199.8 | 263.8 |
| **$NO_2$ (ppb)** | **SEARCH** | 6.4 | 7.9 | 0.5 | 24.3 | 81.8 | 7.5 | 9.4 | 0.5 | 24.9 | 79.0 | 9.0 | 3.0 | 0.5 | -67.2 | 71.0 |
| **$HNO_3$ (ppb)** | **SEARCH** | 0.2 | 0.4 | 0.2 | 111.3 | 149.1 | 0.2 | 0.4 | 0.2 | 114.7 | 155.9 | 0.2 | 0.0 | 0.0 | -90.7 | 91.4 |
| **Max 1-hr $O_3$ (ppb)** | **AIRNow** | 38.3 | 33.8 | 0.5 | -11.6 | 19.5 | 36.8 | 33.8 | 0.4 | -8.1 | 17.5 | 36.7 | 30.1 | 0.2 | -18.1 | 23.4 |
| | **AQS** | 38.0 | 33.1 | 0.5 | -12.8 | 21.7 | 33.9 | 31.7 | 0.5 | -6.5 | 20.5 | N/A | N/A | N/A | N/A | N/A |
| | **CASTNET** | 38.4 | 33.9 | 0.6 | -11.7 | 17.8 | 38.6 | 33.7 | 0.5 | -12.8 | 16.8 | 38.0 | 32.5 | 0.5 | -14.3 | 19.2 |
| | **SEARCH** | 37.6 | 32.5 | 0.6 | -13.3 | 20.1 | 36.0 | 31.2 | 0.6 | -13.2 | 20.3 | 30.5 | 27.8 | 0.3 | -9.0 | 22.8 |
| **Max 8-hr $O_3$ (ppb)** | **AIRNow** | 35.6 | 30.8 | 0.2 | -13.5 | 22.6 | 33.8 | 30.8 | 0.2 | -9.1 | 19.6 | 35.0 | 28.8 | 0.0 | -17.7 | 24.6 |
| | **AQS** | 33.8 | 29.4 | 0.5 | -13.0 | 24.3 | 29.8 | 27.8 | 0.4 | -6.5 | 23.9 | N/A | N/A | N/A | N/A | N/A |
| | **CASTNET** | 32.5 | 29.7 | 0.5 | -8.7 | 18.2 | 32.6 | 29.2 | 0.6 | -10.2 | 16.7 | 29.6 | 28.6 | 0.6 | -3.2 | 16.5 |
| | **SEARCH** | 33.6 | 29.6 | 0.6 | -12.0 | 21.3 | 32.2 | 28.1 | 0.6 | -13.0 | 21.6 | 26.4 | 25.9 | 0.2 | -1.7 | 25.4 |
| **24-hr Avg $PM_{2.5}$ ($\mu g\ m^{-3}$)** | **AIRNow** | 9.3 | 10.0 | 0.4 | 8.3 | 42.6 | 8.9 | 9.5 | 0.3 | 6.7 | 47.2 | 9.2 | 9.2 | 0.3 | 0.8 | 44.0 |
| | **IMPROVE** | 5.6 | 8.9 | 0.4 | 57.4 | 72.2 | 5.2 | 8.4 | 0.4 | 59.3 | 83.2 | N/A | N/A | N/A | N/A | N/A |
| | **STN** | 9.7 | 10.2 | 0.5 | 4.9 | 37.7 | 10.2 | 11.1 | 0.7 | 8.3 | 45.9 | N/A | N/A | N/A | N/A | N/A |
| | **SEARCH** | 7.8 | 12.4 | 0.2 | 59.7 | 85.3 | 7.6 | 12.9 | 0.3 | 68.4 | 89.0 | N/A | N/A | N/A | N/A | N/A |
| **$NH_4^+$ ($\mu g\ m^{-3}$)** | **CASTNET** | 0.9 | 0.5 | 0.6 | -42.7 | 48.2 | 0.8 | 0.5 | 0.6 | -32.0 | 45.2 | 1.0 | 0.4 | 0.6 | -55.7 | 58.8 |
| | **IMPROVE** | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| | **STN** | 0.8 | 0.6 | 0.4 | -17.9 | 63.5 | 0.7 | 0.7 | 0.7 | -9.6 | 63.6 | N/A | N/A | N/A | N/A | N/A |
| | **SEARCH** | 0.8 | 0.6 | 0.2 | -24.9 | 62.8 | 0.7 | 0.6 | 0.2 | -19.1 | 61.4 | N/A | N/A | N/A | N/A | N/A |
| **$SO_4^{2-}$ ($\mu g\ m^{-3}$)** | **CASTNET** | 1.8 | 1.5 | 0.2 | -21.1 | 34.1 | 1.7 | 1.5 | -0.2 | -14.6 | 39.2 | 1.9 | 1.3 | 0.3 | -31.0 | 40.9 |
| | **IMPROVE** | 1.5 | 1.5 | 0.2 | -3.3 | 51.3 | 1.5 | 1.4 | 0.1 | -7.5 | 55.6 | N/A | N/A | N/A | N/A | N/A |
| | **STN** | 1.7 | 1.6 | 0.1 | -5.8 | 53.7 | 1.7 | 1.6 | 0.5 | -4.0 | 47.9 | N/A | N/A | N/A | N/A | N/A |
| | **SEARCH** | 1.5 | 1.8 | 0.1 | 24.7 | 68.8 | 1.5 | 1.7 | 0.1 | 16.3 | 62.4 | N/A | N/A | N/A | N/A | N/A |
| **$NO_3^-$ ($\mu g\ m^{-3}$)** | **CASTNET** | 1.5 | 0.6 | 0.5 | -59.8 | 66.8 | 1.1 | 0.6 | 0.6 | -40.8 | 58.9 | 1.5 | 0.2 | 0.6 | -90.2 | 90.2 |
| | **IMPROVE** | 0.9 | 0.6 | 0.2 | -33.4 | 86.4 | 0.8 | 0.6 | 0.3 | -23.7 | 91.9 | N/A | N/A | N/A | N/A | N/A |
| | **STN** | 1.5 | 0.7 | 0.4 | -54.3 | 73.9 | 1.6 | 1.1 | 0.5 | -32.4 | 71.8 | N/A | N/A | N/A | N/A | N/A |
| | **SEARCH** | 0.6 | 0.5 | 0.0 | -12.8 | 119.2 | 0.6 | 0.7 | 0.3 | 16.4 | 121.1 | N/A | N/A | N/A | N/A | N/A |
| **EC ($\mu g\ m^{-3}$)** | **IMPROVE** | 0.3 | 0.4 | 0.5 | 37.3 | 74.9 | 0.3 | 0.3 | 0.4 | 24.4 | 74.0 | N/A | N/A | N/A | N/A | N/A |
| | **SEARCH** | 1.0 | 0.6 | 0.1 | -37.5 | 78.1 | 1.0 | 0.7 | 0.1 | -29.5 | 87.3 | N/A | N/A | N/A | N/A | N/A |
| **OC ($\mu g\ m^{-3}$)** | **IMPROVE** | 1.1 | 2.0 | 0.4 | 80.8 | 102.6 | 1.0 | 1.9 | 0.3 | 88.7 | 118.3 | N/A | N/A | N/A | N/A | N/A |
| | **SEARCH** | 2.1 | 2.8 | 0.2 | 33.0 | 85.7 | 2.2 | 2.8 | 0.2 | 24.0 | 79.5 | N/A | N/A | N/A | N/A | N/A |
| **TC ($\mu g\ m^{-3}$)** | **IMPROVE** | 1.3 | 2.3 | 0.4 | 72.5 | 95.0 | 1.3 | 2.3 | 0.3 | 75.0 | 106.4 | N/A | N/A | N/A | N/A | N/A |
| | **STN** | 2.6 | 2.6 | 0.4 | 1.5 | 44.5 | 2.5 | 2.6 | 0.6 | 5.4 | 55.9 | N/A | N/A | N/A | N/A | N/A |
| | **SEARCH** | 2.2 | 3.2 | 0.3 | 41.2 | 87.2 | 2.8 | 3.4 | 0.2 | 20.5 | 73.6 | N/A | N/A | N/A | N/A | N/A |
| **Column CO ($10^{18}$ molec. $cm^{-2}$)** | **MOPITT** | 2.3 | 1.2 | 0.3 | -50.7 | 50.7 | 2.3 | 1.2 | 0.1 | -48.4 | 48.4 | 2.3 | 1.2 | 0.1 | -48.2 | 48.2 |
| **Column $NO_2$ ($10^{15}$ molec. $cm^{-2}$)** | **OMI** | 2.7 | 2.7 | 0.9 | 1.0 | 20.1 | 2.3 | 2.9 | 0.9 | 26.2 | 32.8 | 2.5 | 2.3 | 0.7 | -7.9 | 32.7 |
| **Column $SO_2$ (DU)** | **OMI** | 0.39 | 0.09 | 0.5 | -77.2 | 77.5 | 0.41 | 0.09 | 0.5 | -77.1 | 77.3 | 0.40 | 0.11 | 0.5 | -73.2 | 73.7 |
| **Column HCHO ($10^{15}$ molec. $cm^{-2}$)** | **OMI** | 5.4 | 2.6 | 0.0 | -51.5 | 51.6 | 6.3 | 2.6 | 0.1 | -59.0 | 59.0 | N/A | N/A | N/A | N/A | N/A |
| **Column $O_3$ (DU)** | **OMI** | 25.7 | 35.8 | -0.2 | 39.3 | 42.4 | 27.1 | 35.2 | -0.2 | 29.9 | 36.6 | 25.1 | 36.4 | -0.4 | 45.1 | 47.9 |
| **AOD** | **MODIS** | 0.1 | 0.1 | 0.8 | 59.4 | 62.2 | 0.1 | 0.1 | 0.7 | 95.7 | 95.8 | 0.1 | 0.1 | 0.7 | 75.0 | 76.5 |

[1] Data pairs only include simulated and observed data during May, June, and July in 2014 because of loss of simulated data in August and September, 2014.
[2] Mean Obs: Mean observed data; Mean Sim: Mean simulated data; Corr: Correlation coefficient; NMB: Normalized mean bias; NME: Normalized mean error; N/A: Data not available.

Table 3. Discrete evaluation of RT-AQF results for $O_3$ and $PM_{2.5}$ predictions

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | **3** | | | | |
| NE US | 8/5-29,2002 | 1.4 | 14.6 | 2.2 | 18.0 | MAQSIP-RT | KA05 |
| NE US | 8/5-29,2002 | 9.5 | 21.3 | 15.0 | 25.8 | MM5/Chem | KA05 |
| NE US | 8/5-29,2002 | 3.2 | 19.1 | 5.1 | 23.4 | Hysplit/CheM | KA05 |
| E US | 7/1-8/15,2004 | 4.3–8.5 | 14.8–16.9 | 7.0–16.4 | 25.3 | Eta/CMAQ | YU07 |
| SE US | 5/1-9/30,2009 | 4.5 | 16.8 | 9.5 | 26.7 | WRF/Chem-MADRID | MT11 |
| SE US | 5/1-9/30,2009-2011 | -3.0 – 4.6 | 13.4 – 17.0 | -5.5 – 9.6 | 19.9–26.7 | WRF/Chem-MADRID | YA14[b] |
| SE US | 5/1-9/30,2009-2011 | -3.0 – 7.3 | 11.6-17.0 | -5.5 – 15.5 | 17.6–27.4 | WRF/Chem-MADRID | YA14[c] |
| SE US | 12/1-02/28,2009-2012 | -4.6 – -2.2 | 8.0 – 9.7 | -11.9 – -6.0 | 16.1–19.0 | WRF/Chem-MADRID | YA14[b] |
| SE US | 12/1-02/28,2009-2012 | -5.6 – 3.7 | 7.6-10.6 | -13.8 – 10.8 | 15.0–20.3 | WRF/Chem-MADRID | YA14[c] |
| SE US | 5/1-9/30,2012-2014 | -0.1 – 6.9 | 14.0 – 15.1 | 0 – 15.0 | 22.0–26.0 | WRF/Chem-MADRID | This work[b] |
| SE US | 5/1-9/30,2012-2014 | -1.5 – 8.0 | 12.1–15.9 | -3.0 – 17.0 | 19.0–27.0 | WRF/Chem-MADRID | This work[c] |
| SE US | 12/1-02/28,2012-2015 | -6.6 – -3.0 | 9.7–11.1 | -18.1 – -8.1 | 17.5–23.4 | WRF/Chem-MADRID | This work[b] |
| SE US | 12/1-02/28,2012-2015 | -6.2 – -4.2 | 8.0–11.1 | -18.1 – -6.5 | 16.8–23.4 | WRF/Chem-MADRID | This work[c] |
| | | | **Maximum 8-hr average $O_3$** | | | | |
| NE US | 8/5-29,2002 | 8.3 | 18.2 | 15.1 | 25.4 | MAQSIP-RT | KA05 |
| NE US | 8/5-29,2002 | 2.8 | 13.0 | 5.0 | 18.6 | MM5/Chem | KA05 |
| NE US | 8/5-29,2002 | -1.2 | 15.8 | -2.1 | 22.5 | Hysplit/Chem | KA05 |
| NE US | 6/1-9/30,2004 | 10.2 | 15.7 | 22.8 | 28.1 | Eta/CMAQ | ED06 |
| E US | 7/1-8/15,2004 | 6.5–10.4 | 13.9–16.6 | 11.9–22.6 | 19.7–28.8 | Eta/CMAQ | YU07 |
| NY | 7/1-9/30,2004 | 6.5 | 12.8 | — | — | Eta/CMAQ | HO07 |
| NY | 1/1-3/31,2005 | 1.4 | 8.7 | — | — | Eta/CMAQ | HO07 |
| NY | 6/1-9/30,2005 | 4.7 | 13.0 | — | — | Eta/CMAQ | HO07 |
| NE US | 7/14-8/17,2004 | 3.4–14.3 | 11.6–20.9 | — | — | WRF/chem | MK07 |
| and | | 17.0 | 23.2 | — | — | CHRONOS | MK07 |
| SE CA | | 5.9 | 16.2 | — | — | AURAMS | MK07 |
| | | 26.4 | 31.0 | — | — | STEM-2K3 | MK07 |
| | | 13.4 | 17.9 | — | — | ET/CMAQ | MK07 |
| E US | 6/1-9/30,2005 | 10.9 | 16.3 | 22.4 | 27.1 | WRF-NMM/CMAQ | ED09 |
| E US | 6/1-9/30,2006 | 10.5 | 15.6 | 25.2 | 30.4 | WRF-NMM/CMAQ | ED09 |
| E US | 6/1-9/30,2007 | 7.9 | 14.5 | 16.5 | 24.1 | WRF-NMM/CMAQ | ED09 |
| SE US | 5/1-9/30, 2009 | 3.5 | 13.6 | 8.3 | 25.0 | WRF/Chem-MADRID | MT11 |
| SE US | 5/1-9/30,2009-2011 | -1.8 –3.6 | 11.7 – 13.7 | -3.7 –8.5 | 19.6–25.0 | WRF/Chem-MADRID | YA14[b] |
| SE US | 5/1-9/30,2009-2011 | -2.2 –6.1 | 10.5–13.9 | -4.5 –14.6 | 17.8–26.1 | WRF/Chem-MADRID | YA14[c] |
| SE US | 12/1-02/28,2009-2012 | -4.9 – 2.0 | 8.1 – 12.2 | -13.5 – -5.8 | 16.9–33.8 | WRF/Chem-MADRID | YA14[b] |
| SE US | 12/1-02/28,2009-2012 | -4.9 – 2.0 | 6.7–12.2 | -13.5 – -0.3 | 16.9–21.5 | WRF/Chem-MADRID | YA14[c] |
| SE US | 5/1-9/30,2012-2014 | -0.2 –6.9 | 13.2–14.2 | 0.0 –17.0 | 22.0–27.0 | WRF/Chem-MADRID | This work[b] |
| SE US | 5/1-9/30,2012-2014 | -0.8 –8.8 | 10.3–15.1 | -2.0 –22.0 | 21.0–29.0 | WRF/Chem-MADRID | This work[c] |
| SE US | 12/1-02/28,2012-2015 | -6.2 – -3.1 | 11.0–17.3 | -17.7 – -9.1 | 19.6–24.6 | WRF/Chem-MADRID | This work[b] |
| SE US | 12/1-02/28,2012-2015 | -6.2 – -0.5 | 6.1–17.3 | -17.7 – -1.7 | 18.2–25.4 | WRF/Chem-MADRID | This work[c] |
| | | | **24-hr average $PM_{2.5}$** | | | | |
| NY | 7/1-9/30,2004 | 5.4 | 13.2 | — | — | Eta/CMAQ | HO07 |
| NY | 1/1-3/31,2005 | 6.2 | 14.5 | — | — | Eta/CMAQ | HO07 |
| NY | 6/1-7/31,2005 | 4.4 | 13.6 | — | — | Eta/CMAQ | HO07 |
| PN | 8/1-11/30,2004 | 2.1–2.2 | — | 17–32 | 70–81 | MM5/CMAQ | CH08 |
| E US | 7/14-8/18,2004 | -3.2 | 8.8 | -21.0 | 41.2 | Eta/CMAQ | YU08 |
| E Texas | 8/31-10/12,2006 | -1.3 | 5.5 | — | — | 7-model ensemble [a] | DJ10 |
| NA | Summer 2008 | -2.08 | 12.8 | — | — | GEM-CHRONOS | MA09 |
| NA | Winter 2008 | 0.86 | 14.1 | — | — | GEM-CHRONOS | MA09 |
| NA | Summer 2009 | -0.70 | 12.9 | — | — | GEM-CHRONOS | MA09 |
| NA | Summer 2008 | 0.69 | 13.5 | — | — | GEM-MACH15 | MA09 |
| NA | Winter 2008 | -0.18 | 15.9 | — | — | GEM-MACH15 | MA09 |
| NA | Summer 2009 | 2.08 | 13.6 | — | — | GEM-MACH15 | MA09 |
| SE US | 5/1-9/30,2009 | -0.6 | 5.9 | -5.6 | 37.0 | WRF/Chem-MADRID | MT11 |
| SE US | 5/1-9/30,2009-2011 | -1.3 – -0.6 | 5.9 – 8.7 | -10.1 – -5.2 | 36.7 – 38.9 | WRF/Chem-MADRID | YA14[b] |
| SE US | 5/1-9/30,2009-2011 | -1.3 – 3.6 | 4.8 – 20.1 | -10.1 – 34.3 | 35.2 – 65.5 | WRF/Chem-MADRID | YA14[c] |
| SE US | 12/1-02/28,2009-2012 | -1.1 – 0.2 | 5.4 – 6.8 | -10.2 – 1.4 | 39.9 – 41.6 | WRF/Chem-MADRID | YA14[b] |
| SE US | 12/1-02/28,2009-2012 | -2.9 – 3.1 | 4.9 – 9.3 | -20.6 – 36.6 | 0.6 – 65.5 | WRF/Chem-MADRID | YA14[c] |
| SE US | 5/1-9/30,2012-2014 | -0.5 – 1.4 | 5.1–5.7 | -4.0 – 15.0 | 36.0 – 40.0 | WRF/Chem-MADRID | This work[b] |
| SE US | 5/1-9/30,2012-2014 | -0.5 – 4.8 | 3.8–10.8 | 2.0 – 53.0 | 33.0 – 74.0 | WRF/Chem-MADRID | This work[c] |
| SE US | 12/1-02/28,2012-2015 | 0.2 – 0.8 | 5.5–6.1 | 0.8 – 8.3 | 42.6 – 47.4 | WRF/Chem-MADRID | This work[b] |
| SE US | 12/1-02/28,2012-2015 | 0.1 – 5.2 | 4.9–10.5 | 4.9 – 68.4 | 37.3 – 89.0 | WRF/Chem-MADRID | This work[c] |

1. MB: Mean Bias; RMSE: Root Mean Square Error; NMB: Normalized Mean Bias; NME: Normalized Mean Error. SE US: Southeastern U.S.; E US: Eastern U.S., NE US: Northeastern U.S.; SE CA: southeastern Canada; PN: Pacific Northwest; NY: New York State; E Texas: eastern Texas; NA: North America. The unit for MB and RSME are ppb for $O_3$ and $\mu g\ m^{-3}$ for $PM_{2.5}$.
2. Superscript a: the 7 models include: WRF/Chem-2 (27-km), WRF/Chem-2 (12-km), CHRONOS, AURAMS, STEM-2K3, BAMS (15-km), and NMM/CMAQ; b: statistics based on evaluation against AirNow; c: statistics based on evaluation against all datasets.
3. MT11: Chuang et al. (2011); KA05: Kang et al. (2005); ED06: Eder et al. (2006); HO07: Hogrefe et al. (2007); MK07; McKeen et al. (2007); YU07: Yu et al. (2007); CH08: Chen et al. (2008); MA09: Makar et al., 2009; ED09: Eder et al. (2009); YU08: Yu et al. (2008); DJ10: Djalalova et al. (2010); YA14: Yahya et al. (2014).

Table 4. Categorical evaluation of RT-AQF results against AirNow for $O_3$ and $PM_{2.5}$ predictions.

| Area | Period | A (%) | CSI (%) | POD (%) | B | FAR (%) | Model | Reference |
|---|---|---|---|---|---|---|---|---|
| | | | | **Maximum 1-hr average $O_3$** | | | | |
| NE US | 8/5-29,2002 | 99.2 | 9.7 | 14.0 | 0.6 | 76 | MAQSIP-RT | KA05 |
| NE US | 8/5-29,2002 | 97.0 | 9.8 | 29.8 | 2.3 | 87.2 | MM5/Chem | KA05 |
| NE US | 8/5-29,2002 | 99.0 | 8.3 | 18.2 | 1.4 | 86.7 | Hysplit/CheM | KA05 |
| SE US | 5/1-9/30,2009 | 94.0 | 5.2 | 31.3 | 5.3 | 94.1 | WRF/Chem-MADRID | MT11 |
| SE US | 5/1-9/30,2009-2011 | 94.6-96 | 5.2-13.8 | 17-31.3 | 0.6-5.3 | 67-94.1 | WRF/Chem-MADRID | YA14 |
| SE US | 12/1-02/28,2009-2012 | 100 | 0 | 0 | 0 | 0 | WRF/Chem-MADRID | YA14 |
| SE US | 5/1-9/30,2012-2014 | 94.2-97.7 | 3.6-15.5 | 18.8-30.5 | 0.9-7.9 | 71.5-96.1 | WRF/Chem-MADRID | This work |
| SE US | 12/1-02/28,2012-2015 | 100 | 0 | 0 | 0 | 0 | WRF/Chem-MADRID | This work |
| | | | | **Maximum 8-hr average $O_3$** | | | | |
| NE US | 8/1-10,2001 | 80.0 | 34.0 | 49.0 | 1.1 | 13.0 | MM5/MAQSIP_RT | MC04 |
| NE US. | 8/5-29,2002 | 85.8 | 18.1 | 26.7 | 0.7 | 64.0 | MAQSIP-RT | KA05 |
| NE US | 8/5-29,2002 | 76.2 | 17.6 | 36.4 | 1.4 | 74.6 | MM5/Chem | KA05 |
| NE US | 8/5-29,2002 | 89.5 | 5.8 | 7.1 | 0.3 | 76.3 | Hysplit/Chem | KA05 |
| NE US | 6/1-9/30,2004 | 98.9 | 14.2 | 41.0 | 2.3 | 82.1 | Eta/CMAQ | ED06 |
| NY | 7/1-9/30,2004 | 84.0-95.2 | 31.4-53.2 | 46.5-84.8 | — | 32.9-55.2 | Eta/CMAQ | HO07 |
| | 1/1-3/31, 6/1-9/30,2005 | 96.1-99.8 | 0.0-29.0 | 0.0-58.3 | — | 36.7-82.5 | Eta/CMAQ | HO07 |
| NE US | 8/12,2005 | 91.6 | 23.4 | 31.3 | 0.7 | 51.6 | Eta/CMAQ | LE08 |
| E US | 8/12,2005 | 90.4 | 24.3 | 37.5 | 0.9 | 59.1 | Eta/CMAQ | LE08 |
| CONUS | 8/12,2005 | 87.4 | 26.0 | 54.2 | 1.6 | 66.7 | Eta/CMAQ | LE08 |
| SE US | 5/1-9/30, 2009 | 85.6 | 14.0 | 33.3 | 1.7 | 80.6 | WRF/Chem-MADRID | MT11 |
| CONUS | 6/1-8/31,2010 | 86-91 | 0.17-0.21 | 0.71-0.76 | | 0.77-0.82 | WRF-NMM/CMAQ | CH13 |
| CONUS | 01/01-12/31,2010 | 93-96 | 0.17-0.21 | 0.64-0.67 | | 0.76-0.81 | WRF-NMM/CMAQ | CH13 |
| SE US | 5/1-9/30,2009-2011 | 81.4-85.7 | 14-24.9 | 29.1-33.3 | 0.6-1.7 | 48.6-80.6 | WRF/Chem-MADRID | YA14 |
| SE US | 12/1-02/28,2009-2012 | 98.7-100 | 0 | 0 | 0 | N/A[a] | WRF/Chem-MADRID | YA14 |
| SE US | 5/1-9/30,2012-2014 | 80.2-85.3 | 9.9-25.3 | 26.6-46.7 | 0.8-4.2 | 54.9-88.9 | WRF/Chem-MADRID | This work |
| SE US | 12/1-02/28,2012-2015 | 98.7-99.2 | 0 | 0 | 0 | N/A[a] | WRF/Chem-MADRID | This work |
| | | | | **24-hr average $PM_{2.5}$** | | | | |
| NY | 7/1-9/30,2004 | 60.8-89.7 | 22.5-53.7 | 24.3-90.9 | — | 25.0-55.0 | Eta/CMAQ | HO07 |
| | 1/1-3/31, 6/1-7/31,2005 | 91.4-99.7 | 0-3.6 | 0-44.7 | — | N/A[a], 96.2-100 | Eta/CMAQ | HO07 |
| E Texas | 8/31-10/12,2006 | — | 0.0-8.0 | 0.0-14 | — | 80-100 | 7-model ensemble [b] | DJ10 |
| SE US | 5/1-9/30,2009 | 76.2 | 22.3 | 31.5 | 0.7 | 56.6 | WRF/Chem-MADRID | MT11 |
| SE US | 5/1-9/30,2009-2011 | 70.7-76.2 | 22.3-27.9 | 31.5-36 | 0.6-0.7 | 44.6-56.7 | WRF/Chem-MADRID | YA14 |
| SE US | 12/1-02/28,2009-2012 | 82.2-85.9 | 14.8-22.2 | 27.7-38.3 | 0.7-1.2 | 61.3-76.6 | WRF/Chem-MADRID | YA14 |
| SE US | 5/1-9/30,2012-2014 | 77.5-83.2 | 10.3-21.3 | 15.3-40.1 | 0.6-1.3 | 68.3-75.9 | WRF/Chem-MADRID | This work |
| SE US | 12/1-02/28,2012-2015 | 83.5-85.3 | 14.7-17.1 | 25.5-31.8 | 1.0-1.2 | 72.1-74.1 | WRF/Chem-MADRID | This work |

1. A: Accuracy; CSI: Critical Success index; POD: Probability Of Detection; B: Bias; FAR: False Alarm Ratio. SE US: Southeastern U.S.; NE US: Northeastern U.S.; E US: eastern U.S.; E Texas: eastern Texas; NY: New York State, CONUS: continental U.S.
2. Superscript a: An FAR of N/A indicates that no exceedances were predicted by the AQF model; b: the seven models include: WRF/chem-2 (27-km), WRF/chem-2 (12-km), CHRONOS, AURAMS, STEM-2K3, BAMS (15-km), and NMM/CMAQ;
3. MT11: Chuang et al. (2011); MC04: McHenry et al. (2004); KA05: Kang et al. (2005); ED06: Eder et al. (2006); HO07: Hogrefe et al. (2007); LE08: Lee et al. (2008); DJ10: Djalalova et al. (2010); CH13: Chai et al. (2013); YA14: Yahya et al. (2014).

**List of Figure Captions**

Figure 1. Discrete evaluation of the maximum 1-hr and 8-hr $O_3$ and 24-hr average $PM_{2.5}$ for (a) $O_3$ seasons and (b) winters during 2009-2015.

Figure 2. Spatial distributions of maximum 1-hr $O_3$ and 8-hr $O_3$ during the $O_3$ seasons and average 24-hr $PM_{2.5}$ concentrations during the $O_3$ and winter seasons during 2012-2015. The observations are symbolled as circles, they are taken from AIRNow, AIRS-AQS, CASTNET, and SEARCH for $O_3$ and from AIRNow, IMPROVE, STN, and SEARCH for $PM_{2.5}$.

Figure 3. Time series of the observed and forecasted maximum 1-hr $O_3$ and 8-hr $O_3$ concentrations for $O_3$ seasons during 2012-2015. The observations are taken from AIRNow.

Figure 4. Time series of the average 24-hr average $PM_{2.5}$ concentrations for (a) $O_3$ seasons and (b) winters during 2012-2015. The observations are taken from AIRNow.

Figure 5. Categorical evaluation of the maximum 1-hr and 8-hr $O_3$ and 24-hr average $PM_{2.5}$ for (a) $O_3$ seasons and (b) winters during 2009-2015.

Figure 6. Spatial distributions of satellite-derived and simulated column $NO_2$, TOR, and AOD during (a) the 2012 $O_3$ season (rows 1 and 2), and (b) the winter of 2012-2013 (rows 3 and 4).

Figure 7. Comparison of CO spatial distributions in Aug. 2012: (a) satellite observation from MOPPIT, (b) baseline simulation, and (c) sensitivity simulation.

Figure 8. Comparison of TOR spatial distributions in Dec. 2012: (a) satellite observation from OMI, (b) baseline simulation, and (c) sensitivity simulation.

Figure 9. Changes in observed and forecasted T2, Precip, and WS10 (relative to 2009) and SWDOWN, LWDOWN, and CF (relative to 2011) during 2010-2015.

Figure 10. Changes in observed and forecasted surface $O_3$ and $PM_{2.5}$ concentrations (relative to 2009) and column CO, $NO_2$, $SO_2$, TOR, AOD (relative to 2011) during 2010-2015.

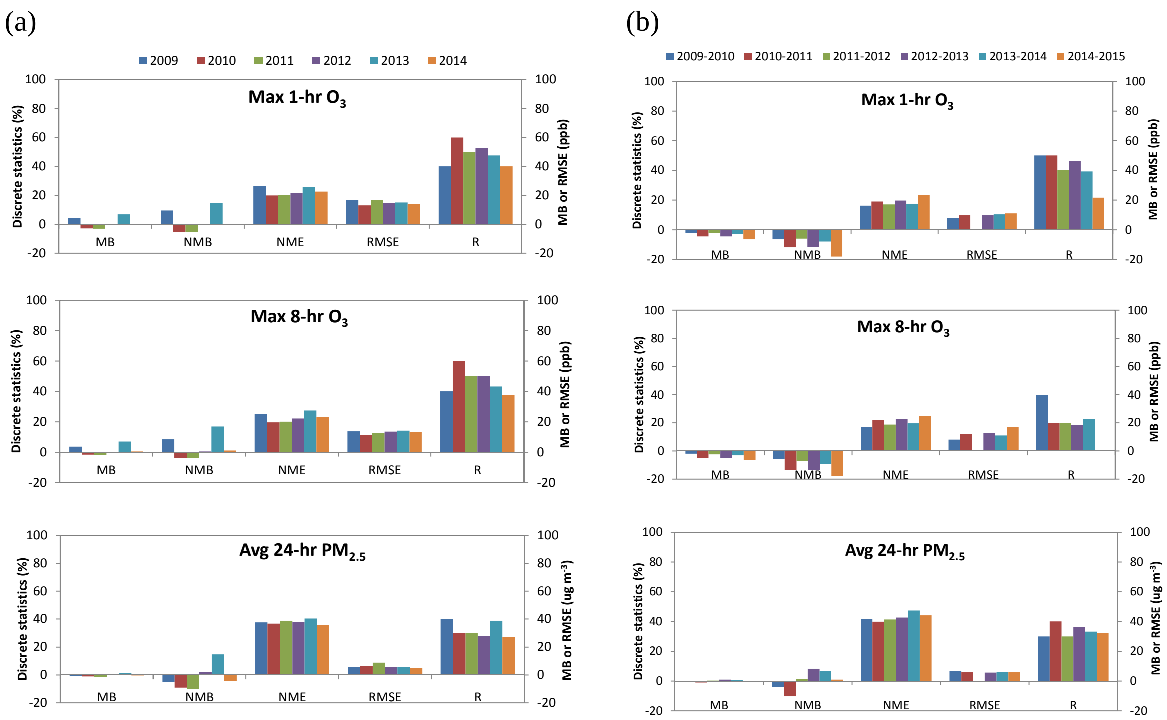


Figure 1. Discrete evaluation of the maximum 1-hr and 8-hr $O_3$ and 24-hr average $PM_{2.5}$ for (a) $O_3$ seasons and (b) winters during 2009-2015.

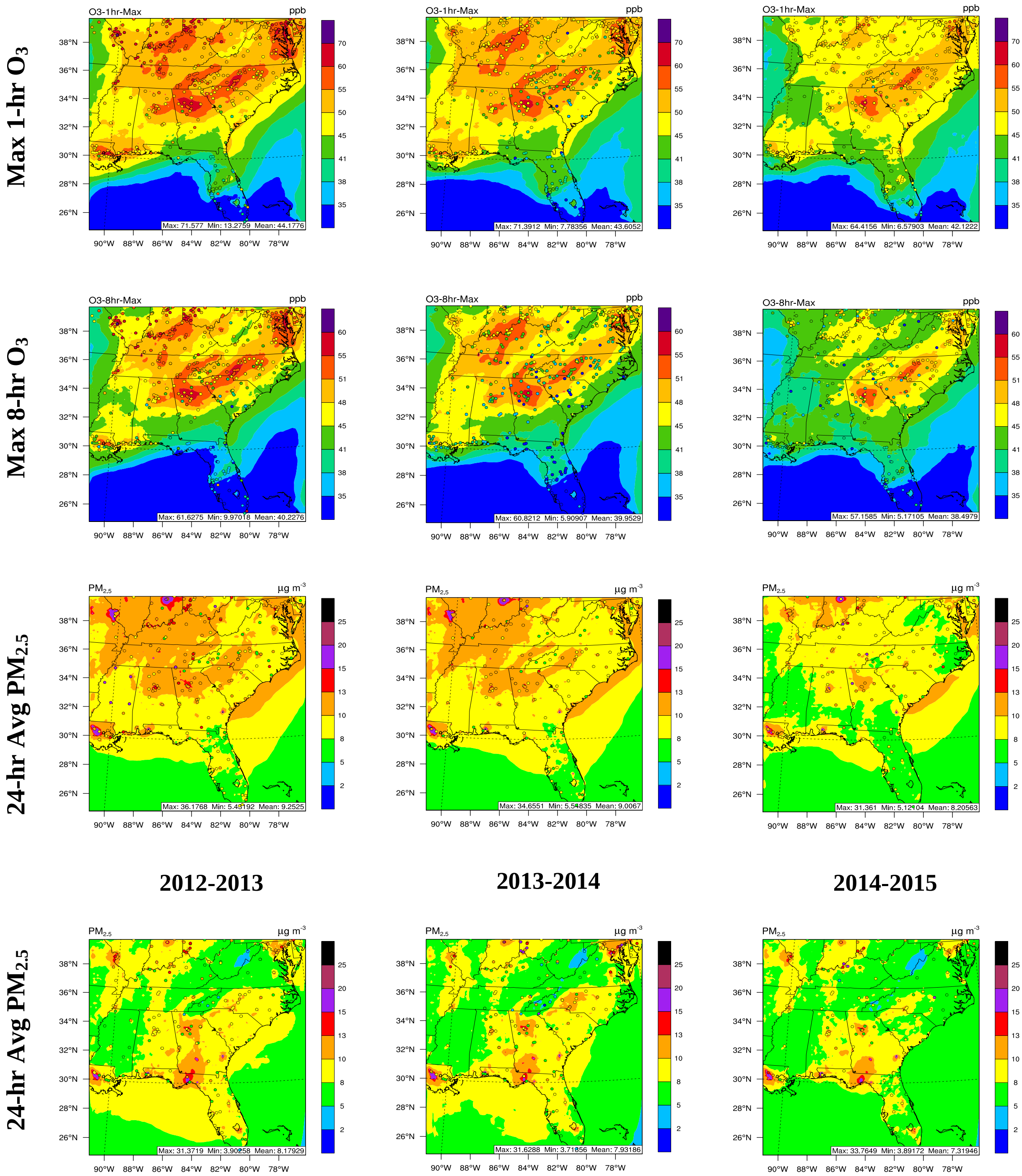


Figure 2. Spatial distributions of maximum 1-hr $O_3$ and 8-hr $O_3$ during the $O_3$ seasons and average 24-hr $PM_{2.5}$ concentrations during the $O_3$ and winter seasons during 2012-2015. The observations are symbolled as circles, they are taken from AIRNow, AIRS-AQS, CASTNET, and SEARCH for $O_3$ and from AIRNow, IMPROVE, STN, and SEARCH for $PM_{2.5}$.

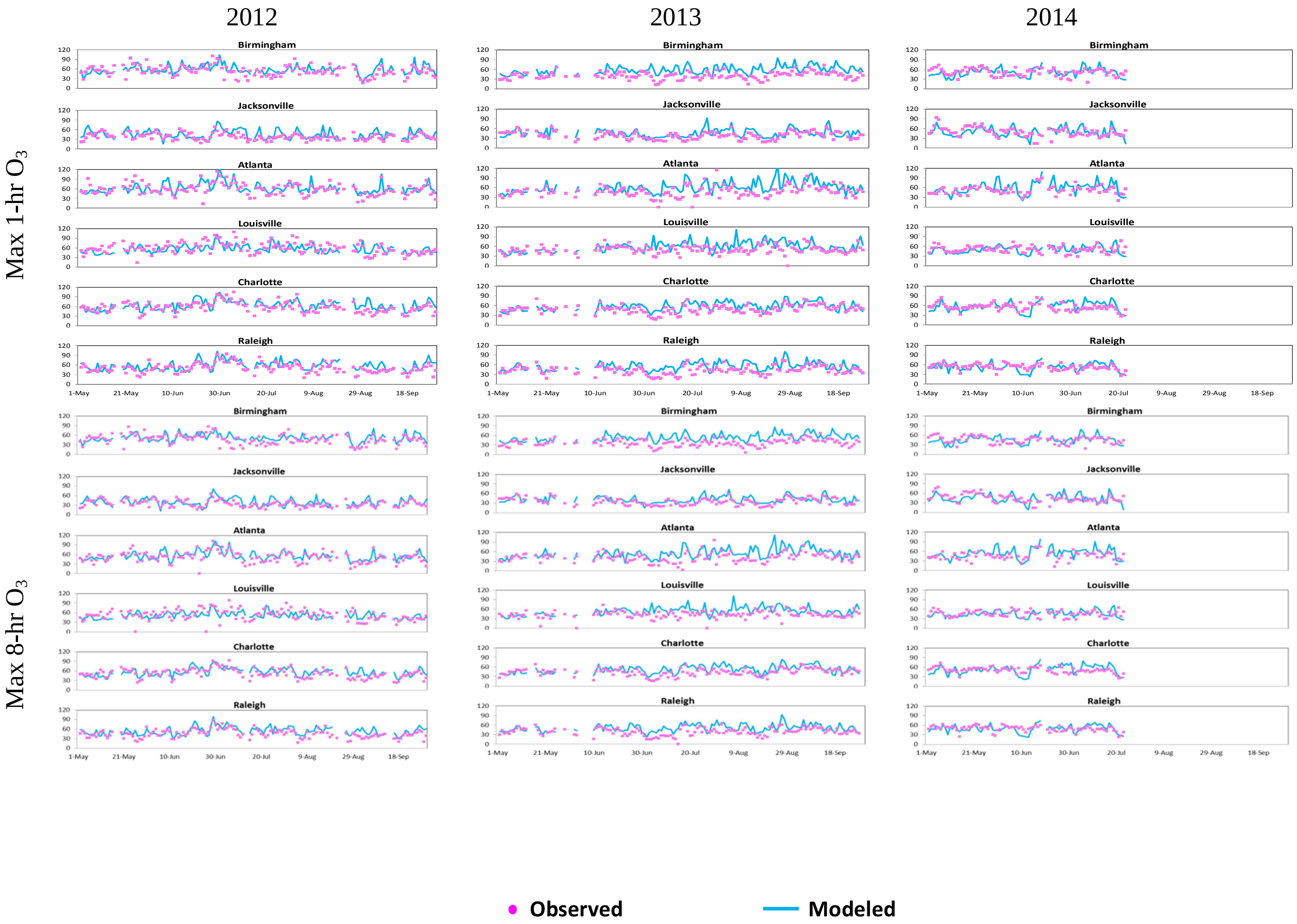


Figure 3. Time series of the observed and forecasted maximum 1-hr $O_3$ and 8-hr $O_3$ concentrations for $O_3$ seasons during 2012-2015. The observations are taken from AIRNow.


Figure 4. Time series of the average 24-hr average $PM_{2.5}$ concentrations for (a) $O_3$ seasons and (b) winters during 2012-2015. The observations are taken from AIRNow.

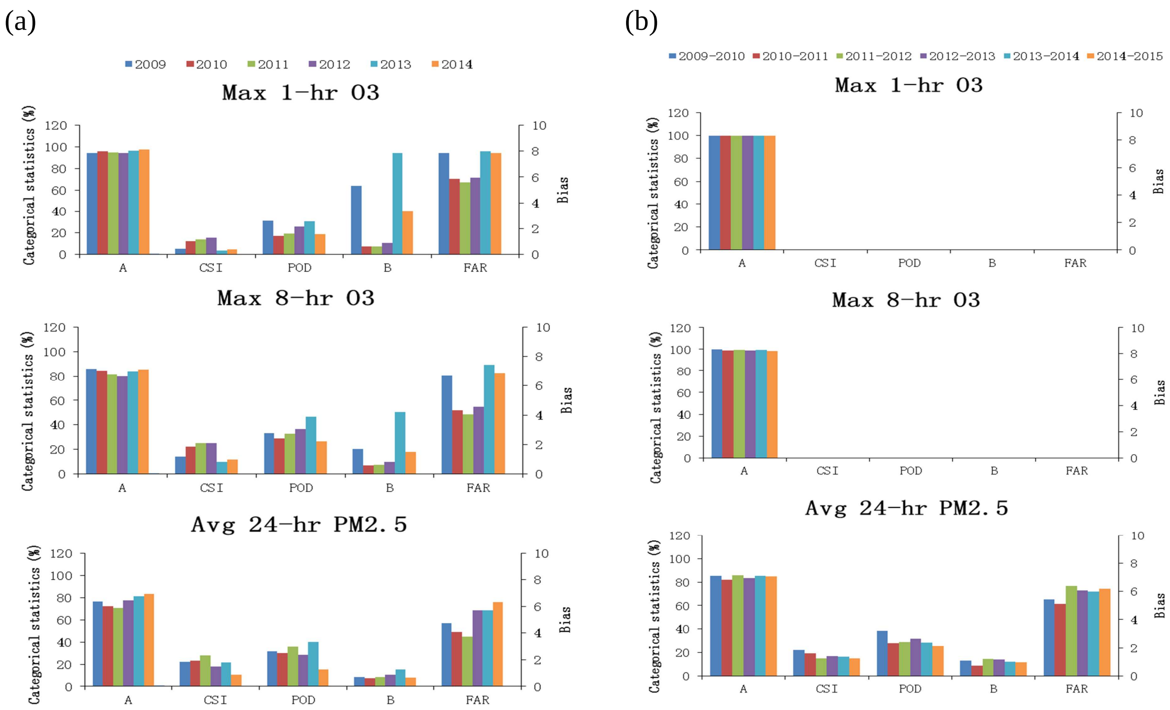


Figure 5. Categorical evaluation of the maximum 1-hr and 8-hr $O_3$ and 24-hr average $PM_{2.5}$ for (a) $O_3$ seasons and (b) winters during 2009-2015.

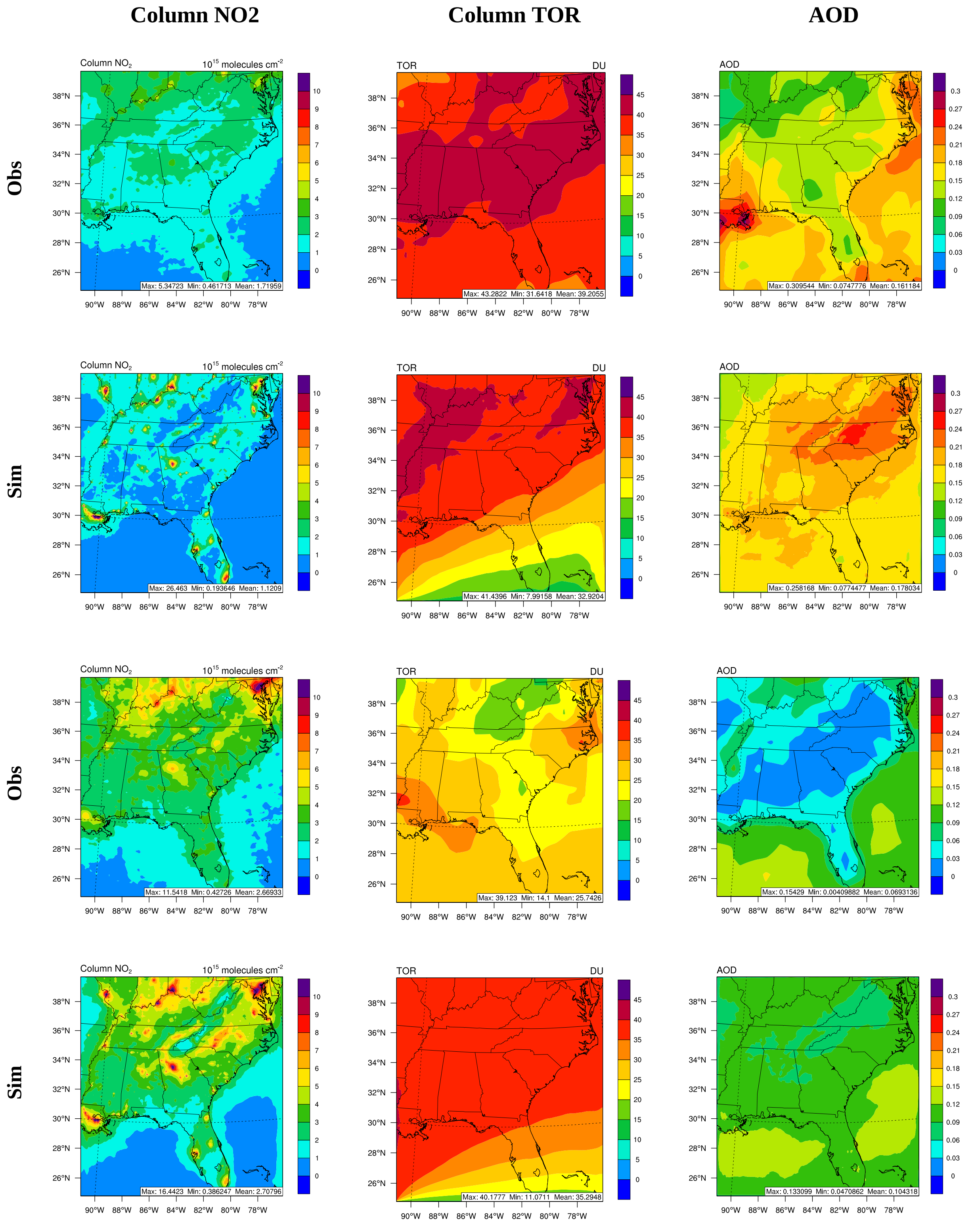


Figure 6. Spatial distributions of satellite-derived and simulated column $NO_2$, TOR, and AOD during (a) the 2012 $O_3$ season (rows 1 and 2), and (b) the winter of 2012-2013 (rows 3 and 4).

(a) Obs from MOPPIT (b) Sim_base (c) Sim_sen

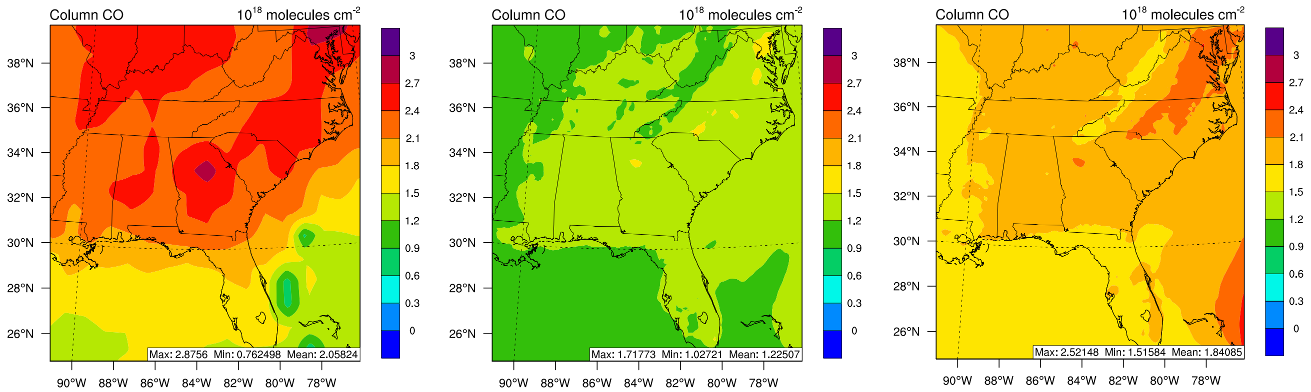


Figure 7. Comparison of CO spatial distributions in August. 2012: (a) satellite observation from OMI, (b) baseline simulation, and (c) sensitivity simulation.

(a) Obs from OMI (b) Sim_base (c) Sim_sen

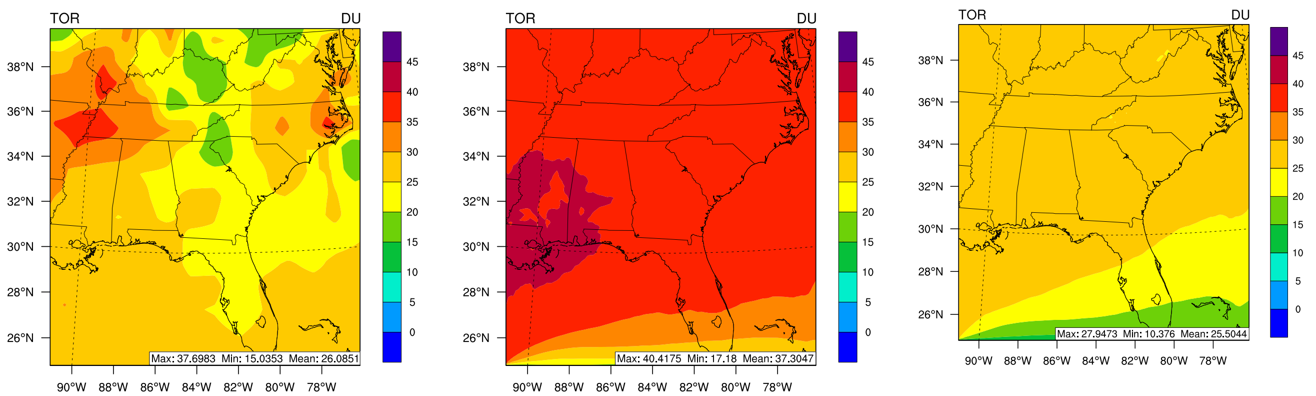


Figure 8. Comparison of TOR spatial distributions in Dec. 2012: (a) satellite observation from OMI, (b) baseline simulation, and (c) sensitivity simulation.

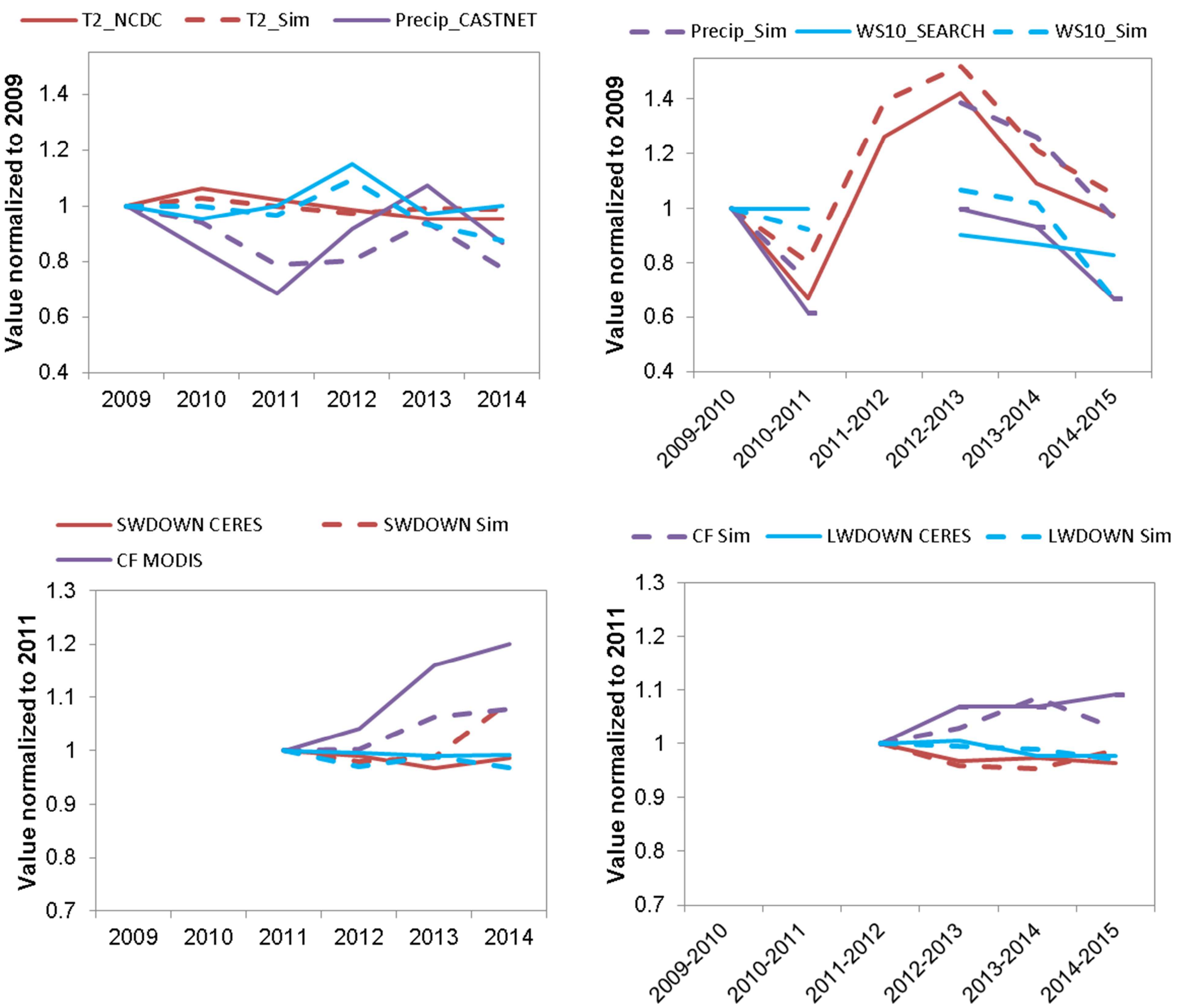


Figure 9. Changes in observed and forecasted T2, Precip, and WS10 (relative to 2009) and SWDOWN, LWDOWN, and CF (relative to 2011) during 2010-2015.

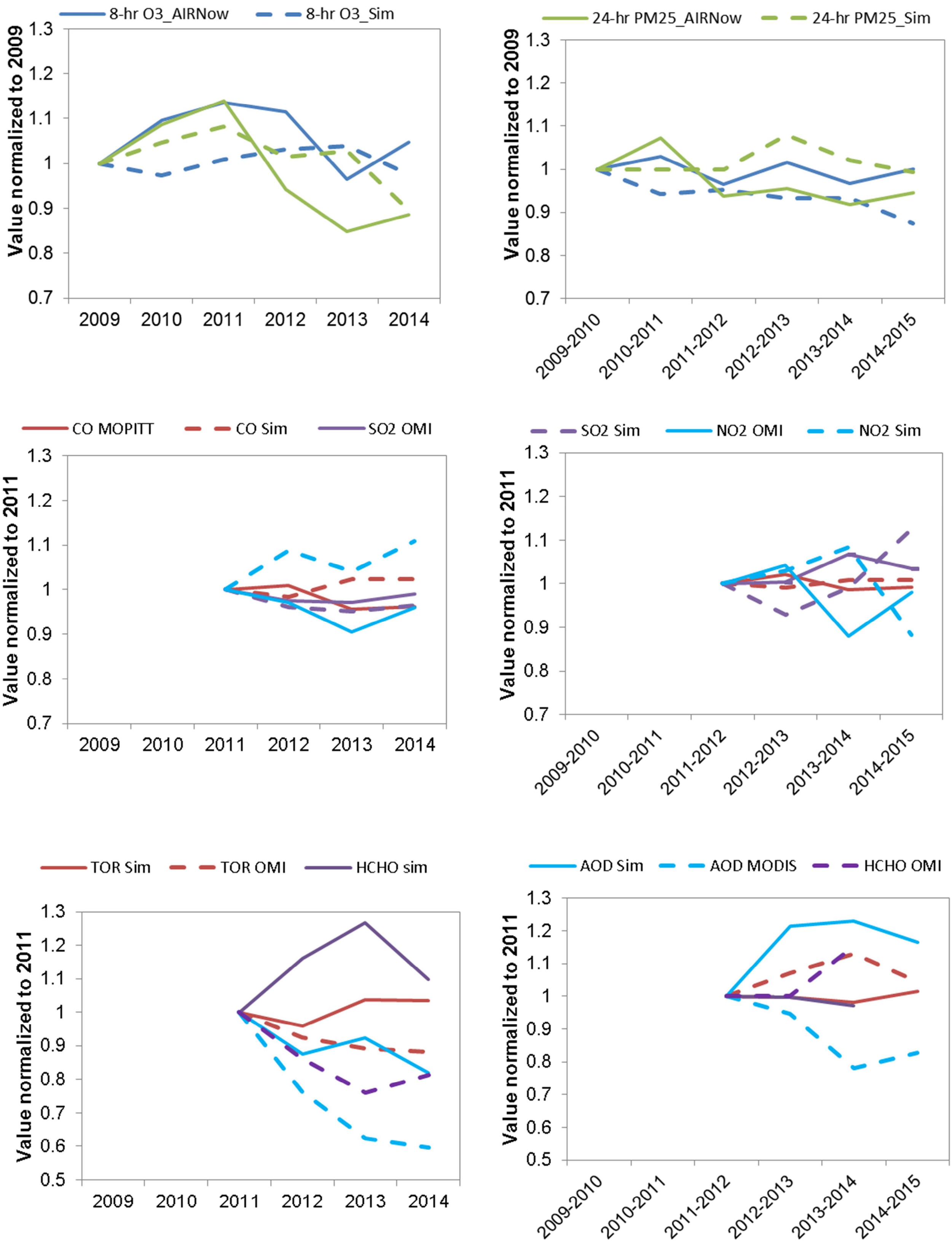


Figure 10. Changes in observed and forecasted surface $O_3$ and $PM_{2.5}$ concentrations (relative to 2009) and column CO, $NO_2$, $SO_2$, TOR, AOD (relative to 2011) during 2010-2015.

Highlights:

(1) A comprehensive evaluation of multi-year forecasts using surface and satellite data
(2) The model shows good skills for multi-year trends and inter-seasonal variability at surface
(3) Satellite-constrained boundary conditions can improve forecasts of column variables.