https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

Operational chemical weather forecasting with the ECCC online Regional Air Quality Deterministic Prediction System version 023 (RAQDPS023) - Part 2: Multi-year prospective and retrospective performance evaluation

Michael D. Moran[1*], Alexandru Lupu[1], Verica Savic-Jovcic[1], Junhua Zhang[1], Qiong Zheng[1], Elisa I. Boutzis[1], Rabab Mashayekhi[2], Craig A. Stroud[1], Sylvain Ménard[2], Jack Chen[1], Konstantinos Menelaou[2], Rodrigo Munoz-Alpizar[2], Dragana Kornic[2], and Patrick M. Manseau[2]

[1]Air Quality Research Division, Environment and Climate Change Canada (ECCC), Toronto, Ontario, Canada
[2]Canadian Centre for Meteorological and Environmental Prediction, ECCC, Montreal, Quebec, Canada
*Retired

Correspondence to: Michael Moran (mike.moran@ec.gc.ca), Alexandru Lupu (alexandru.lupu@ec.gc.ca), or Craig Stroud (craig.stroud@ec.gc.ca)

Abstract

The online version of the Regional Air Quality Deterministic Prediction System (RAQDPS) is a chemical weather forecast system that has been employed operationally by Environment and Climate Change Canada (ECCC) since 2009. It is run twice daily to produce 72 hour forecasts of hourly 10 km abundance fields of three key predictands, NO_2, O_3, and $PM_{2.5}$ total mass, as well as other gas-phase chemical species, $PM_{2.5}$ chemical components, and dry and wet deposition for Canada, the contiguous U.S., and northern Mexico. Version 023 of the RAQDPS (RAQDPS023) went into service at ECCC in December 2021 and was replaced by the RAQDPS025 in June 2024. A companion paper by Moran et al. (2025) describes the RAQDPS023 in detail. In this paper we present the results of a five-year performance evaluation of prospective and retrospective annual air quality (AQ) simulations made with the RAQDPS023. The annual simulations considered were the first year of RAQDPS023 forecasts in 2021/22 and four years of retrospective annual simulations for the 2013–2016 period that used historical, year-specific emissions. Forecasts made by the RAQDPS-FW023, a duplicate operational system to the RAQDPS023 except for the addition of near-real-time (NRT) biomass burning (BB) emissions, were also evaluated for the 2021/22 period. A NRT measurement data set consisting of hourly NO_2, O_3, and $PM_{2.5}$ surface measurements for Canada and the U.S. was used for the 2021/22 evaluation whereas a much more extensive set of air-chemistry and precipitation-chemistry measurements was used for the 2013–2016 evaluations. Some evaluation results were also compared with results for the 2010–2019 period for forecasts made by earlier operational versions of the RAQDPS and with evaluation results for several peer AQ forecast models. In addition to looking at a number of highly aggregated "headline" scores, many stratified analyses were also performed, including evaluations by network, season, month, hour of day, region, and land-use type. Consideration of simulations for multiple years with the same model and year-specific input emissions helped to identify systematic model errors by reducing the influence of year-to-year variations in meteorology, and a comprehensive evaluation for many species for 2013–2016 supported by stratified analyses provided diagnostic insights that allowed the scientific basis for the RAQDPS023 forecasts to be assessed (i.e., "right answers for the right reasons?"). Although one confounding factor for this study was the sizable reduction in the emissions of some pollutants in North America that occurred from 2013 to 2021, it was found that the trends in AQ observations over this period agreed with the year-specific description of emissions used for the five annual simulations from a rank-ordered perspective.

While RAQDPS023 evaluation scores for hourly NO_2 and O_3 volume mixing ratio forecasts were found to be competitive with peer models and often met suggested performance benchmarks for the five simulation years, another key finding was that the RAQDPS023 forecasts consistently underpredicted hourly $PM_{2.5}$ total mass concentrations for all months in 2021/22 and for the majority of months in 2013–2016. The largest underpredictions occurred in summer and at rural stations whereas overpredictions often occurred in the cold season at urban stations. The model also missed the observed bimodality in monthly $PM_{2.5}$ concentrations and exaggerated the observed diurnal variations in hourly $PM_{2.5}$ concentrations. Additional evaluations with daily $PM_{2.5}$ chemical composition measurements and daily gravimetric $PM_{2.5}$ total mass measurements were also examined to better understand the hourly $PM_{2.5}$ underpredictions. Consistent overpredictions of elemental carbon and sea salt concentrations and underpredictions of sulfate concentration were identified, but scores for predictions of daily gravimetric $PM_{2.5}$ total mass were better than those for hourly $PM_{2.5}$ total mass, directing attention to differences in measurement methods. SO_2 and HNO_3 levels were also found to be overpredicted in general while NH_3 levels were underpredicted: these three gas-phase species are all $PM_{2.5}$ precursors, which raises concerns about some process representations such as those for sulfur oxidation and gas-phase dry deposition. As well, springtime O_3 levels were underpredicted while isoprene levels were consistently overpredicted. The impact of BB emissions on predictions of NO_2, O_3, and $PM_{2.5}$ was also characterized in detail by comparing evaluation results for the 2021/22 RAQDPS023 and RAQDPS-FW023 forecasts. Negligible impact was found for monthly NO_2 forecasts when BB emissions were included, but monthly O_3 forecast scores were modestly improved and monthly $PM_{2.5}$ forecast scores were markedly improved from July to September 2021, as well as summer and annual scores. Taken together, the results of this comprehensive multi-year evaluation point to a number of RAQDPS023 system components where improvements are desirable. These results also provide a strong benchmark against which to compare the performance of future versions of the RAQDPS.

1 Introduction

The use of operational short-range air quality (AQ) forecast systems to predict tomorrow's air quality, also referred to as chemical weather, has expanded rapidly over the last two decades (e.g., Kukkonen et al., 2012; Zhang et al., 2012a,b; WMO, 2020; Brasseur and Kumar, 2021). Environment and Climate Change Canada (ECCC), Canada's federal environment ministry, which is responsible for operational weather forecasting in Canada, began to make operational regional AQ forecasts in 2001. Since that time, numerous upgrades and improvements have been made to this system (Moran et al., 2025). Version 23 of the ECCC Regional Air Quality Deterministic Prediction System (RAQDPS023) became the Canadian operational, continental-scale chemical weather forecast system for North America on 1 December 2021 (Moran et al., 2021b) and continued in this role until June 2024 (CMC-RAQDPS-025, 2024). There was also a clone of the RAQDPS023 forecast system named the RAQDPS-FW023, which was identical except for the addition of near-real-time (NRT) biomass burning (BB) emissions (Pavlovic et al., 2016; Chen et al., 2019; Chen and Menelaou, 2021). The RAQDPS023 and RAQDPS-FW023 were both run twice per day on a 10 km continental grid to produce 72 hour forecasts of hourly surface concentration fields of ozone (O_3), nitrogen dioxide (NO_2), particulate matter with aerodynamic diameter smaller than 2.5 µm ($PM_{2.5}$), and other chemical species and compounds. These forecasts were disseminated to ECCC forecast offices and also directly to the public via a public ECCC website (https://weather.gc.ca/firework/index_e.html). This goal of this paper is to present the results of a multi-year prospective and retrospective performance evaluation of RAQDPS023 AQ predictions, which both quantifies predictive skill and provides an evaluation benchmark against which the performance of future RAQDPS versions can be compared.

The comparison of AQ model predictions with AQ measurements for a chosen simulation period allows the modelling system's performance to be assessed, weaknesses to be identified, and, for some more comprehensive evaluations, potential improvements to be suggested. Initially, such model performance evaluations considered retrospective simulations (or hindcasts) for individual or multiple AQ models that were being used in a regulatory environment (e.g., Dennis and Downton, 1984; Venkatram et al., 1988; Dennis et al., 1993; Tesche et al., 2006; van Loon et al., 2007; Smyth et al., 2009; Solazzo et al., 2012a,b; Yahya et al., 2014; Im et al., 2015a,b; Appel et al., 2021). In the regulatory context, however, there is typically interest in both model skill for the simulation period considered and model skill in predicting AQ changes in response to changes in input emissions or meteorological conditions (e.g., Dennis and Downton, 1984; Gilliland et al., 2008; Pun et al., 2008; Dennis et al., 2010; Foley et al., 2015; Koo et al., 2015; Colette et al., 2017). For AQ forecasting, by contrast, forecast skill under current conditions is the primary concern (Steyn and Galmarini, 2008; Dennis et al., 2010).

Zhang et al. (2012a,b) have provided a review of the history of both regional AQ forecasting in North America and Europe and global AQ forecasting, including performance evaluation approaches, up to 2012. Kukkonen et al. (2012) provided a similar overview for the same period but focussed on operational European regional-scale AQ forecasting models. Zhang et al. (2012a) noted that the 1998 development in the U.S. of the Aerometric Information Retrieval Now (AirNow) program (www.airnow.gov), a NRT data repository and dissemination hub for North American AQ measurements supplied by more than 100 monitoring agencies in the U.S. and Canada, was revolutionary for North American AQ forecasting since it allowed forecasting teams to obtain immediate feedback on model performance (e.g., McKeen et al., 2005, 2007, 2009; Eder et al., 2006, 2009, 2010; Mathur et al., 2008; Chuang et al., 2011; Chai et al., 2013; Lee et al., 2017; Chen et al., 2021; Campbell et al., 2022; Williams et al., 2022). A current example of the use of AirNow data for short-term model performance evaluation is in an ongoing multi-model AQ forecast evaluation for North America that is led by ECCC under the umbrella of the World Meteorological Organization (WMO) Global Air quality Forecasting and Information System (GAFIS) initiative (see **Sect. 4.3**). AirNow data are also used for objective analyses (e.g., Robichaud and Ménard, 2014; Robichaud et al., 2016) and for chemical data assimilation (e.g., Pagowski et al., 2010; Ma et al., 2021).

The NRT measurements available from AirNow, however, have three important disadvantages. First, measurements are only available for six chemical compounds: NO_2, O_3, CO, SO_2, $PM_{2.5}$, and PM_{10}. Second, AirNow is a "meta-network" since the multiple agencies contributing measurement data may employ different instruments and sampling techniques, each with their own biases and errors, to measure the same chemical species. As a consequence, there can be considerable heterogeneity in a combined AirNow measurement data set vs. the uniformity expected of a typical measurement network data set. And third, the AirNow measurements must be viewed as preliminary since they have not undergone the quality assurance/quality control (QA/QC) procedures normally applied by the monitoring agencies before they release new data sets.

The more traditional source of AQ measurement data is to obtain them directly from the lead agency for a monitoring network or from an AQ measurement data clearinghouse such as AQS or NAtChem (see **Table S2a**). However, these finalized network data sets suffer from the significant disadvantage of only being available anywhere from three months to years after sampling, since some AQ measurements require post-sampling calibration while others (e.g., from filterpacks, annular denuders, passive samplers, and precipitation samplers) must undergo laboratory analysis after collection followed by network QA/QC procedures. But such finalized data sets do have three important advantages over the NRT AirNow data. First, they include measurements of many additional chemical species, including more trace gases such as HNO_3, NH_3, and some individual volatile organic compounds (VOCs), $PM_{2.5}$ chemical components, and major inorganic ions in precipitation. Second, even for the six pollutants that are reported to AirNow, not all North American stations that measure these species report to the AirNow data centre. And third, these data sets have been QA/QCed before release. For example, Chai et al. (2013) compared AirNow and AQS hourly O_3 measurements for 2010 and showed scatterplots of differences between the two data sets for a one-month period. The issue of availability, however, means that these finalized AQ network data sets cannot be used for the immediate evaluation of AQ forecasts, that is, prospective AQ simulations, but they are preferable for the evaluation of historical or retrospective AQ simulations since they permit more comprehensive evaluations of predictions of the atmospheric chemical environment using a broader range of QA/QCed measurement data.

A paper by Dennis et al. (2010) proposed a framework for evaluating AQ model performance that consists of four evaluation types: operational; diagnostic; dynamic; and probabilistic. The first two evaluation types are the most relevant for evaluating deterministic AQ forecasts. Operational evaluations address the basic question of how well model predictions of concentration and deposition agree with observations of chemical concentrations and deposition. To do this they use routine measurements of a small set of air-chemistry species, and, infrequently, additional air-chemistry, precipitation-chemistry, and meteorological parameters to calculate standard statistical performance measures (e.g., Table A2). Diagnostic evaluations, on the other hand, are less common and are used to evaluate model inputs and process representations by considering many additional relevant observations such as precursor concentrations, pollutant concentrations aloft, PM composition and size distributions, and meteorological parameters that have a direct impact on pollutant concentrations such as temperature, planetary boundary layer (PBL) height, vertical wind profiles, cloud cover, and precipitation (e.g., Vautard et al., 2012). Diagnostic evaluations can address

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

three additional important questions. First, is agreement between model predictions and observations the result of chance or of good scientific understanding and representation of atmospheric dynamics, physics, chemistry, and emissions? Put another way, is the model getting the right answers for the right reasons? Second, are differences between model predictions and observations due to errors in model input fields or to gaps or errors in model process representations or to computational factors? And third, can the identification of the sources of differences between the model predictions and observations be used to improve the model?

Many operational evaluations have considered only a small number of observed species even if finalized measurement data sets were used (e.g., Chai et al., 2013; Pan et al., 2014; Marécal et al., 2015; Wagner et al., 2015; Lee et al., 2017; Campbell et al., 2022; and Williams et al., 2022). Given the complexity of atmospheric chemistry related to secondary pollutants such as O_3 and to the multiple chemical components of PM (e.g., Sillman, 1999; Meng et al., 1997; Bachmann, 2013), however, such limited evaluations will not provide insights into the reasons for poor model performance. A comprehensive operational evaluation, on the other hand, which makes use of the full range of available AQ measurements, can consider nearly complete mass budgets for some chemical families such as sulphur species or oxidized nitrogen species and hence may be considered closer to a diagnostic evaluation. Comprehensive operational evaluations, however, are relatively uncommon. For example, Huang et al. (2021) reviewed over 300 peer-reviewed articles that reported evaluation results from AQ modelling studies for China and found that very few considered more than seven pollutants. Nevertheless, examples of comprehensive operational evaluations include Biswas et al. (2001), Hogrefe et al. (2001a,b), Zhang et al. (2006a,b), Cai et al. (2008), Yu et al. (2008), Zhang et al. (2009a), Hogrefe et al. (2015), Yahya et al. (2014, 2015), Tessum et al. (2015), Zhang et al. (2016), Chen et al. (2021), and Wang et al. (2021). Lastly, examples of diagnostic evaluations include Zhang et al. (2006c, 2009b), Godowitch et al. (2011), Gan et al. (2015), Knote et al. (2015), Galmarini et al. (2021), and Clifton et al. (2023).

This paper presents the results of an operational performance evaluation of both AQ forecasts and AQ hindcasts made by the RAQDPS023 chemical weather forecast system. Evaluations were performed for five simulation years: (i) the first year of RAQDPS023 (and RAQDPS-FW023) forecasts from 1 June 2021 to 31 May 2022, which used projected anthropogenic input emissions files; and (ii) four years of retrospective annual simulations for the 2013–2016 period performed with the equivalent RAQDPS024 forecast system (same system but ported to a new computer; see Moran et al., 2025) but using historical, year-specific input emissions files. Note that from 1 June to 30 November 2021 the RAQDPS023 and RAQDPS-FW023 systems were run in a parallel (i.e., pre-operational) mode beside the RAQDPS022 and RAQDPS-FW022 systems that were operational at that time before being promoted to operational status on 1 December 2021. In addition, the performance of a decade of operational forecasts made by earlier RAQDPS versions from 1 January 2010 to 30 June 2019 is examined both to show the evolution of forecast skill over this period and to allow comparison with RAQDPS023 scores. Note that RAQDPS-FW023 retrospective simulations for 2013–2016 have not been considered here due to the incompatibility with this period of version 4.1 of the Canadian Forest Fire Emissions Prediction System (CFFEPS) used by the RAQDPS-FW023. CFFEPS v4.1, which depends on a satellite instrument launched in 2017, was not introduced until 2021 (Chen and Menelaou, 2021; Moran et al., 2025).

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

Given that BB emissions also have large year-to-year variations (e.g., Table A4), their neglect may complicate identification of systematic model errors, especially for the summer months. The impact of this omission is examined for 2021/22 in Sect. 4.2.

AirNow data have been used for the performance evaluation of the 2021/22 forecasts since not all finalized network measurement data sets were available for that period during the preparation of this paper. The use of AirNow data does reflect common practice for AQ forecast performance evaluations in the near term and is also consistent with evaluation results for previous RAQDPS operational versions for the 2010–2019 period, which also employed AirNow data (Sect. 4.1**)**. On the other hand, the use of AirNow data limits the number of chemical species that can be considered, and 2021/22 analyses were only performed for NO_2, O_3, and $PM_{2.5}$ total mass. For the four years of retrospective annual runs, however, a much broader set of finalized AQ measurement data, including $PM_{2.5}$ speciation measurements and precipitation-chemistry measurements, was available and was used to carry out as broad and comprehensive an evaluation of model performance as possible. The performance evaluation reported here includes analyses stratified by different measurement characteristics to identify which network, species, month, hour of day, region, and land-use type resulted in the most skillful and the least skillful model predictions. Both Canadian and U.S. AQ measurement data sets were considered for all five years in order to expand the spatial coverage of the evaluation. This differs from many past evaluations of AQ model performance over North America that have only considered U.S. AQ measurements (e.g., Tessum et al., 2015; Yahya et al., 2015; Appel et al., 2017; Toro et al., 2021), although there are exceptions (e.g., Appel et al., 2021). One complicating factor for this study was that emissions of some anthropogenic pollutants decreased materially between 2013 and 2021 (see **Sect. 2.2**), but this factor was also positive in that it allowed examination of the representativeness of the input model emissions that were used and constituted a dynamic evaluation of opportunity (e.g., Gilliland et al., 2008; Godowitch et al., 2010; Foley et al., 2015). The consideration of a total of 15 simulation years facilitated the identification of systematic model biases and errors by revealing common patterns across years and reducing the importance of year-to-year variations in emissions and in meteorology, including its impact on biogenic emissions.

The rest of this paper is organized as follows. Section 2 describes the study methodology, including the model configuration, run setup, and input emissions used to perform the 2013–2016 retrospective annual runs, the AQ measurement data sets used for the evaluation, the data processing and data filtering applied for model-measurement pairing, and the techniques and evaluation metrics used for the performance evaluation. Section 3 and the Supplement (S) present results of the RAQDPS023 performance evaluation for 2021/22 and 2013–2016, where Sect. 3 focuses on aggregate annual analyses for air– and precipitation–chemistry measurements and the Supplement presents more detailed analyses stratified by network, season or month, hour of day, region, or land-use. Section 4 then compares RAQDPS023 performance relative to 2010–2019 RAQDPS forecast performance and 2021/22 RAQDPS-FW023 performance, summarizes RAQDPS023 and RAQDPS-FW023 performance vs. four peer AQ forecast systems, and discusses RAQDPS023 shortcomings revealed by the evaluations. Lastly, Sect. 5 presents a summary and conclusions.

2 Methodology

2.1 Modelling system configurations and setups

The model configuration and run setup of the RAQDPS023 for the 2021/22 forecasts has been summarized in CMC-RAQDPS-023 (2021) and described in detail by Moran et al. (2025). Only a short overview will be given here. Some key aspects include the use of the following: (1) version 5.1.0 of the ECCC Global Environmental Multiscale (GEM) numerical weather prediction (NWP) model code and version 3.1.0.0 of the Modelling Air quality and Chemistry (MACH) chemical weather module code, which is embedded within the GEM code; (2) a rotated limited-area latitude-longitude grid covering North America and adjacent oceans (e.g., Fig. 13) with 10-km horizontal grid discretization and 84 staggered vertical hybrid levels capped by a model lid at 0.1 hPa; (3) a two-time-level iterative-implicit time integration scheme and three-dimensional semi-Lagrangian advection scheme used with a 300 s meteorological time step and a 900 s chemistry time step; (4) imposed tracer mass conservation with an iterative, locally mass-conserving monotonicity correction and a Bermejo-Conde (2002) global mass fixer; (5) a simplified two-bin sectional representation of the PM_{10} size distribution (diameter ranges of 0-2.5 µm and 2.5-10 µm); (6) PM dry chemical composition represented by eight chemical compounds [sulfate (SO_4), nitrate (NO_3), ammonium (NH_4), elemental carbon (EC), primary organic matter (POM), secondary organic matter (SOM), crustal material (CM), and sea salt (SS)]; (7) 41 prognostic gas-phase chemical compounds and 16 prognostic particle-phase section-compounds (i.e., 2 size bins x 8 compounds); (8) ADOM-2 gas-phase chemistry mechanism, ADOM aqueous-phase chemistry mechanism, HETV inorganic heterogeneous chemistry mechanism, and Instantaneous secondary organic Aerosol Yield (IAY) scheme; (9) parameterizations of aerosol particle nucleation, condensation/evaporation, coagulation, dry deposition and sedimentation, hygroscopic growth, and activation; and (10) parameterizations of gas-phase dry deposition and in-cloud and below-cloud scavenging of particles and soluble gases.

The configuration and setup used for the 2013–2016 retrospective annual runs followed those of the RAQDPS023 operational 2021/22 forecasts as closely as possible, but some differences could not be avoided as the retrospective simulations were performed later and outside of the operational environment. One major (and deliberate) difference was the replacement of the RAQDPS023 operational projected input emissions files with year-specific input emissions files based on historical year-specific emissions inventories for 2013–2016 (see next section). A minor (but unavoidable) difference was the need to use an equivalent modelling system (RAQDPS024) for the 2013–2016 hindcasts due to the migration with minimum changes of all ECCC operational and research computing to a new supercomputer in late June 2022. In addition, there were a few minor differences related to near-surface vertical diffusion, model initialization and spin-up, meteorological piloting, and simulation run strategy, whose impacts were small.

First, the RAQDPS023 operational runs employed two adjustments related to near-surface vertical diffusion to avoid the possibility of predicting extremely high surface concentrations due to the conjunction of high surface emissions, an extremely stable PBL, and very low wind conditions such as might occur during northern winter nights under a strong anticyclone. As described by Moran et al. (2025), one preemptive adjustment was to impose a minimum PBL height of 100 m when calculating the vertical diffusion of chemical tracers (where free-atmosphere convection applies above the PBL top); the other was to inject surface emissions into the lowest two model layers instead of the lowest layer (61 m thickness vs. 20 m thickness). For the four years of retrospective runs, however, these two adjustments were removed so that whatever PBL height was forecast by GEM was used in defining the vertical diffusivity profile and surface emissions were injected into the lowest (20 m thick) model layer. The reason for doing so was to test over a four-year period whether the two operational adjustments were needed.

A different approach was also used to maximize the dynamical balance between the mass and momentum fields in the meteorological initialization step. The operational forecast runs for 2021/22 employed an hourly incremental analysis update (IAU) approach from T-3 to T+3 hours, where T=0 is the run start time (e.g., Bloom et al., 1996). For the retrospective runs, on the other hand, a digital filter was employed at T=0 (Fillion et al., 1995). This difference was necessary because archived analyses for T-3 hours were not available for the 2013–2016 period.

The hourly meteorological lateral boundary conditions (LBCs) supplied by a meteorological "piloting" model for the retrospective runs also had a different source. For the 2021/22 operational forecasts these were supplied by version 8.0.0 of the operational 10-km Regional Deterministic Prediction System (RDPS), a limited-area-model configuration of GEM v5.1.0 that was run by ECCC to make meteorological forecasts for North America in advance of the RAQDPS023 run (Moran et al., 2025). The RDPS8.0.0 horizontal grid was a superset of the RAQDPS023 horizontal grid and its vertical levels were identical with those of the RAQDPS023 (CMC-RDPS-8.0.0, 2021). For the retrospective runs, on the other hand, the meteorological LBCs were supplied from special runs of a 15-km global configuration of GEM 5.1.0. This change avoided the need to run both global and regional versions of GEM, and previous tests had shown that the use of a meteorological piloting model with 10 km vs. 15 km grid spacing had very little impact on RAQDPS forecasts.

Lastly, simulation run length was the source of one more difference. RAQDPS023 operational forecast runs were 72 hours in length and were initialized at T-3 hours using the T+9 hour forecast fields from the previous RDPS run launched 12 hours earlier. To save computer time the retrospective runs were only 18 hours in length and were initialized at T=0 hours using the T+12 hour forecast fields from the previous global GEM run. In both cases, though, annual sequences of hourly predicted fields were prepared by concatenating hourly predictions for only the first 12 forecast hours of each run (i.e., T+1 to T+12 hours).

2.2 Input emissions

The RAQDPS023 2021/22 forecasts used the SET4.0.0 anthropogenic emissions data set described by Moran et al. (2021b, 2025). The SET4.0.0 emissions were based on a projected 2020 Canadian national emissions inventory and projected 2023 U.S. and Mexican national emissions inventories, which were roughly in temporal alignment with the forecast period. Note, though, that it was not possible to modify the SET4.0.0 emissions in near-real time to account for rapidly evolving emissions changes in North America associated with the COVID-19 pandemic (cf. Mashayekhi et al., 2021).

To provide year-specific input emissions files for each of the 2013–2016 retrospective annual runs, however, a concerted effort was made to use recently available and consistent national emissions trend data sets for Canada, the U.S., and Mexico. Each of these national emissions trend data sets provides a multi-decadal sequence of annual anthropogenic national emission inventories that were generated using largely consistent emissions estimation methodologies for all of the years considered by each data set. Four year-specific annual anthropogenic national emission inventories were extracted for the 2013–2016 period for Canada from the ECCC TREND16 emissions trend data set. Similarly, four year-specific annual anthropogenic national emission inventories were extracted for the 2013–2016 period for the U.S. from the U.S. EPA EQUATES (Epa's air QUAlity TimE Series) national emissions trend data set. The EQUATES data set also includes a set of annual anthropogenic national emissions inventories and annual wildfire emissions inventory files for 2002-2019 for Mexico, so that year-specific Mexican inventories for the 2013–2016 period were also available. More details about these data sets are provided in **Sect. S2.2** of the Supplement.

Model-ready emissions files for the years 2013–2016 were then prepared from these national inventories using version 4.8 of the Sparse Matrix Operator Kernel (SMOKE) emissions processing system (https://www.cmascenter.org/smoke/; Zhang et al., 2018b). The use of an emissions processing system was necessary because while the national inventories of the three countries report annual emissions of seven criteria air pollutants by jurisdiction (province, county, or state), the RAQDPS023 requires gridded emissions fields for each emitted model species for every hour of every day of the year (e.g., Dickson and Oliver, 1991; Houyoux et al., 2000; Matthias et al., 2018; Zhang et al., 2018b). Additional details about the processing of these anthropogenic emission inventories with SMOKE are provided in Sect. S2.2.

Several types of natural emissions were also accounted for. Time-varying biogenic emissions were included in all five annual simulations, where biogenic emissions were calculated for each time step of the RAQDPS023 simulations using code for a modified version of the Biogenic Emission Inventory System (BEIS) v3.09 biogenic emissions algorithms with inputs of two GEM-predicted meteorological fields: surface temperature and solar insolation (Moran et al., 2025). Time-varying sea-salt emissions were also included for all five annual simulations; these emissions were calculated for each time step based on surface wind speed. Hourly BB emissions, on the other hand, were only considered in the 2021/22 RAQDPS-FW023 forecast runs (Sect. 4.2). Note that some BB emissions might be due to very large prescribed burns or grass fires as well as wildfires if these were detected by satellite (Moran et al., 2025). Note also that neither system version considered some other types of natural emissions, namely natural wind-blown fugitive dust emissions, lightning emissions, pollen and other biological emissions, other marine emissions, and volcanic emissions.

Table 1 presents a summary of annual inventory emissions of the seven criteria pollutants for Canada, the U.S., and Mexico for the five years for which annual RAQDPS023 runs were performed. The rows named "Total Anthro" and "Total Biogenic" are, respectively, the annual, SMOKE-processed anthropogenic emissions and the annual,

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

dynamically-calculated biogenic emissions within the model domain that the model "sees" (i.e., responds to). The domain-total "Total Anthro" and "Total Biogenic" values thus include all Canadian emissions but only U.S. emissions from the 48 contiguous U.S. states and part of Alaska and exclude emissions from the rest of Alaska, Hawaii, and some U.S. territories in the Caribbean and the Pacific Ocean and only include Mexican emissions from the 340 Mexican counties out of 2,457 that lie completely or partially within the RAQDPS023 domain (e.g., Fig. 13).

Some significant changes are evident in annual emissions over this nine-year time period in Table 1, first over the four-year period from 2013 to 2016, then from 2016 to 2021/22, and in total from 2013 to 2021/22. For example, North American "Total Anthro" SO_2 emissions decreased by 37% from 2013 to 2016 and then by a further 37% from 2016 to 2021/22, for a total decrease of 60% relative to 2013, while "Total Anthro" NO_x emissions decreased by 19%, 22%, and 36% for the same three periods. "Total Anthro" VOC and CO emissions also decreased over the three periods, by 8%, 1%, and 9% for VOC emissions and by 13%, 14%, and 25% for CO emissions. "Total Anthro" NH_3 emissions, on the other hand, were nearly constant during the 2013–2016 period but then increased in Canada while decreasing in the U.S. and Mexico in 2021/22 for an overall domain-total decrease of 6% from 2013 to 2021/22. Lastly, the domain-total SMOKE-processed values for "Total Anthro" $PM_{2.5}$ and PM_{10} emissions are considerably lower than total inventory values due to the impact of adjustments for land-use-dependent, near-source removal due to settling and impaction (i.e., transportable fraction), but further meteorology-dependent emission reductions due to snow cover and wet soil are only applied during the RAQDPS023 simulation (Moran et al., 2025).

Table 1 also compares SMOKE-processed, domain-total annual anthropogenic emissions with calculated domain-total annual biogenic emissions of NO_x and VOCs. Biogenic NO emissions can be seen to have contributed 4% of domain-total NO_x emissions in 2013, rising to 6% in 2021/22 as anthropogenic NO_x emissions decreased and biogenic NO emissions increased. By contrast biogenic VOC emissions contributed 78% of domain-total VOC emissions in 2013, a considerably larger percentage, and then rose to 81% in 2021/22 as anthropogenic VOC emissions declined and biogenic VOC emissions increased. In addition, Table S1 compares the seasonal variation of the SMOKE-processed SET4.0.0 anthropogenic emissions for 32 model species and the biogenic emissions of four model species. Seasonal variations depend strongly on both the pollutant and its source types and can have markedly different cycles. For example, NH_3 is primarily emitted by agricultural activities and can be seen to have a pronounced winter minimum and summer maximum, whereas seasonal emissions of two lumped VOC species, ALD2 (acetaldehyde and higher aldehydes) and CRES (cresols and phenols), have a pronounced winter maximum and summer minimum consistent with their dominant source being residential wood combustion. For most anthropogenic species, however, seasonal variations were considerably smaller, including SO_2, NO, and NO_2. Although fossil-fuel power generation is important for both SO_2 and NO_x emissions, this suggests that at the continental scale the increased power load for space heating in the winter in North America is roughly balanced by the increased power load for air conditioning in the summer. The seasonal variation of biogenic emissions, on the other hand, was more like NH_3; they had strong seasonal cycles with winter minima and summer maxima. In fact, biogenic isoprene emissions were predicted to have the most pronounced domain-level seasonal cycle, increasing from just 2% in the winter to 65% in the summer (Table S1).

2.3 Air-chemistry and precipitation-chemistry observations

Routine air-chemistry and precipitation-chemistry measurements are available from multiple measurement networks operating in Canada and the U.S. The hourly measurements of NO_2, O_3, and $PM_{2.5}$ total mass made by continuous instruments that are reported in near-real time by some agencies to the U.S. EPA's AirNow program (Dye et al., 2004; Wayland et al., 2004; Zhang et al., 2012a) have already been mentioned. AirNow hourly measurement data for NO_2, O_3, and $PM_{2.5}$ from U.S. monitors have been combined in this study with NRT NO_2, O_3, and $PM_{2.5}$ hourly measurement data from Canadian National Air Pollution Surveillance (NAPS) monitors that report directly to ECCC's Canadian Meteorological Centre (CMC). This combined NRT data set has been used to evaluate the 2021/22 RAQDPS023 and RAQDPS-FW023 operational forecasts of these species.

The use of these NRT abundance measurements, which are considered to be preliminary, for model evaluation is consistent with the operational nature of the RAQDPS023 forecasts. Two automated data filters were applied by CMC to the AirNow and NAPS NRT measurements upon receipt before they were used for model evaluation or other purposes such as operational air pollutant objective analyses (e.g., Robichaud et al., 2016). The first filter flagged negative abundance values and above-threshold abundance values as suspicious (NO_2 over 200 ppbv, O_3 over 200 ppbv, $PM_{2.5}$ over 300 $\mu g \cdot m^{-3}$) or invalid (NO_2 over 2000 ppbv, O_3 over 500 ppbv, $PM_{2.5}$ over 1000 $\mu g \cdot m^{-3}$), while the second filter flagged large jumps in abundances between consecutive hours as suspicious (over 30 ppbv for NO_2, over 60 ppbv for O_3, over 90 $\mu g \cdot m^{-3}$ for $PM_{2.5}$). Values flagged as suspicious or invalid were not used for this study. A completeness criterion was also imposed on this data set to ensure temporal representativeness of the evaluation data: individual station data sets were required to have at least 75% valid values out of the total possible values for a one-year evaluation period to be considered complete. Table 2 lists the number of available AirNow and NAPS stations for 2021/22 that measured hourly O_3, NO_2, and $PM_{2.5}$ abundances while Figure S2 shows the locations of these stations. Note that some stations only measured one or two of these species.

The evaluation of the retrospective annual simulations for 2013–2016, on the other hand, was based on finalized AQ network data sets. The CAPMoN and NAPS networks in Canada and the AMoN, AQS, CASTNET, CSN, IMPROVE, NATTS, and PAMS networks in the U.S. provide air-chemistry measurement data sets for various chemical species, including hourly NO_2, O_3, and other gas-phase species (e.g., SO_2, CO, HNO_3, and NH_3), hourly $PM_{2.5}$ and PM_{10} mass, daily FRM (Federal Reference Method) and FEM (Federal Equivalent Method) $PM_{2.5}$ mass (e.g., Noble et al., 2001; Gantt, 2022), and daily $PM_{2.5}$ chemical composition, while the CAPMoN network in Canada and the NADP network in the U.S. provide precipitation-chemistry measurement data sets. Details about each of these networks are given in Table S2a, and Sect. S2.3 provides some additional information about daily $PM_{2.5}$ measurements. Note that all network data sets used in this study were accessed on 24 June 2024 (relevant because these data sets are always subject to change even years after their original release). Table 2 and Tables S2b-d list the number of stations for each network for 2013–2016 with available measurements and with complete measurements of various chemical species (cf. Sect. S2.4), while Figs. S3–S6 show the locations of available stations by network.

Although AQ measurements provide important chemical information about the real atmosphere, these measurements, like AQ model predictions, also have biases and errors. For example, NO_2 measurements made with chemiluminescence monitors frequently have positive biases due to interference from other oxidized nitrogen species (e.g., Dunlea et al., 2007; Lamsal et al., 2015), and NH_3 measurements made with passive monitors have negative biases (Puchalski et al., 2011). Measurements of both $PM_{2.5}$ total mass and semi-volatile $PM_{2.5}$ chemical components are known to have both positive and negative artifacts (e.g., Chow, 1995; Frank, 2006; Watson et al., 2009; Dabek-Zlotorzynska et al., 2011; Malm et al., 2011; Su et al., 2018; Gantt, 2022). Estimates of reconstructed $PM_{2.5}$ total mass based on the sum of $PM_{2.5}$ species mass measurements from the CSN, IMPROVE, and NAPS networks often differ from direct measurements of $PM_{2.5}$ total mass (e.g., Malm et al., 2011; Chow et al., 2015; Hand et al., 2019). As well, different networks measuring the same species often do not use the same instruments or follow the same field and laboratory protocols, which can affect comparability across networks. Examples include the use of different instruments to measure SO_2, CO, and O_3 concentrations by different agencies reporting to AQS (e.g., Demerjian, 2000; Parrish and Fehsenfeld, 2000), SO_2 and HNO_3 concentrations measured by the CAPMoN and CASTNET networks vs. the NAPS network (Dabek-Zlotorzynska et al., 2011; Feng et al., 2020), NH_3 concentrations measured by the AMoN and NAPS networks (Dabek-Zlotorzynska et al., 2011; Puchalski et al., 2011), sulphur and nitrogen species measured by the CASTNET and IMPROVE networks (e.g., Ames and Malm, 2001; Lavery et al., 2009), and $PM_{2.5}$ chemical components measured by the IMPROVE and CSN networks (Hand et al., 2012; Solomon et al. 2014). In order to assess measurement comparability between networks some efforts have been made to co-locate instruments used by different networks at one or more locations, including CSN and IMPROVE (Malm et al., 2011; Hand et al., 2012), CASTNET and CAPMoN (Schwede et al., 2011), and CAPMoN and NADP (Sirois et al., 2000; Wetherbee et al., 2010; Feng et al., 2023).

2.4 Pairing measurements with model predictions

The comparison of AQ measurements and AQ model predictions must be done with care since measurements and model predictions never represent exactly the same quantities (e.g., Seigneur and Moran, 2004; Appel et al., 2008). From a temporal perspective, reported AQ measurements may either be instantaneous values or time averages, and the time averages may in turn represent mean values for an averaging period that begins, ends, or straddles the reporting time. Model values, on the other hand, are nominally instantaneous but correspond to a time that is a discrete time step after the previous integration time and after a particular operator calculation step in a repeating sequence of operators. From a spatial perspective, AQ measurements are typically made at a near-surface "point" location whereas AQ model predictions represent a volume-average value corresponding to the volume of an individual model grid cell. This spatial representativeness discrepancy is sometimes referred to as incommensurability. It is a fundamental source of model uncertainty that can be reduced by reducing model grid spacing but never removed entirely (e.g., Nappo et al., 1982; Venkatram, 1988; McNair et al., 1996; Spicer et al., 1996; Swall and Foley, 2009; Stroud et al., 2011; Schutgens et al., 2016). Estimates of population exposure to air pollutants based on ambient measurements from a small number of AQ measurement stations also suffer from the same problem (e.g., Jerrett et al., 2005; Hystad et al., 2011). And lastly, from a chemical perspective AQ measurements sometimes correspond to a combination of two or more model variables while in other cases they correspond to more detailed chemical species than the AQ model is able to consider. For these reasons some pre-processing is generally required to pair or match AQ measurements and model predictions before they are compared while bearing in mind that some differences will still remain, and the methodology used to perform this pairing should be documented (Simon et al., 2012).

Temporal pairing was relatively straightforward for this study, although it was necessary to examine AQ network documentation carefully to understand the exact temporal nature of the measurements being reported. Model concentration predictions were available every chemistry time step, or every 15 minutes in the case of the RAQDPS023 (Moran et al., 2025), so it was simple to pair model predictions with instantaneous hourly concentration measurements. On the other hand, to pair model predictions with the mean hourly concentration measurements reported by the NAPS and AQS networks (Fig. S2), multiple consecutive sub-hourly model predictions needed to be combined using one of two approaches: an end-value approach, which averages the model values at the beginning and end of the measurement sampling period; or an integration approach, where the trapezoidal rule is used to combine the five RAQDPS023 values available for the hour. The end-value approach was used in this study. To pair model predictions with mean daily air concentration measurements from the AQS, CSN, IMPROVE, and NAPS networks (Table S2a), all hourly model values for each 24-hour sampling period were averaged. To pair with mean weekly air concentration measurements from CASTNET, all hourly model values for each weekly sampling period were averaged; the same weekly averaging was also performed for daily mean CAPMoN and NAPS air concentrations for temporal consistency in order to be able to compare evaluation statistics between networks properly and to pool measurements for these three networks. Similarly, to pair model predictions with mean biweekly AMoN measurements, hourly model NH_3 values were averaged for each biweekly sampling period, and the same was done for NAPS daily mean NH_3 values for temporal consistency. Finally, to pair model predictions with daily precipitation-chemistry measurements (CAPMoN) or weekly precipitation-chemistry measurements (NADP), hourly model deposition forecasts were accumulated for the appropriate period, but then weekly deposition values were also calculated for CAPMoN for temporal consistency with NADP.

More details about pairing related to this study, including spatial and chemical considerations, especially for PM measurements, and completeness screening, are provided in Sect. S2.4. These details include the nuances of evaluating VOC predictions, handling ambient vs. standard temperature and pressure, the measurement proxies used for some PM chemical components such as NH^4, OM (=POM+SOM), CM, and SS, and the combined gas-particle phases implicit in precipitation-chemistry measurements.

2.5 Model performance metrics and evaluation

Once a set of paired observed and model-predicted values is available, model performance metrics can be calculated. For the present study we chose to calculate 12 statistical metrics: observation mean ($\overline{\mathrm{O}}$); model prediction mean ($\overline{\mathrm{M}}$); mean bias (MB); normalized mean bias (NMB); normalized mean absolute error (NME); root mean square error (RMSE); Pearson correlation coefficient (R); fraction of predictions within a factor of 2 of observations (FAC2);

centered root mean square error (CRMSE); standard deviations of observations (σ_O or SDO) and model predictions (σ_M or SDM); and normalized standard deviation (NSD). Definitions of these metrics are provided in Table A2 and some background information about their selection is provided in Sect. S2.5. Inclusion of the first seven of these metrics is consistent with the recommendation of Simon et al. (2012) for a minimum set of performance evaluation statistics that should always be calculated to promote comparability across separate studies. The eighth metric, FAC2, is a dimensionless and bounded (0–1) measure of error or scatter that is not sensitive to outliers (e.g., Chang and Hanna, 2004; Borrego et al., 2008; Derwent et al., 2010; Savage et al., 2013). CRMSE, unlike RMSE, is insensitive to bias and represents the error due to differences in pattern variation or, alternately, the standard deviation of the error; it has been used in many studies (e.g., Bencala and Seinfeld, 1979; Stanski et al., 1989; Taylor, 2001; Chang and Hanna, 2004; Entekhabi et al., 2010; Sakaguchi et al., 2012; Thunis et al., 2012). NSD has been suggested as a metric by Taylor (2001), Chang and Hanna (2004), and Thunis et al. (2012), but by itself it does not provide information about the magnitudes of σ_O or σ_M (which were recommended by Willmott (1981) and reported by Appel et al. (2021)). Note that CRMSE, R, σ_O, and σ_M (and sometimes NSD) are all linked by the Taylor diagram (Taylor, 2001). One other quantity that should be reported with these 12 metrics is N, the number of measurement-model pairs. Many evaluation studies fail to report this quantity, but it is used explicitly in the calculation of many metrics and it provides valuable information about sample size, representativeness, and significance (e.g., Huang et al., 2021).

Since this is an AQ forecasting evaluation, we have focused here on “native” network sampling duration: that is, hourly, daily, weekly, or biweekly, but based on the networks with the longest sampling duration for each species (e.g., biweekly for AMoN for NH_3, but weekly for CASTNET and NADP-NTN) to allow consistent comparisons between networks and pooling of network measurements. Other studies that focused on model performance for regulatory applications have looked at “constructed” predictands such as maximum daily 8-hr average (MDA8) or maximum 1-hourly values of O_3 volume mixing ratios (VMRs), whereas in this study we have only considered network-reported values such as hourly O_3 VMRs. And while we report annual domain-average statistics based on combined network measurements, we also report results for more stratified (i.e., disaggregated) analyses, including network-specific statistics, seasonal statistics, monthly statistics, diurnal statistics, regional statistics, and urban/rural statistics. To calculate urban vs. rural statistics, each measurement site was classified as urban or rural based on its grid-cell population density, where a threshold of 400 persons km^{-2} was applied for Canada and 386 persons km^{-2} (1000 persons per square mile) for the U.S. as the minimum urban population density. The slightly different thresholds are used for consistency with national censuses. We have also discussed our results contextually by reference to the performance benchmarks proposed by Simon et al. (2012), Emery et al. (2017), Kelly et al. (2019), Huang et al. (2021), and Zhai et al. (2024) (see Sect. S2.5). Some additional discussion related to model performance metrics can also be found in Sect. S2.5.

3 Results

This section presents evaluation results, first for one meteorological parameter important for air quality, then for three key chemical species, NO_2, O_3, and $PM_{2.5}$ total mass, and then for other gas-phase species, $PM_{2.5}$ chemical composition, and wet concentration and deposition of three inorganic species. Annual, seasonal, monthly, diurnal, and regional evaluations for the one-year period from 1 June 2021 to 31 May 2022, the first year of RAQDPS023 forecasts, are presented in this section along with selected evaluation results for the 2013–2016 annual simulations. Many additional tables and figures related mainly to the 2013–2016 simulations, which serve to further quantify predictive skill and characterize the temporal and spatial variability of system performance, can be found in the Supplement.

3.1 Operational evaluation of AQ-relevant meteorological predictands

Meteorological processes affect air quality through their influence on emissions, atmospheric transport and diffusion, chemistry, and wet and dry removal. Near-surface temperature, wind speed, and precipitation are three meteorological parameters important for surface air quality (e.g., Vautard et al., 2012; Gilliam et al., 2015; McNider and Pour-Biazar, 2020; Wang et al., 2021; Campbell et al., 2022). As described by Moran et al. (2025), the RAQDPS023 is a regional chemical weather model configured to produce nearly identical meteorological forecasts to the RDPS 8.0.0, the ECCC regional weather forecast model that was operational at the same time (Fillion et al., 2010; Caron et al., 2015; McTaggart-Cowan et al., 2019; CMC-RDPS-8.0.0, 2021). Evaluations of RDPS weather forecasts have been presented in these and other publications. For this paper we have only evaluated RAQDPS023 precipitation forecasts based on precipitation measurements from AQ networks, which are not usually available to or considered in NWP model performance evaluations. This choice also ensures that performance statistics for precipitation are consistent in time and space with those for pollutant concentrations in precipitation and wet deposition (Sect. 3.3.3).

Domain-average annual scores for weekly precipitation forecasts at precipitation-chemistry stations for 2013 to 2016 are listed in Table 6. We have chosen to consider weekly forecasts because the U.S. NADP precipitation-chemistry network only reports weekly accumulated measurements. Interestingly, this set of scores does not show much variation from year to year: for example, annual MB values ranged from -0.1 to 2.2 mm·week^{-1}, NMB values from -0.01 to 0.11, NME values from 0.49 to 0.54, RMSE values from 17.6 to 20.6 mm·week^{-1}, FAC2 values from 0.56 to 0.57, and R values from 0.71 to 0.78. Appel et al. (2011) reported a comparable NMB range but a markedly lower RMSE range for 12-km MM5 model simulations for the 2002-2006 period, but those earlier statistics were calculated for accumulated seasonal and annual precipitation predictions as opposed to weekly predictions, which have greater temporal variability.

Seasonal analyses of RAQDPS023 predictions of near-surface temperature, wind speed, and precipitation can be found in Sect. S3.1 as well as individual-network annual and seasonal evaluation results and subregional annual evaluation results for precipitation. These additional evaluations showed that model skill in predicting weekly precipitation was highest for the winter season and lowest for the summer season (e.g., Fig. S161). The probable explanation is that synoptic-scale precipitation, which is more predictable, is likeliest to occur in the winter whereas small-scale convective precipitation, which is harder for NWP models to predict, is likeliest to occur in the summer (e.g., Appel et al., 2011; Gilliam et al., 2021). Many of the largest overpredictions of weekly precipitation at individual stations occurred in the Rocky Mountain region of the western U.S., where subgrid-scale (SGS) topographical features and station location are likely to be important, whereas underpredictions were common in the southeastern and central U.S. (Fig. S121). In addition, the evaluation statistics for an individual network or month or subregion were sometimes qualitatively different from those for the combined networks, combined months, or combined subregions. For example, model skill in predicting annual and seasonal mean weekly precipitation was found to be higher for the Canadian CAPMoN precipitation-chemistry network than the U.S. NADP network (Tables S6A, S6S). This possibility always needs to be kept in mind when interpreting the most highly aggregated performance statistics (e.g., annual statistics and all-network, all-station statistics), which may obscure systematic network differences or be impacted by compensating errors (e.g., Makar et al., 2014).

3.2 Operational evaluation of three key air quality predictands

This section presents operational evaluation results for RAQDPS023 predictions of NO_2, O_3, and $PM_{2.5}$ total mass for five years. Results are presented first for 2021/22 NO_2 and O_3 forecasts based on NRT measurements, followed by results for 2013–2016 NO_2 and O_3 hindcasts based on QA/QCed network measurements, 2021/22 $PM_{2.5}$ forecasts based on NRT measurements, and 2013–2016 $PM_{2.5}$ hindcasts based on QA/QCed network continuous and gravimetric measurements.

3.2.1 NO_2 and ozone

2021/22 operational forecasts of NO_2 and O_3

Figure 1 shows the spatial distribution over North America of annual mean NO_2 and O_3 hourly surface VMR fields predicted by the RAQDPS023 for the 2021/22 period. Coloured "coffee beans" (i.e., divided dots) are superimposed on the contoured fields to show observed and predicted annual mean values at NRT measurement stations for the same period (see Fig. S2 for station locations). Generally good agreement is evident in Fig. 1 between the observed and predicted annual mean values of both pollutants, although the higher NO_2 VMRs associated with urban centers are smaller-scale features caused by high NO_x emissions over urban centers (cf. Fig. S1a),.

More quantitatively, Table 3 lists values of all-station annual model evaluation statistics for hourly NO_2 and O_3 VMR forecasts for 2021/22. These overall performance scores are generally less good for NO_2 than for O_3, including NMB (-0.19 vs. -0.07), NME (0.56 vs. 0.28), FAC2 (0.52 vs. 0.83), and R (0.65 vs. 0.72). This difference is not surprising given that NO_2 is a primary pollutant whose spatial distribution is dominated by the distribution of emissions with strong spatial gradients whereas O_3 is a secondary pollutant with a smoother spatial pattern and smaller dynamic range. One indicator of the degree of smoothness of a pattern is the coefficient of variation (CV) or relative standard deviation (ratio of standard deviation to arithmetic mean; see Table A2), where a lower value indicates a greater smoothness (i.e., lower variability) (e.g., Fruin et al., 2014; Lee et al., 2018). The observed and predicted annual CV values for 2021/22 calculated from Table 3 were 1.16 and 1.20, respectively, for NO_2 vs. 0.50 and 0.47 for O_3.

In order to judge the level of model skill suggested by the Table 3 scores, Zhai et al. (2024) have recommended benchmark goals for NMB, NME, and R of ±0.20, 0.40, and 0.60 for "good" NO_2 performance scores (i.e., scores above 67th percentile relative to the scores for a historical multi-model ensemble) while Emery et al. (2017) have recommended benchmark goals for NMB, NME, and R of ±0.05, 0.15, and 0.75 as good O_3 performance scores. These benchmark goals were met by RAQDPS023 forecasts for 2021/22 except for NO_2 annual NME scores and O_3 annual R scores. When interpreting these benchmark comparisons, however, it should also be noted that Simon et al. (2012) found performance scores for retrospective model applications to be better on average than those for forecast applications due to the use by the former of year-specific emissions, meteorological reanalyses, day-specific chemical lateral boundary conditions, and other retrospective data sets that are not available to AQ forecasting applications.

Figures 2 and 3 provide a disaggregation of the all-station annual scores from Table 3 by showing spatial distributions of station-specific annual values of four statistics (MB, NMB, CRMSE, and R) for hourly NO_2 and O_3 measurements, respectively, for 2021/22. Such plots can reveal regional patterns in the evaluation statistics. For example, annual NMB values for NO_2 tend to be negative everywhere but they are more negative (i.e., worse) in general at western stations while CRMSE and R values for NO_2 are lower in the continental interior than the coastal areas (Fig. 2). Annual MB and NMB values for O_3, on the other hand, are generally negative at western stations but positive at eastern stations, particularly at coastal stations (Fig. 3), while both annual CRMSE and R scores are higher overall across the continent for O_3 than for NO_2.

Figure 5 adds temporal detail to the annual analysis shown in Fig. 1. It shows the corresponding predicted spatial distributions of seasonal mean NO_2 and O_3 hourly surface VMR fields for 2021/22, again with superimposed coloured divided dots to show observed and predicted seasonal mean values at NRT measurement stations for each season. By inspection predicted domain-scale, seasonal mean NO_2 levels appear to be highest in the winter season (DJF) and lowest in the summer season (JJA), whereas O_3 levels are predicted to be highest in the spring season (MAM) and lowest in the autumn (SON) season.

Another perspective on temporal variation of model performance is provided by Figs. 6 and 7, which show time series of observed and predicted all-station monthly mean VMR values of hourly NO_2 and O_3, respectively, for 2021/22 for all NRT measurement stations in the model domain as well as time series of monthly NMB, CRMSE, and R scores. Observed and predicted monthly mean NO_2 VMRs are both highest in January and lowest in June and July; observed and predicted monthly mean O_3 VMRs are both highest in April and lowest in November. Monthly NMB values for hourly NO_2 are negative for all months, monthly CRMSE values for NO_2 peak in January, and monthly R values for NO_2 do not vary much from month to month but are slightly higher in the winter. Monthly NMB values for hourly O_3 are negative for all months except December and January, monthly CRMSE values for O_3 peak in July, and monthly R values for O_3 also do not vary much but are slightly higher in the summer. It should be noted that seasonal variations in NO_x emissions are very small at the domain scale (Table S1), suggesting that the observed and predicted variations in monthly mean NO_2 levels evident in Fig. 6 are controlled by other factors than emissions, such as monthly variations in temperature, photolysis, PBL height, vegetation phenology, and dry deposition. For O_3, on the other hand, biogenic emissions of VOC, its other main precursor, have very large seasonal variations (Table S1). Interestingly, although the largest predicted monthly O_3 values occur in April, the model still underpredicts the well-known springtime O_3 maximum in the Northern Hemisphere (e.g., Penkett and Brice, 1986; Monks, 2000; Liudchik et al., 2015) by about 5 ppbv.

It is also of interest to examine model performance by time of day since many emission source sectors and meteorological and chemical processes vary diurnally. Figures 9 and 10 show all-station, annual-mean diurnal time series in local time (LT) of five statistics for hourly NO_2 and O_3 surface VMRs, respectively, for the 2021/22 period. The annual-mean diurnal time series of observed and predicted VMRs for both species display a strong dependence on time of day as do the diurnal time series of annual NMB, CRMSE, and R scores. Model predictions of annual-mean hourly values of both NO_2 and O_3 surface VMRs agree well overall with observations, including the times of the observed daily maxima and minima. The annual-mean diurnal time series of NO_2 VMR and the associated evaluation statistics in Fig. 9 display extrema close to the times of morning and afternoon rush hours, suggesting that diurnal variation of on-road emissions plays an important role in driving the diurnal pattern, while the maximum annual-mean hourly O_3 VMR occurs at 14 LT and the minimum at 05 LT (Fig. 10). For O_3 the smallest annual-mean hourly NMB and CRMSE values and highest annual-mean hourly R values occur close to the mid-day O_3 peak.

Figures 2–4 showed how model performance can vary geographically. A complementary result is presented in Fig. 12, which compares regional time series of observed and predicted monthly means of hourly NO_2 and O_3 VMRs for 2021/22 for the four continental quadrants shown in Fig. S7. Both observed and predicted time series exhibit a regional dependence. For NO_2 the agreement between observed and predicted monthly means was closest for western Canada while for O_3 it was closest for the eastern U.S. Interestingly, peak observed monthly mean NO_2 values were slightly higher in the west than in the east, at least for these regional sets of stations (see Fig. S2). Observed monthly mean NO_2 peaks also occurred in January in three of the four regions and in December in the eastern U.S., in overall agreement with Fig. 6. Similarly, peak observed monthly mean O_3 values occurred in April in three of the four regions and in May in the western U.S., in overall agreement with Fig. 7, but peak predicted monthly mean O_3 values occurred in both March and April. Note too that monthly mean NO_2 VMRs were also underpredicted in all months in the western and eastern U.S., in agreement with Fig. 6, but some monthly overpredictions can also be seen in western and eastern Canada.

2013–2016 hindcasts of NO_2 and O_3

The RAQDPS023 annual hindcasts for 2013–2016 can also be evaluated to look for consistencies in model performance across multiple years. Figure 13 shows plots of predicted spatial distributions of annual mean NO_2 and O_3 surface VMR fields for 2013–2016 and 2021/22. For both species the broad spatial patterns are very similar over land for the five years despite year-to-year variations in meteorology and the monotonic decrease of 18% in domain-total NO_x emissions over the 2013–2016 period and the further decrease of 20% from 2016 to 2021/22 (Table 1). Nevertheless, year-to-year decreases in annual NO_2 levels are visible over this near-decadal period, including in Texas, the Ohio Valley, and the Washington, D.C.–Boston corridor. Latitudinal gradients in the spatial distributions of O_3 surface VMR can be seen over land in Fig. 13 for all five years, with a east–west band of elevated values stretching across the continental U.S. and peaking in the elevated terrain of the U.S. Rocky Mountain and Great Basin regions. A recent analysis of surface O_3 observations showed a similar pattern (Gaudel et al., 2018). However, trends in O_3 from 2013 to 2021/22 are not obvious in this figure, unlike those for NO_2.

Table 3 lists values of observed and predicted all-station annual mean NO_2 and O_3 surface VMRs for 2013–2016 as well as 2021/22. Note that the statistics for the 2013–2016 period are based on quality-assured, retrospective observation data sets released by individual agencies rather than the NRT measurements used to evaluate 2021/22 forecasts (see Fig. S3 for station locations). Consistent with Fig. 13, observed and predicted all-station annual mean NO_2 VMR values both exhibit a monotonic decrease from 2013 to 2021/22, though for a smaller set of stations in 2021/22 (Table 2), whereas observed all-station annual mean O_3 VMR values exhibit little change over this period vs. a small upward trend for predicted all-station annual mean O_3 VMR. Table 3 also lists domain-wide annual values of 10 other model performance statistics for the five years. Statistics for the hindcasts might be expected to be better than the forecasts due to the use of year-specific emissions. In fact, the scores are mixed and are comparable overall for the five years. For example, annual NMB values for NO_2 are negative for 2021/22 but positive and smaller in magnitude for the other four years, annual RMSE and NME values for NO_2 are better for 2021/22 than for 2013–2016, but FAC2 and R values are better for 2013–2016 than for 2021/22. The annual NMB and R values for NO_2 for 2013–2016 all exceed the benchmark goals of 0.20 and 0.60 for good performance recommended by Zhai et al. (2024), but annual NME scores do not meet either of their recommended thresholds (0.40 or 0.55). Annual NMB values for O_3, on the other hand, are more negative for 2013–2016 than for 2021/22 and FAC2 scores are also lower while NME and R scores are comparable. In addition, the annual NMB and R scores for O_3 for 2013–2016 are all above the acceptable benchmarks of ±0.15 and 0.50 for these statistics recommended by Emery et al. (2017) but fall below the more stringent benchmark goals of ±0.05 and 0.75, while NME scores do not meet either recommended threshold (0.15 or 0.25).

Results from additional data analyses for NO_2 and O_3 with a focus on the 2013–2016 hindcasts can be found in Sect. S3.2.1. These results include tables of separate annual and seasonal scores for the AQS and NAPS networks as well as regional scores for all five years, spatial plots of both annual station scores and predicted seasonal mean surface VMR fields for 2013–2016, seasonal and regional diurnal analyses, and monthly time series, monthly density scatterplots, and urban vs. rural monthly time series for 2013–2016. One finding from these supplemental analyses is the high level of consistency between the aggregated annual statistical scores across years for both species for the AQS and NAPS networks individually (Table S3A) and annual scores at the individual station level (e.g., Fig. S42). Another is the clear consistency across the 2013–2016 hindcasts of the seasonal variations in the NO_2 and O_3 seasonal mean VMR fields, statistical scores, diurnal time series, and monthly mean time series (e.g., Fig. S140), which helps to identify systematic model errors. Third, the overall agreement in observed and predicted temporal trends also provides support for the representativeness of the year-specific emissions used for these hindcasts (e.g., Fig. S141). Fourth, there are some striking differences between the time series of monthly NO_2 and O_3 statistics for urban stations vs. rural stations that underline the importance of emissions forcing (e.g., Fig. S209 vs. Fig. S210). Fifth, the larger annual, seasonal, and monthly negative biases for NO_2 that were found in the 2021/22 forecasts for U.S. regions vs. Canadian regions (e.g., Fig. S168) suggest that the U.S. NO_x emissions used for these forecasts may have been too low whereas the Canadian NO_x emissions that were used were more representative of 2021/22 conditions. The same pattern was not seen in the 2013-2016 hindcasts, which used different emissions. To detect this possible issue, however, statistics for the individual networks had to be computed and then compared (Table S3A). Sixth, the annual regional analysis found that the two western regions had the largest negative NMB values for O_3, consistent with Fig. 12 and pointing to possible issues with the O_3 lateral boundary conditions (Table S7). And seventh, the monthly density scatterplots reveal an obvious precision limitation (only whole numbers) in the reported hourly NO_2 measurements (Fig. S169).

3.2.2 $PM_{2.5}$ total mass

2021/22 operational forecasts of $PM_{2.5}$ total mass

Figure 1 also shows the spatial distribution over North America of the annual mean $PM_{2.5}$ hourly surface concentration field (without sea salt) predicted by the RAQDPS023 for 2021/22. Coloured divided dots are again superimposed to show observed and predicted annual values at NRT hourly $PM_{2.5}$ measurement stations for the same period (see Fig. S2 for station locations). It is clear from this figure that the RAQDPS023 underpredicts annual $PM_{2.5}$ levels for 2021/22 at a majority of measurement stations.

Table 3 lists values of observed and predicted all-station annual mean $PM_{2.5}$ surface concentrations (including sea salt) and 10 other annual evaluation statistics for hourly $PM_{2.5}$ forecasts for all stations measuring hourly $PM_{2.5}$ (including both Class III FEM and non-FRM/FEM monitors: see Sect. S2.3) for 2021/22. All-station annual MB and NMB values for forecast hourly $PM_{2.5}$ concentrations were -2.5 $\mu g \cdot m^{-3}$ and -0.31, respectively, consistent with Fig. 1. Other statistics in Table 3 for $PM_{2.5}$ were also poorer (e.g., NME=0.66, FAC2 =0.46, R=0.24, NSD=0.69) than corresponding scores for NO_2 and O_3. Emery et al. (2017) proposed "acceptable" score benchmarks (i.e., above 33rd percentile for a historical multi-model ensemble) for predicted $PM_{2.5}$ total mass for NMB, NME, and R scores of ±0.30, 0.50, and 0.40, but none of these benchmarks were met.

Figure 4 shows the spatial distribution of station-specific annual values of MB, NMB, CRMSE, and R for hourly $PM_{2.5}$ total mass for 2021/22 based on NRT hourly measurements at AirNow and NAPS stations. Consistent with Fig. 1 and Table 3, annual MB was negative for most stations, but values were small at a minority of stations and even positive at a handful of stations. Annual NMB values, on the other hand, were greater than -0.40 at the majority of stations, especially in the west. Annual CRMSE values were highest in the western U.S away from the coast; as discussed in Sect. 4.2 these western scores were influenced by the lack of BB emissions in the RAQDPS023 runs. Lastly, annual R scores were highest in the northeast and along the U.S. west coast and are lowest for many Rocky Mountain, Great Plains (central U.S.), and Prairie (central Canada) stations.

Figure 5 shows the predicted spatial distributions of *seasonal* mean $PM_{2.5}$ hourly surface concentration fields (without sea salt) for 2021/22, again with superimposed divided dots that show observed and predicted seasonal mean $PM_{2.5}$ concentration values at NRT hourly measurement stations. Similar to Fig. 1, observed seasonal mean values were higher in general than predicted seasonal mean values, especially in the summer and in the west (see Sect. 4.2 for a discussion of the impact of the inclusion of wildfire emissions).

Figure 8 adds further temporal detail by showing time series of observed and predicted monthly mean $PM_{2.5}$ concentration values for 2021/22 as well as time series of monthly NMB, CRMSE, and R values based on all NRT measurements of hourly $PM_{2.5}$ surface concentration in the model domain. The observed all-station monthly mean $PM_{2.5}$ concentration time series has a large August peak and a lower January peak but the reverse is true for the predicted monthly mean $PM_{2.5}$ time series. The RAQDPS023 also underpredicted monthly mean $PM_{2.5}$ concentrations in all months, with the largest underpredictions occurring in the summer and the smallest in the winter. While monthly NMB was negative for all months, it was most negative for spring and summer, with an extreme value close to -0.6 in August. Monthly CRMSE was quite variable, but with a peak in August and a secondary peak in January. Monthly R values did not vary much for winter and spring, but they were considerably lower for summer and autumn and were close to zero for July and August. The poor summer scores were in large part due to the neglect of BB emissions (Sect. 4.2).

Figure 11 shows annual-mean diurnal time series of five statistics for hourly $PM_{2.5}$ surface concentration for 2021/22. Model performance for $PM_{2.5}$ clearly varied with time of day. Predicted annual-mean hourly concentrations were biased low at all hours of the day, and they displayed more diurnal variability than the measurements. The largest negative annual-mean hourly NMB value occurred in the early afternoon (14 LT) while the smallest values occurred at the beginning of the morning rush hour (06 LT) and towards the end of the evening rush hour (20 LT). Annual-mean hourly CRMSE and R values, on the other hand, exhibited relatively small diurnal variations.

It is also of interest to examine how model performance varied with geography. Figure 12 compares regional time series of observed and predicted monthly means of hourly $PM_{2.5}$ surface concentration for 2021/22 for the four continental quadrants (Fig. S7). Peak observed monthly mean $PM_{2.5}$ concentrations occurred in July or August and were higher in the west than in the east, but as in Fig. 8 there was also a secondary peak in the cold season in December or January for all four regions (cf. Fig. 5). In contrast the predicted monthly mean $PM_{2.5}$ concentrations had a primary cold-season peak in December or January and a weak warm-season peak in July or August, again consistent with Fig. 8. Note that the predicted warm-season peak occurred without any contribution from BB emissions (see Sect. 4.2), suggesting a second warm-season emissions source such as biogenic secondary organic aerosol (SOA). Another regional difference is that predicted monthly mean $PM_{2.5}$ concentrations were higher than the observed values in early winter for eastern Canada but were lower for all months for the other three regions.

2013–2016 hindcasts of $PM_{2.5}$ total mass

Figure 13 shows the predicted spatial distributions of annual mean $PM_{2.5}$ surface concentration fields (including sea salt) over North America for 2013–2016 and 2021/22. The spatial patterns of annual mean $PM_{2.5}$ concentration for these five years are broadly similar, although some minor interannual variations can be seen over the ocean regions due to variations in sea-salt emissions due to interannual differences in near-surface wind speed (cf. Fig. S9). Annual mean $PM_{2.5}$ surface concentrations, like annual mean NO_2 surface VMRs, are highest over the eastern U.S. and California and lower over most of Canada and the rest of the western U.S., with the exception of isolated urban areas in the western U.S. and a tongue of elevated $PM_{2.5}$ levels over the Canadian province of Alberta. The impact of the large decreases in domain-total anthropogenic emissions of two $PM_{2.5}$ precursors, SO_2 and NO_x, of 60% and 31% from 2013 to 2021 (Table 1) is also reflected in this figure by a decrease in $PM_{2.5}$ levels with time over North America (see also multi-year plots of annual mean $PM_{2.5}$-SO_4 and $PM_{2.5}$-NO_3 concentration fields in Fig. 14).

Table 3 lists annual values of 12 evaluation statistics for $PM_{2.5}$ hourly surface predictions for all stations measuring hourly $PM_{2.5}$ for the 2013–2016 hindcasts in addition to the 2021/22 forecasts. The values of predicted annual mean $PM_{2.5}$ surface concentrations show a downward trend across the five simulation years: the observed annual mean surface concentrations also show a downward trend but it is much weaker. Overall, the 2013–2016 scores were very similar amongst themselves, suggesting consistent behaviour in RAQDPS023 $PM_{2.5}$ predictions from year to year, but some differences are evident between the 2013–2016 scores and the 2021/22 scores. Annual NMB values were negative for all five years, but the range for 2013–2016 was -0.06 to -0.09, considerably better than the 2021/22 value of -0.31. Annual NME scores, on the other hand, were slightly worse for 2013–2016 (0.71–0.73) than 2021/22 (0.66), while annual FAC2 scores were slightly better (0.48–0.50 vs. 0.46) and annual R scores were comparable (0.17–0.29 vs. 0.24). Note that all of the annual NMB scores for 2013–2016 (unlike 2021/22) but none of the NME and R scores met the $PM_{2.5}$ benchmarks for acceptable performance recommended by Emery et al. (2017).

In addition to hourly $PM_{2.5}$ total mass concentration measurements there are also roughly 900 daily $PM_{2.5}$ monitors operating in North America that make 24-hour $PM_{2.5}$ mass measurements using a filter-based, gravimetric approach (see Sect. S2.3 and Table S2c). These daily gravimetric $PM_{2.5}$ measurements mainly support regulatory applications and are not available in near-real time, but they are notable because they represent an independent data source for model evaluation that can supplement the Class III FEM and non-FRM/FEM hourly $PM_{2.5}$ concentration measurements used for NRT evaluations and the Table 3 statistics. Daily gravimetric $PM_{2.5}$ measurements also have different uncertainties than continuous $PM_{2.5}$ measurements, including negative artefacts due to the volatilization of semi-volatile species such as ammonium, nitrate, and particle water during transport to and inside the controlled laboratory environments where filter analyses are performed (e.g., Frank, 2006; Dabek-Zlotorzynska et al., 2011; Malm et al., 2011; Chow et al., 2015; Hand et al., 2019). Gravimetric $PM_{2.5}$ monitor locations from the AQS, IMPROVE, and NAPS networks are shown in Fig. S6a (vs. Fig. S3c for continuous $PM_{2.5}$ monitors). Note that these combined networks include monitors with every-day, one-day-in-three, and one-day-in-six sampling frequencies, but roughly 65% of the daily gravimetric $PM_{2.5}$ mass measurements in the U.S. are made by non-speciation, mass-only monitors (Tables 5 and S2c; Malm et al., 2011).

Table 5 lists all-station annual evaluation statistics for gravimetric measurements of daily $PM_{2.5}$ mass for the 2013–2016 hindcasts. First note that observed annual $PM_{2.5}$ total mass concentration values for the gravimetric data set in Table 5 have a range from 7.2 to 8.2 µg·m^{-3}, similar to but smaller than the corresponding range of 7.4 to 8.7 µg·m^{-3} for the continuous $PM_{2.5}$ data set from Table 3. The range of annual NMB scores for the gravimetric data set, however, is 0.0 to 0.07, slightly better and of opposite sign to the -0.09 to -0.06 range for the continuous $PM_{2.5}$ measurements in Table 3. Other annual scores for the 2013–2016 hindcasts are also better for the gravimetric measurement data set. The range of annual NME scores for the gravimetric $PM_{2.5}$ measurements is 0.51 to 0.54 (vs. 0.71 to 0.73 for the continuous $PM_{2.5}$ measurements), the range of annual FAC2 scores is 0.68 to 0.70 (vs. 0.48 to 0.50), and the range of annual R scores is 0.43 to 0.47 (vs. 0.17 to 0.29). Note the narrow range of each of these annual statistics for the four years, suggesting a high level of consistency in model performance. Note too that the range of annual NSD scores for the gravimetric measurements is 1.41 to 1.61, higher than the range of 0.78 to 1.36 for the continuous $PM_{2.5}$ measurements.

Some differences in scores for these two independent data sets are to be expected due to differences in monitor locations, in temporal aggregation (daily vs. hourly), and in measurement technologies. One reason for some of the better scores for the gravimetric $PM_{2.5}$ mass measurements may be their daily sampling period, which will reduce the influence of short-term model errors compared to the hourly measurements (e.g., Appel et al., 2008, 2021). The sampling period would not, however, affect annual means or MB and NMB scores. One technical difference between the gravimetric $PM_{2.5}$ mass measurements and continuous mass measurements is that the filter analysis for the former is performed under constant-temperature, low-humidity conditions in a laboratory after transport from the field and storage prior to analysis whereas the hourly continuous mass measurements are made under ambient conditions where temperature and humidity can vary widely. This means that gravimetric $PM_{2.5}$ mass values are likely to be lower due to loss of some semi-volatile mass from ammonium nitrate, organic matter, or particle water, especially for high-humidity or low-temperature ambient conditions. The fact that annual NMB scores for gravimetric $PM_{2.5}$ mass measurements are more positive than those for continuous $PM_{2.5}$ measurements is thus surprising.

The results of additional analyses for hourly continuous and daily gravimetric $PM_{2.5}$ total mass measurements with a focus on the 2013–2016 hindcasts are presented in Sect. S3.2.2. These results include tables of annual and seasonal scores for the individual AQS, IMPROVE, and NAPS networks as well as regional scores for all five years, spatial plots of both annual station scores and predicted seasonal mean $PM_{2.5}$ surface concentration fields for 2013–2016, seasonal and regional diurnal analyses, and monthly time series, monthly density scatterplots, and urban vs. rural monthly time series. One insight from these additional analyses are the considerable variations between seasons that are evident across all five years in the seasonal mean $PM_{2.5}$ mass fields, in the seasonal statistical scores, and in monthly mean time series (e.g., Fig. S142). Another is the high level of consistency between seasonal scores for $PM_{2.5}$ for the 2013–2016 hindcasts (and in many cases for the 2021/22 forecasts) at the aggregated all-station level and annual scores at the individual station level, which allows systematic model errors to be identified. For example, all-station monthly NMB scores were most negative in summer for all four years (Figs. S142 and S198). Third, some of the analyses suggest that the neglect of BB emissions by the RAQDPS023 was an important contributing factor to its overall underpredictions of $PM_{2.5}$ total mass. Fourth, there were some striking differences for 2013-2016 between the time series of monthly $PM_{2.5}$ statistics for urban stations (Fig. S213) vs. rural stations (Fig. S214) that underline the importance of emissions forcing. For example, monthly NMB values for the urban stations were positive for some months whereas they were uniformly negative for the rural stations, suggesting that $PM_{2.5}$ underprediction is mainly a rural issue. Fifth, additional differences are shown between evaluation statistics for 2013-2016 for daily gravimetric $PM_{2.5}$ total mass measurements vs. hourly continuous $PM_{2.5}$ measurements. These differences include positive monthly NMB scores for the cold-season months for the gravimetric measurements (Fig. S198) vs. lower scores for the continuous measurements (Fig. S142). And sixth, some of the evaluation scores between individual networks were also very different, and certain differences in the overall characteristics of the individual networks, in particular the dominance of either urban stations or rural stations, can help to explain why network scores were different. For example, for the daily gravimetric $PM_{2.5}$ measurements the seasonal MB and NMB scores for one measurement network (NAPS) were positive for all seasons and for a second network (AQS) they were positive for most seasons (Table S5S). For the hourly continuous $PM_{2.5}$ measurements, on the other hand, the seasonal MB and NMB scores were negative for both networks for all seasons. The fact that some scores point to model overpredictions of $PM_{2.5}$ make it clear that the negative all-station MB and NMB scores presented in Table 3 and Figs. 5, 8, and 11 do not tell the whole story. Multiple factors must be considered in addition to model formulation to explain these different scores, including the magnitude and distribution of emissions, seasonal and regional variations in meteorology, differences in network composition, and differences in measurement instrument characteristics.

3.3 Expanded evaluation for additional species and processes

As described in Moran et al. (2025), the RAQDPS023 predicts abundances of 47 gas-phase chemical species and 16 size bin-chemical components. While direct atmospheric measurements are not available for all of these species and components, this section describes evaluation results based on 2013–2016 QA/QCed measurements of nine atmospheric gases in addition to NO_2 and O_3, seven $PM_{2.5}$ chemical components, and three aqueous-phase inorganic ions.

3.3.1 Other gases

Table 4 extends the all-station annual statistical scores reported in Table 3 for two gas-phase species, NO_2 and O_3, to nine other individual or lumped ADOM-2 gas-phase species predicted by the RAQDPS023 for 2013–2016: NO, NO_x, HNO_3, NH_3, SO_2, CO, ETHE, HCHO, and ISOP. The measurements considered were provided by five networks: AMoN, AQS, CAPMoN, CASTNET, and NAPS. As discussed in Sect. 2.3 and illustrated by Figs. S4 and S5, the available sets of measurements for these other gas-phase species are different from those for NO_2 and O_3 in terms of the numbers and the locations of surface monitors, and also, in some cases, the sampling period, which ranged from hourly to biweekly (see also Tables S2a and S2b). Looking at values of N, the number of complete measurements per year, in Table 4, we see that NO, NO_x, and CO have the most measurements, followed by ETHE and ISOP, then SO_2, HCHO and HNO_3, and lastly NH_3 with the fewest measurements. However, due to the different sampling period lengths, the number of measurements does not necessarily reflect the number of measurement stations. For example, there are more monitors measuring HNO_3, NH_3, and HCHO than there are for ETHE and ISOP (Table S2b) even though N is smaller for the first three species. Note that the evaluation scores for HNO_3, NH_3, and SO_2 will also be referred to in the next section, since these three species are precursors to three $PM_{2.5}$ chemical components.

Compared to the annual model performance for NO_2 and O_3 predictions for 2013–2016 summarized in Table 3, the overall model skill for these other gas-phase species presented in Table 4 is more varied. For example, all-station annual mean NO VMRs, like those for NO_2, were overpredicted for all four years. All-station annual NMB, NME, FAC2, and R scores for hourly NO VMR, however, were less good than those for hourly NO_2 VMR for all four years, but this difference is at least partly due to the limited precision at which NO VMR measurements were reported (see Sect. S3.3.1). All-station annual NMB, NME, FAC2, and R scores for hourly CO and daily HCHO VMRs, on the other hand, were comparable to those for NO_2. All-station annual mean SO_2, HNO_3, and ISOP VMRs were overpredicted for all four years (and all months: see Sect. S3.3.1), whereas all-station annual mean NH_3 VMR was underpredicted for all four years (and all months: Fig. S145). All-station annual NMB scores for hourly ETHE predictions vs. hourly ethene measurements ranged from 0.68 to 1.19 (i.e., overpredictions), but these scores were confounded by the model's inclusion of isoprene oxidation products in this lumped VOC species (Moran et al., 2025), suggesting that overpredictions should be expected. The impact of decreasing SO_2 and NO_x emissions in 2013–2016 (Table 1) can also be clearly seen in corresponding decreases in Table 4 in observed and predicted all-station annual mean SO_2 and HNO_3 VMR values.

More evaluation results for 2013–2016 for these other gas-phase species are provided in Sect. S3.3.1, including tables of annual and seasonal scores for individual Canadian and U.S. networks as well as all-station regional scores, spatial plots of both annual station scores and predicted seasonal mean surface VMR fields, time series of all-station monthly statistics, and monthly density scatterplots. It is clear from these additional analyses that all nine species exhibit strong seasonal variations: four species (HNO_3, NH_3, HCHO, ISOP) were observed and predicted to have summer maxima, four (NO, NO_x, SO_2, CO) were observed and predicted to have winter maxima, and ETHE was observed to have a winter maximum but predicted to have a summer maximum (Table S4S). As noted above, the disagreement for ETHE was not surprising due to the inconsistency between measured ethene (C_2H_4) and lumped model ETHE, and winter ETHE scores, when isoprene emissions were low, were better than those for the other three seasons (Table S4S). As was the case for NO_2 and O_3, there were also marked similarities in the scores for these other gas-phase species across the four annual hindcast simulations, which supports identification of systematic model errors. For example, HNO_3, SO_2 and ISOP were overpredicted and NH_3 was underpredicted in all seasons and months (Table S4S; Figs. S146, S150, and S145).

The consistency in scores for the four years also suggests that the year-specific emissions used for these hindcasts were representative. However, some scores raise concerns that Canadian SO_2 emissions and biogenic ISOP emissions may have been too high and NH_3 emissions too low for the 2013–2016 period. The decreases in SO_2 and NO_x emissions from 2013 to 2016 were also reflected in time series of both observed and predicted monthly mean VMRs for NO,

HNO_3, and SO_2 (Figs. S143, S144, S146) as well as NO_2 (Fig. S140). Performance benchmarks for CO and SO_2 proposed by Zhai et al. (2024) are also discussed in Sect. S3.3.1, and more CO scores than SO_2 scores met these benchmarks. Scores for individual networks can also be quite different owing to different network characteristics. SO_2 provides a good illustration of this behaviour as SO_2 measurements were available from four networks: scores for the CAPMoN and CASTNET networks were significantly better overall than those for the AQS and NAPS networks for both annual and seasonal evaluations (Tables S4A, S4S). It was noted that evaluation scores also tended to be similar for species with similar characteristics, such as primary species vs. secondary species and species having similar emissions sources (e.g., combustion). Observed and predicted annual and seasonal CV values for the other gas-phase species were also considered. Based on CV values, they fell naturally into three groups, depending on whether the pollutants were primary, secondary, or mixed primary-secondary in nature. Lastly, it was noted that a few scores for 2016 appeared to be outliers and monthly density scatterplots for hourly NO and CO revealed obvious precision problems with reported measurements for these two species (Figs. S172, S176).

3.3.2 $PM_{2.5}$ chemical components

As emphasized by Bachmann (2013), $PM_{2.5}$ is a multi-pollutant. This means that accurate forecasts of $PM_{2.5}$ total mass require accurate forecasts of its underlying chemical components, each of which has different emissions sources and different formation pathways. Forecast errors for $PM_{2.5}$ total mass can thus be better understood by examining forecast errors for its underlying chemical components. Fortunately, three North American networks (CSN, IMPROVE, NAPS) make measurements of $PM_{2.5}$ composition (Sect. 2.3 and Table S2a). Station locations for these three networks are plotted in Fig. S6b, while Table S2c summarizes the number of stations for 2013–2016 for which $PM_{2.5}$ speciation measurements were available and the smaller number for which temporally representative seasonal and annual measurements were available (see Sect. S2.4).

Figure 14 shows plots of the spatial distributions of annual mean surface concentrations of nine $PM_{2.5}$ chemical components over North America for 2013–2016 and 2021/22 predicted by the RAQDPS023. Note that the sets of contour intervals used vary by component, with $PM_{2.5}$-EC, $PM_{2.5}$-NH_4, and $PM_{2.5}$-CM having the smallest ranges and $PM_{2.5}$-SS and $PM_{2.5}$-TOM (= $PM_{2.5}$-POM + $PM_{2.5}$-SOM) having the largest ranges. It is clear from this figure that the predicted spatial distributions vary markedly between chemical components. It is also clear that the predicted annual spatial distributions of each of these chemical components for these five years are broadly similar, pointing to the anchoring effect of relatively constant emissions to the atmosphere of primary $PM_{2.5}$ chemical components and $PM_{2.5}$ gas-phase precursors (e.g., Fig. S1), which for most pollutants changed relatively little from year to year (Table 1). The annual mean surface concentrations of $PM_{2.5}$ total mass shown in Fig. 13 also displayed this broad similarity. However, comparison of the spatial distributions of annual mean $PM_{2.5}$-SO_4 and $PM_{2.5}$-NO_3 surface concentrations from 2013 to 2021/22 does suggest decreasing concentrations of these two chemical components over this period, consistent with the large decreases in domain-total anthropogenic emissions of SO_2 and NO_x from 2013 to 2021/22 (Table 1). Interestingly, annual mean $PM_{2.5}$-NH_4 surface concentration fields can also be seen to decrease from 2013 to 2021/22 despite nearly constant NH_3 emissions over this period. This downward trend is due instead to the reduced availability of gaseous H_2SO_4 and HNO_3 in the atmosphere, which form sulfate and nitrate particulate salts with NH_3. Domain-wide primary $PM_{2.5}$ emissions also decreased by 8% from 2013 to 2016, and some decline is evident from 2013 to 2016 for $PM_{2.5}$-EC and $PM_{2.5}$-POM in Fig. 14. The annual spatial distributions of other $PM_{2.5}$ components display only small year-to-year variations with the exception of sea salt, for which interannual variations are driven by interannual variations in surface wind speed (see Fig. S9).

Although the above discussion suggests the close connection between the spatial distributions of emissions of primary $PM_{2.5}$ components and $PM_{2.5}$ gas-phase precursors and the resulting spatial distributions of $PM_{2.5}$ chemical components in the atmosphere, it is not a simple relationship since many chemical and physical processes in the atmosphere modulate the chemical transformation and removal pathways of these multiple chemical components. For example, the annual spatial distributions of $PM_{2.5}$-SO_4 shown in Figure 14 are quite smooth, which reflect its origin as a secondary pollutant resulting from gas-phase or aqueous-phase oxidation of North American SO_2 emissions, even though the SO_2 emissions shown in Fig. S1d are largely emitted by isolated point sources (e.g., ECCC, 2018; Foley et al., 2023). The spatial distribution in Fig. 14 of $PM_{2.5}$-SOM, another secondary pollutant, is also smooth, but its maximum is located over the southeastern U.S. where biogenic VOC emissions are high, especially in the summer season (cf. Fig. S27). It is also clear from Fig. S1 that there are marked differences in the spatial distributions of some of the anthropogenic emissions associated with different $PM_{2.5}$ chemical components. For example, the majority of NO_x emissions are located in the eastern half of North America, but the locations of some major highways and large urban areas in western North America are visible in the $PM_{2.5}$-NO_3 surface concentration panels in Fig. 14, which suggests the important contributions of on-road mobile sources and population centres to NO_x emissions. Emissions of primary $PM_{2.5}$ and of NH_3 gas, the precursor to $PM_{2.5}$-NH_4, on the other hand, are stronger over the North American interior (Figs. S1e,f). Elevated NH_3 emissions from agricultural activities in the midwestern U.S. are visible in Fig. S1e as well as fertilizer application in the San Joaquin Valley of California, large animal feedlot operations in Texas and Oklahoma, and extensive swine production in North Carolina. While NH_3 emissions are dominated by agricultural activities, some NH_3 emissions are also associated with population centres due to on-road mobile emissions (e.g., Toro et al., 2024). In addition, the chemical composition of primary $PM_{2.5}$ emissions depends on the emissions source type. Combustion sources dominate $PM_{2.5}$-EC and $PM_{2.5}$-POM emissions, as can be seen in Fig. 14 by their association with major highways and population centres, while the $PM_{2.5}$-TOM surface concentration field displays characteristics of both the $PM_{2.5}$-POM and $PM_{2.5}$-SOM fields. Fugitive dust emissions from paved and unpaved roads are the main source of $PM_{2.5}$-CM, and $PM_{2.5}$-CM surface concentrations can be seen in Fig. 14 to be elevated in both urban centres and rural areas. Lastly, the spatial distribution of $PM_{2.5}$-SS is dominated by its oceanic sources, but the limited transport of sea salt from the oceans inland over most of North America evident in Fig. 14 should also be noted.

Table 5 presents all-station annual scores for seven $PM_{2.5}$ chemical components for 2013–2016 for the three networks combined. The scores for each component tend to be similar from year to year, but these scores can also vary considerably between components. For example, all-station annual NMB scores for $PM_{2.5}$-SO_4 and $PM_{2.5}$-NO_3 were negative for all four years whereas all-station annual NMB scores for $PM_{2.5}$-NH_4, EC, CM, and SS were positive for all four years. Only $PM_{2.5}$-TOM had small annual NMB values of both signs for this period. Note that the $PM_{2.5}$-NH_4 overpredictions are inconsistent with the underpredictions of both $PM_{2.5}$-SO_4 and $PM_{2.5}$-NO_3. One possible explanation is that this is an artefact due to the lack of available IMPROVE $PM_{2.5}$-NH_4 measurements so that only CSN and NAPS measurements were considered for $PM_{2.5}$-NH_4 in Table 5. However, the same inconsistency can be seen in Table S5A for just CSN measurements. Another possible explanation is that the RAQDPS023 does not consider the neutralization of $PM_{2.5}$-SO_4 and $PM_{2.5}$-NO_3 by base cations, which would reduce $PM_{2.5}$-NH_4 concentrations and increase NH_3 VMR (e.g., Vasilakos et al., 2018; Miller et al., 2024). All-station annual NME scores in Table 5 were lowest for $PM_{2.5}$-SO_4 (~0.47), followed in rank order by annual NME scores for $PM_{2.5}$-NO_3, TOM, EC, NH_4, CM, and SS (~1.45). All-station annual FAC2 scores were highest for $PM_{2.5}$-SO_4 (~0.64), followed in decreasing order by those for $PM_{2.5}$-EC, TOM, NH_4, NO_3, SS, and CM (~0.31). All-station annual R scores were also highest for $PM_{2.5}$-SO_4 (~0.66), followed in decreasing order by those for $PM_{2.5}$-NO_3, EC and SS, NH_4, TOM, and CM (~0.17). Based on the annual NME, FAC2, and R scores taken together, the RAQDPS023 showed the most skill for $PM_{2.5}$-SO_4 followed by $PM_{2.5}$-NO_3, EC, TOM, NH_4, SS, and CM. Note also that all-station annual NSD scores fell into two groups: values for $PM_{2.5}$-SO_4, NO_3, and NH_4 were less than one whereas values for $PM_{2.5}$-EC, TOM, CM, and SS were greater than one. The three components in the first group are all secondary components whereas those in the second group are all primary components or a mixed primary-secondary component in the case of $PM_{2.5}$-TOM. While this difference might appear to suggest that the model is overemphasizing the contribution of temporal variations due to emissions, the $PM_{2.5}$ speciation measurements are 24-hour samples so that the observed and predicted temporal variation at measurement locations can only be due to interday and longer variations. For anthropogenic emissions this would point to the day-of-week and month-of-year temporal profiles that have been assumed by the emissions processing system for different source sectors (Sect. S2.2), but emissions for the two primary components that had the largest NSD values, $PM_{2.5}$-CM and SS, are also the ones most affected by meteorology.

It is worth noting that downward trends can be seen in Table 5 in the values of both observed and predicted annual mean concentrations for $PM_{2.5}$-SO_4, NO_3, and NH_4, consistent with the monotonic decreases in SO_2 and NO_x emissions that occurred over this period and with Fig. 14. For the $PM_{2.5}$-SO_4 evaluation statistics, the values of annual RMSE and R also decreased from 2013 to 2016. This is similar to decreases in these two statistics reported by Kelly et al. (2019) for the CMAQ model over the 2007 to 2015 period, which they attributed to decreasing SO_2 emissions and lower summertime $PM_{2.5}$-SO_4 peaks, which in turn reduced the $PM_{2.5}$-SO_4 “signal” (e.g., Chan et al., 2018). Note also that 2016 annual $\overline{O}$, $\overline{M}$, MB, NMB, RMSE, R, σ_O, and σ_M scores for $PM_{2.5}$-SO_4, NO_3, EC, and OC for two recent versions of the CMAQ model (with wildfire emissions) were given in Appel et al. (2021). Although minor methodological differences such as inclusion or exclusion of NAPS measurements are suggested by comparisons of the 2016 $\overline{O}$ and σ_O scores for the two models, there is rough agreement between the CMAQ scores and the RAQDPS023 2016 scores in Table 5 for $PM_{2.5}$-SO_4, NO_3, and EC while CMAQ $PM_{2.5}$-OC scores cannot be compared directly with RAQDPS023 $PM_{2.5}$-TOM scores.

$PM_{2.5}$ speciation measurements can also be used to calculate $PM_{2.5}$ reconstructed mass as a weighted sum of the individual $PM_{2.5}$ chemical species (e.g., Malm et al., 2011; Chow et al. 2015; Hand et al., 2019). This is often done using the IMPROVE formula, which uses some measured species as proxies for the OM, CM, and SS components, which are not measured directly. The RAQDPS023 $PM_{2.5}$ total mass forecasts, on the other hand, which are the sum of the seven predicted $PM_{2.5}$ speciated chemical components shown in Table 5 (including SS), are thus also a reconstructed mass but do not require the use of any proxies. A slightly modified version of the IMPROVE formula has been used in this study to calculate $PM_{2.5}$ reconstructed mass from speciation measurements (see Sect. S3.3.2 for more details).

Figure 15 compares observed and predicted all-station seasonal mean $PM_{2.5}$ reconstructed mass and chemical composition for 2013–2016 based on combined CSN, IMPROVE, and NAPS measurements that were complete (Sect. S3.3.2). There is good agreement overall between observed and predicted $PM_{2.5}$ reconstructed mass seasonal means. Predicted $PM_{2.5}$ total mass was greater than observed $PM_{2.5}$ reconstructed total mass for all four winters, while the opposite was true for all four summers, with closer agreement for spring and autumn. This good agreement might seem surprising given the consistent model underpredictions of hourly $PM_{2.5}$ mass measurements described in Sect. 3.2.2, but it is more consistent with the better evaluation results for daily gravimetric $PM_{2.5}$ mass measurements discussed in that section. Similar comparisons for the CMAQ model against CSN and IMPROVE measurements have been presented for 2011 and 2016 simulations by Appel et al. (2017, 2021). Interestingly, both models overpredicted $PM_{2.5}$ total mass in winter 2016 and underpredicted it in summer 2016.

In addition, the stars plotted in Fig. 15 indicate the seasonal means of observed gravimetric $PM_{2.5}$ total mass for 2013–2016. It is natural to compare gravimetric $PM_{2.5}$ mass and $PM_{2.5}$ reconstructed mass since they both attempt to measure of the same quantity. Good agreement can be seen in the observations between these two measurements for three seasons but not for the summer, for which the gravimetric total mass was larger, consistent with findings by Malm et al. (2011) for the CSN and IMPROVE networks. Predicted seasonal means of $PM_{2.5}$ total mass, on the other hand, were greater than the gravimetric seasonal means in winter and autumn but were even smaller than the observed $PM_{2.5}$ reconstructed mass in summer. A closely related quantity, the $PM_{2.5}$ residual mass, is defined to be the difference between gravimetric $PM_{2.5}$ mass and reconstructed $PM_{2.5}$ mass (e.g., Hand et al., 2019). Observed seasonal mean $PM_{2.5}$ residual mass was greater than 0.5 $\mu g \cdot m^{-3}$ for summer but was small otherwise, whereas predicted seasonal mean $PM_{2.5}$ residual mass was greater than 1 $\mu g \cdot m^{-3}$ for summer but was negative for winter, with values ranging from -1.47 to -0.90 $\mu g \cdot m^{-3}$ (see Table S5S-mr). Hand et al. (2019) found observed seasonal $PM_{2.5}$ residual mass to be mostly positive after 2011 with a strong summer peak. Given the marked model underpredictions of summer mean gravimetric $PM_{2.5}$ mass but overpredictions of winter mean gravimetric $PM_{2.5}$ mass (**Sect. S3.2.2**), the factors that contribute to the non-negligible observed summer $PM_{2.5}$ residual mass may be of interest because they may also be relevant to the model. Malm et al. (2011), Chow et al. (2015), and Hand et al. (2019) have suggested that some of these factors may be the neglect of particle-bound water in calculating $PM_{2.5}$ reconstructed mass, ammonium and nitrate volatilization under laboratory conditions, and seasonal variations in the OM:OC scaling ratio (lower in winter, higher in summer).

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

Figure 15 also shows good agreement overall between observed and predicted seasonal mean $PM_{2.5}$ chemical composition. The dominant contribution of the $PM_{2.5}$-EC and TOM carbonaceous components to $PM_{2.5}$ total mass is evident in both the observed and predicted stacked bar graphs as are the anticorrelated seasonal variations in $PM_{2.5}$-SO_4 and $PM_{2.5}$-NO_3. Overpredictions of $PM_{2.5}$-TOM concentration in the winter but underpredictions in the summer can be seen for all four years. The decrease in the observed and predicted contribution of the three major inorganic ions (SO_4, NO_3, NH_4) to $PM_{2.5}$ total mass from 2013 to 2016 due to decreases in annual SO_2 and NO_x emissions can also be seen. Some of the annual mean biases for the $PM_{2.5}$ chemical components noted in Table 5 are also reflected in almost all seasons in Fig. 15, including the underpredictions of $PM_{2.5}$-SO_4 and NO_3 and overpredictions of $PM_{2.5}$-EC and SS. In fact the reduction of bias for predictions of $PM_{2.5}$ total mass associated with the summation of these components with their individual underpredictions and overpredictions is an example of the positive impact that compensating errors can have on model skill, and it demonstrates the value of a more comprehensive evaluation of model performance, in this case the prediction of $PM_{2.5}$ chemical components in addition to $PM_{2.5}$ total mass.

Since Fig. 15 is based on measurements from sampling sites located mainly over the continental U.S. (see Fig. S6a), it does not account for $PM_{2.5}$ composition over northern Mexico and most of Canada. Figure 16, by contrast, shows a monthly time series of the predicted mean $PM_{2.5}$ chemical composition averaged over the land portion of the domain and the 2013–2016 simulations, with $PM_{2.5}$-TOM separated into POM and SOM. Seasonal variations of $PM_{2.5}$-SO_4, NO_3, POM, SOM, and SS can be clearly seen whereas seasonal variations of $PM_{2.5}$-NH_4, EC, and CM are less pronounced. $PM_{2.5}$-NO_3 and POM are predicted to have winter maxima and summer minima whereas $PM_{2.5}$-SO_4 and SOM are predicted to have winter minima and summer maxima. Note that the total inorganic component (sum of $PM_{2.5}$-SO_4, NO_3, and NH_4) only has a relatively small monthly variation. The predicted peak monthly mean $PM_{2.5}$ total mass occurs in August, driven by monthly maximum values of $PM_{2.5}$-SO_4 and SOM. This is different from Fig. 15, where the highest $PM_{2.5}$ total mass was predicted to occur in the winter, but note that predicted $PM_{2.5}$ total mass in Fig. 15 is roughly 5 $\mu g \cdot m^{-3}$ vs. 1.5 $\mu g \cdot m^{-3}$ in Fig. 16, suggesting that the former overweights the influence of urban areas. One other interesting feature in Fig. 16 is the SS maximum in August, which is in apparent contradiction to the SS wintertime maximum evident in Fig. S30. However, while Fig. S30 showed that inland penetration of sea salt over North America is limited, some seasonal variations are evident near Florida and the U.S. Gulf coast that may be associated with the occurrence of sea-land breezes in the warm season (and observed seasonal-mean SS values in Table 5 are largest in the spring). Lastly, Fig. S203 presents a similar analysis to Fig. 16 but for averaging over the full domain. The SS component clearly dominates $PM_{2.5}$ bulk mass in this figure, unlike Fig. 16, reflecting predicted high levels of sea salt over the Pacific and Atlantic Oceans.

It is also informative to look at the diurnal variation of the $PM_{2.5}$ chemical components. Figure 17 shows the predicted diurnal variation of eight $PM_{2.5}$ chemical components for each season after averaging over the 2013–2016 simulations and all North American continental grid cells. It is clear from this figure that both $PM_{2.5}$ total mass and chemical composition are predicted to vary with time of day. $PM_{2.5}$ total mass has a maximum for three seasons at 12 UTC (=7 EST), near sunrise and morning rush hour, and a minimum in all seasons at 21 UTC (=16 EST). The wintertime

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

maximum, on the other hand, occurs at 04 UTC (i.e., near local midnight), pointing to a different balance between surface emissions and vertical stability. Note too that the individual $PM_{2.5}$ chemical components display different diurnal behaviours. Hourly $PM_{2.5}$-SO_4 concentration is predicted to be effectively constant, consistent with a nonvolatile, secondary regional pollutant. Hourly $PM_{2.5}$-NO_3 and NH_4 concentrations, by contrast, are lowest in the afternoon when near-surface temperature is highest, and highest at night, especially before sunrise when near-surface temperature is lowest. This behaviour is consistent with the semi-volatile nature of ammonium nitrate, for which lower temperatures favour the particle phase (e.g., Malm et al., 2004; Yu et al., 2005). Hourly $PM_{2.5}$-EC, POM, and CM concentrations are also predicted to be highest during the night and lowest in the afternoon, but this is likely due to greater vertical mixing of local emissions during the day, which reduces the near-surface buildup of these three primary species. Hourly $PM_{2.5}$-SOM, which is assumed to be nonvolatile in the RAQDPS023 (see Moran et al., 2025), behaves like $PM_{2.5}$-SO_4 and displays little diurnal variation. Finally, sea-salt concentrations tend to be higher at night and lower during the day, likely due to diurnal variations in surface wind speed and hence in sea-salt emissions.

Note that the daily measurements made by the $PM_{2.5}$ speciation networks are not able to confirm these diurnal variations of $PM_{2.5}$ chemical components predicted by the model. However, the all-station, annual-mean diurnal analyses of observed and predicted hourly $PM_{2.5}$ total mass shown in Fig. 11 for North America for 2021/22, in Fig. S167 for four seasons, and in Fig. S168 for four sub-continental regions, all suggest that measured diurnal variations were smaller than the predicted variations. Two contributing factors to this difference might be the diurnal allocation of primary $PM_{2.5}$ emissions used by the RAQDPS023 and the parameterization of $PM_{2.5}$ chemical volatility, including ammonium, nitrate, and water components. Interestingly, Fig. S167 shows that the observed diurnal variation was largest in the winter and smallest in the summer, which is consistent with Fig. 17. In addition, both the observed and predicted all-station seasonal-mean diurnal curves of $PM_{2.5}$ total mass in Fig. S167 have one peak at about the time of morning rush hour and sunrise and a second peak at about the time of evening rush hour and sunset. By contrast the seasonal continent-wide diurnal time series in Fig. 17 only have one peak, in the morning near sunrise, but the majority of grid cells sampled for this figure will contain little vehicular activity, unlike the urban areas in which many monitors are located.

More evaluation results for the daily $PM_{2.5}$ speciation measurements and gravimetric $PM_{2.5}$ total mass measurements for 2013–2016 can be found in Sect. S3.3.2. These results include tables of annual and seasonal scores for the individual CSN, IMPROVE, and NAPS networks as well as regional scores, spatial plots of annual MB, NMB, CRMSE, and R station scores for each $PM_{2.5}$ chemical component, spatial plots of predicted seasonal mean $PM_{2.5}$ component concentration fields (cf. Fig. 14), monthly time series of $PM_{2.5}$ component statistics, monthly density scatterplots, and additional stacked bar graphs stratified by network and by region. The discussion of Table 5 noted consistent annual underpredictions or overpredictions for some of the seven $PM_{2.5}$ chemical components for the 2013–2016 hindcasts. Similarly consistent biases were found throughout the year by season or month for the combined networks for three $PM_{2.5}$ components, namely underpredictions for $PM_{2.5}$-SO_4 (Fig. S151) and overpredictions for $PM_{2.5}$-EC (Fig. S154) and $PM_{2.5}$-SS (Fig. S157). In addition, $PM_{2.5}$-TOM, the component found to have the smallest annual bias, was shown to have pronounced seasonal biases, with marked overpredictions in winter and underpredictions in summer (Fig. S155). Nevertheless, performance benchmarks for five $PM_{2.5}$ chemical components (SO_4, NO_3, NH_4, EC, TOM) proposed by Emery et al. (2017) were also compared with both annual and seasonal scores. Nearly all $PM_{2.5}$-SO_4, NO_3, NH_4, and TOM annual NMB scores and many seasonal scores met acceptable NMB benchmarks, all $PM_{2.5}$-SO_4 and NO_3 annual and most seasonal NME scores met acceptable NME benchmarks, and all $PM_{2.5}$-SO_4, NO_3, NH_4, and EC annual and seasonal R scores met acceptable R benchmarks. Some biases were also shown to be network-dependent. $PM_{2.5}$-SO_4 was underpredicted and $PM_{2.5}$-SS was overpredicted throughout the year for all three speciation networks, but $PM_{2.5}$-EC was overpredicted for the urban-focused CSN and NAPS networks in all seasons but not for the rural-focused IMPROVE network (Table S5S). Similarly, annual NMB values for $PM_{2.5}$-TOM ranged from 0.20 to 0.31 for CSN and from 0.52 to 0.80 for NAPS vs. -0.46 to -0.30 for IMPROVE (Table S5A). And for $PM_{2.5}$-CM the network-dependent biases were even more pronounced: annual NMB values ranged from 1.23 to 1.59 for CSN and from 1.96 to 2.34 for NAPS vs. -0.59 to -0.50 for IMPROVE. $PM_{2.5}$-EC, POM, and CM are all primary pollutants, suggesting that some of these model biases may be connected to the representation of primary $PM_{2.5}$ emissions.

The systematic biases by network that were identified for some $PM_{2.5}$ chemical components also affect predictions of $PM_{2.5}$ composition and total mass. The discussion of Fig. 15 noted that the best agreement was for spring and autumn. However, when the same analysis was applied to only CSN measurements and to only IMPROVE measurements, the agreement was less good for these two seasons, with the model overpredicting $PM_{2.5}$ total mass for CSN (Fig. S200) and underpredicting $PM_{2.5}$ total mass for IMPROVE (Fig. S201). The findings were similar for a regional analysis, where predicted $PM_{2.5}$ composition and total mass were in very good agreement with the combined measurements for WUS and EUS (Fig. S205), but the same regional analysis for CSN-only measurements (Fig. S206) and IMPROVE-only measurements (Fig. S207) revealed errors of opposite sign. The presence of these compensating errors for urban-focused and rural-focused stations again points to the benefits of a more disaggregated analysis. They also agree with the finding in Sect. 3.2.2 that $PM_{2.5}$ total mass underprediction appears to be mainly a rural issue. Predicted peak seasonal concentrations for $PM_{2.5}$-NO_3, NH_4, POM, and SS occurred in the winter (Figs. S23, S24, S26, and S30) vs. summer peaks for $PM_{2.5}$-SO_4, SOM, and TOM (Figs. S22, S27, and S28). This complementarity can help to explain the observed bimodality in monthly $PM_{2.5}$ total mass (Fig. S198). Close links are also evident between the monthly mean concentration time series for $PM_{2.5}$-SO_4 (Fig. S151) and SO_2 (Fig. S146), which were seasonally anticorrelated. $PM_{2.5}$-SO_4 had an observed and predicted summer maximum and winter minimum and was underpredicted while SO_2, its precursor, had an observed and predicted winter maximum and summer minimum and was overpredicted. A similar anticorrelation was evident for $PM_{2.5}$-NO_3 (Fig. S152) and HNO_3 (Fig. S144), which were also closely related. Lastly, downward time trends for 2013-2016 were also visible in observed monthly surface concentrations of $PM_{2.5}$-SO_4 and NH_4 (Figs. S151, S153) due to the SO_2 and NO_x emissions decreases that took place over this period.

3.3.3 Precipitation chemistry

Precipitation-chemistry measurements are valuable for model evaluation because they are quite different from the air-chemistry measurements considered in the previous sections. One difference is that they quantify a removal process, wet deposition, as opposed to near-surface ambient concentrations determined by dispersion and chemistry. In addition, they represent removal through a column extending from the Earth's surface to cloud top rather than a pure surface-level quantity. They are also an integrated measurement in terms of both gas-particle partitioning and aerosol particle size distribution because they include contributions from both gas and particle phases and from all aerosol particle sizes. Precipitation-chemistry measurements can thus provide some important insights into sulfur, oxidized nitrogen, reduced nitrogen, and base cation removal processes and atmospheric mass budgets. As described in Sect. 2.3, two large national networks are responsible for North American precipitation-chemistry measurements, CAPMoN in Canada and NADP in the U.S.

Figure 18 shows the predicted spatial distributions of annual mean concentration in precipitation over North America of the three major inorganic ions, $SO_4^{=}$, NO_3^-, and NH_4^+, for 2013–2016 and 2021/22. Several features are evident. First, the spatial distributions are different for each of these ions, largely in response to the different spatial distributions of their respective precursor emissions (cf. Fig. S1). Annual mean $SO_4^{=}$ concentration in precipitation tends to be larger over the eastern U.S. whereas annual mean NO_3^- and NH_4^+ concentrations in precipitation are larger over the western and central U.S. Another feature is that predicted $SO_4^{=}$ and NO_3^- annual mean concentrations in precipitation appear to decrease from 2013 to 2016, consistent with the 37% and 18% decreases in North American SO_2 and NO_x emissions, respectively, over this period (Table 1). No time trend is evident, however, for annual mean NH_4^+ concentration in precipitation, which differs from the downward trend for annual mean $PM_{2.5}$-NH_4 concentration visible in Fig. 14. Two reasons are that NH_3 emissions only decreased slightly (7%) from 2013 to 2016 and both NH_3 and PM-NH_4 are removed by precipitation scavenging regardless of the exact balance for gas-particle partitioning while NH_3 VMR displayed a slight upward trend from 2013 to 2016 (cf. Table 4, Fig. S16). Figure 18 also agrees qualitatively with comparable plots of observed annual mean $SO_4^{=}$, NO_3^-, and NH_4^+ concentrations in precipitation for these four years produced by the U.S. NADP network (e.g., National Acid Deposition Program, 2014, 2017), except for elevated values of annual mean NO_3^- and NH_4^+ concentrations in precipitation predicted over California in 2013. Note too that the corresponding annual wet deposition fields predicted by the RAQDPS023 for 2014–2016 (see Figs. S34–S36) have also been used in a study by Cathcart et al. (2025) to calculate total deposition and critical-load exceedance fields for Canada.

Table 6 lists all-station annual evaluation statistics for 2013–2016 for predicted weekly concentrations in precipitation of $SO_4^{=}$, NO_3^-, and NH_4^+ and for the corresponding predicted weekly wet deposition of these ions. The reduction in concentrations in precipitation of $SO_4^{=}$ and NO_3^- from 2013 to 2016 suggested visually by Fig. 18 is confirmed by Table 6. Observed and predicted annual mean $SO_4^{=}$ weekly concentration in precipitation for the combined networks decreased from 0.81 to 0.62 mg·L^{-1} and from 0.82 to 0.61 mg·L^{-1}, respectively, over this period, while observed and predicted annual mean $SO_4^{=}$ weekly wet deposition decreased from 15.6 to 10.7 mg·m^{-2} and from 15.9 to 10.1 mg·m^{-}

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

^{2}, respectively. Observed and measured annual mean NO_3^- weekly concentration in precipitation and wet deposition also declined from 2013 to 2016, though by smaller percentages, from 0.96 to 0.90 mg·L^{-1} and 0.94 to 0.78 mg· L^{-1} and from 16.5 to 14.1 mg·m^{-2} and 16.2 to 11.6 mg·m^{-2}, respectively. By contrast annual mean NH_4^+ weekly concentration in precipitation increased slightly while annual mean NH_4^+ weekly wet deposition decreased slightly.

Table 6 also reveals that model performance can vary considerably by ionic species. For example, all-station annual NMB values for $SO_4^=$ and NO_3^- weekly concentrations in precipitation ranged from -0.01 to 0.14 and from -0.14 to -0.01, respectively, but corresponding all-station annual NMB values for NH_4^+ weekly concentration in precipitation ranged from 0.23 to 0.41, pointing to a considerable and consistent overprediction for this quantity. All-station annual NMB scores for $SO_4^=$, NO_3^-, and NH_4^+ weekly wet deposition ranged from -0.06 to 0.05, -0.18 to -0.02, and 0.13 to 0.16, respectively, suggesting small biases for $SO_4^=$ wet deposition, a small but consistent underprediction for NO_3^- wet deposition, and a consistent overprediction for NH_4^+ wet deposition. In comparison, Simon et al. (2012) reported (10%, 90%) quantile values for NMB scores for accumulated seasonal and annual $SO_4^=$, NO_3^-, and NH_4^+ wet deposition of (-0.09, 0.38), (-0.45, 0.19), and (-0.33, 0.28), respectively, in a review of North American AQ model performance evaluations. The RAQDPS023 annual NMB scores fit comfortably within these reported ranges despite their much lower level of temporal aggregation (i.e., weekly vs. annual). Zhang et al. (2018c) reported large negative annual NMB values for $SO_4^=$, NO_3^-, and NH_4^+ accumulated annual wet deposition of -0.05, -0.32, and -0.31, respectively, for the 1990–2010 period using the CMAQ model without post-processing adjustments, and Benish et al. (2022) obtained annual NMB values for $SO_4^=$, NO_3^-, and NH_4^+ accumulated annual wet deposition of -0.12, -0.10, and -0.20 for the 2002–2017 period using a newer version of the CMAQ model and also without post-processing adjustments. Note that the negative bias in NO_3^- concentration in precipitation and wet deposition is consistent with the neglect of lightning NO emissions in both models (Zhang et al., 2018c).

Corresponding all-station annual NME values from Table 6 for $SO_4^=$, NO_3^-, and NH_4^+ weekly concentrations in precipitation for 2013–2016 ranged from 0.59 to 0.68, 0.52 to 0.57, and 0.78 to 0.95, respectively, and annual NME values for $SO_4^=$, NO_3^-, and NH_4^+ weekly wet deposition ranged from 0.61 to 0.65, 0.52 to 0.55, and 0.68 to 0.72. The NME scores for weekly wet deposition compare favourably with top-quartile NME scores for accumulated seasonal and annual $SO_4^=$, NO_3^-, and NH_4^+ wet deposition of 0.61, 0.51, and 0.57 compiled by Simon et al. (2012). All-station annual FAC2 values for weekly concentration in precipitation were slightly higher (better) for NO_3^- (0.68–0.70) than for $SO_4^=$ or NH_4^+ (0.63–0.65; 0.57–0.60), and the same was true for weekly wet deposition (0.62–0.64 vs. 0.53–0.56 and 0.55–0.57). All-station annual R scores for all stations were also slightly higher for NO_3^- weekly concentration in precipitation (0.46-0.52) than for $SO_4^=$ or NH_4^+ weekly concentration in precipitation (0.25–0.41, 0.38–0.47), whereas all-station annual R scores for $SO_4^=$, NO_3^-, and NH_4^+ weekly wet deposition were comparable (0.49–0.58, 0.46–0.59, 0.47–0.59). By comparison, Zhang et al. (2018c) obtained higher annual R scores for accumulated $SO_4^=$, NO_3^-, and NH_4^+ annual wet deposition for the 1990–2010 period of 0.92, 0.89, and 0.77, respectively, and Benish et al. (2022) obtained annual R scores for accumulated $SO_4^=$, NO_3^-, and NH_4^+ annual wet deposition for the 2002–2017 period of 0.88, 0.88, and 0.78, but these higher scores may again be explained by greater temporal aggregation. Note also that the values of the statistics in Table 6 were fairly constant across the four years in spite of the large SO_2 and NO_x emission reductions, suggesting that the year-specific input emissions that were used for the retrospective simulations were representative of each year.

More evaluation results for predicted $SO_4^=$, NO_3^-, and NH_4^+ weekly concentrations in precipitation and wet deposition can be found in Sect. S3.3.3. These include tables of separate annual and seasonal scores for the CAPMoN and NADP networks for 2013–2016 as well as regional scores, spatial plots of annual MB, NMB, CRMSE, and R scores at individual stations, spatial plots of predicted seasonal concentration in precipitation and wet deposition fields for 2013–2016 and 2021/22, monthly time series of statistics for weekly concentrations in precipitation and wet deposition, and monthly density scatterplots. A number of additional insights can be found in these supplemental analyses. For example, the separate annual and seasonal scores for the CAPMoN and NADP networks presented in Tables S6A and S6S revealed some differences that tended to favour CAPMoN. For example, annual R scores were consistently higher for CAPMoN than NADP for $SO_4^=$, NO_3^-, and NH_4^+ weekly concentrations in precipitation and weekly wet deposition. Some regional differences were also evident in many of spatial distributions of station-specific annual values of MB, NMB, CRMSE, and R. For example, the station-level annual MB scores for $SO_4^=$ weekly concentration in precipitation tended to be positive in eastern Canada and the northeastern U.S. but negative elsewhere (Fig. S109). Annual R values for $SO_4^=$ and NO_3^- weekly wet deposition also tended to be lower in the west and higher in the east (Figs. S128 and S132). It is also clear from Table S6S and Figs. S158-S163 that seasonal model skill was consistent overall from year to year but sometimes varied markedly between seasons and species. For example, NH_4^+ weekly wet deposition had the largest variations in observed and predicted seasonal mean values of the three major ions, with maximum seasonal values two to three times higher than minimum seasonal values. Scores for $SO_4^=$ weekly concentration in precipitation and weekly wet deposition were worse overall in the winter whereas scores for NO_3^- and NH_4^+ weekly concentration in precipitation and weekly wet deposition were worse overall in the summer. Predicted monthly mean concentrations in precipitation for $SO_4^=$, NO_3, and NH_4^+ all had both negative and positive biases, but the majority of monthly NMB values for $SO_4^=$ and NO_3^- weekly concentrations in precipitation fell in the ±0.20 range for 2013–2016 (Figs. S158-S159) whereas monthly NMB values for weekly NH_4^+ concentration in precipitation had peak values from 0.70 to 1.70 in July (Fig. S160). Lower monthly NMB values were found for wet deposition. The predicted seasonal mean precipitation fields were shown to link but geographically shift seasonal concentration in precipitation fields and the seasonal wet deposition fields. Lastly, observed and predicted monthly mean values of $SO_4^=$ weekly concentration in precipitation and $SO_4^=$ weekly wet deposition all declined from 2013 to 2016 (Figs. S158 and S162), providing observational support for the specified decreases in SO_2 annual emissions used for the hindcast annual simulations.

4 Discussion

4.1 Comparison with RAQDPS forecast performance for 2010–2019

The RAQDPS operational AQ forecast system underwent 22 upgrades between 2010 and 2021 (Moran et al., 2025). These included eight major upgrades: four that implemented significant changes to the modelling system code and configuration, two that involved major changes to the input emissions files, and two that included changes to both (Table A3). RAQDPS operational performance was routinely monitored over this period using hourly surface measurements of a small number of chemical species, in particular NO_2, O_3, and $PM_{2.5}$, from North American air-chemistry stations that report their measurements in near real time to AirNow or NAPS (Sect. 2.3). In order to assess the impact of this succession of system upgrades on forecast performance, Moran et al. (2021a) used hourly operational forecast outputs and NRT hourly AQ measurements archived at CMC to evaluate 9.5 years of RAQDPS operational forecasts of hourly surface NO_2, O_3, and $PM_{2.5}$, starting from RAQDPS001 forecasts at the beginning of 2010 through to July 2019, when the RAQDPS021 became operational (and before 2020 COVID-19 pandemic impacts: see Mashayekhi et al., 2021). This analysis aligned with suggestions made by Kelly et al. (2019) by providing as much consistency as possible in the calculation of evaluation statistics, in the model domain and grid resolution, and in the weighting of different time periods.

While some data filtering is performed routinely on the NRT measurements received at CMC before use (Sect. 2.3), additional filtering was applied to this 2010–2019 AQ measurement data set (Moran et al., 2021a). More stringent lower and upper cutoff thresholds of 0 and 150 ppbv, 0 and 150 ppbv, and 0 and 200 $\mu g \cdot m^{-3}$ were imposed on the NO_2, O_3, and $PM_{2.5}$ hourly observations, respectively. This additional filtering removed less than 0.1% of the available hourly measurements but avoided the impacts on the evaluation statistics of some extreme outliers. Measurement-model pairing was then performed for this 9.5-year period (cf. Sect. 2.4) followed by the calculation of seasonal performance statistics. Seasonal performance statistics for the 2013–2016 annual hindcast runs and 2021/22 forecast runs made with the RAQDPS023 (Table S3S) could then be compared with these earlier results to examine the impact of using the new version of the forecast system along with year-specific emissions for 2013–2016 and projected emissions for 2021/22.

Figure 19 shows time series of all-station seasonal R scores for NO_2, O_3, and $PM_{2.5}$ for the 2010–2019 period. The time series of 2010–2019 seasonal R scores for NO_2 and O_3 shown in this figure are broadly comparable: both time series fall in a numerical band from 0.50 to 0.70, both exhibit cyclical seasonal variations, and both exhibit some overall improvement with time. These positive trends were anticipated since proposed RAQDPS upgrades are only accepted for operational implementation if they can demonstrate at least equal or better expected forecast skill relative to the existing operational version (Moran et al., 2025). There is also a suggestion of some anticorrelation to the seasonal R scores for NO_2 and O_3, with R scores for NO_2 tending to be highest in the winter and lowest in the summer whereas R scores for O_3 tended to be highest in the summer but lowest in the spring. The 2010–2019 seasonal R scores for $PM_{2.5}$, on the other hand, are lower than those for NO_2 and O_3, are less cyclical seasonally, and do not show any improvement with time.

Figure 19 also shows seasonal R scores for the five RAQDPS023 annual simulations for 2013–2016 and 2021/22 that are the focus of this paper. The RAQDPS023 seasonal R scores for 2013–2016 were higher for both NO_2 and O_3 but lower for $PM_{2.5}$ compared to the previous operational RAQDPS versions. RAQDPS023 scores for 2021/22 were slightly lower for NO_2, comparable for O_3, and higher for $PM_{2.5}$ compared to the hindcast scores. These differences could be due either to changes to the treatments of atmospheric chemistry in GEM-MACH or to the use of more representative emissions in the RAQDPS023 simulations, but neither change was intended to produce poorer $PM_{2.5}$ forecasts. Each panel of Fig. 19 also shows species-specific values of "acceptable" and "good" benchmarks for R scores recommended by Emery et al. (2017) and Zhai et al. (2024). Seasonal R scores for NO_2 surpassed the more stringent "good" benchmark of 0.60 for later versions of the RAQDPS, seasonal R scores for O_3 for the RAQDPS023 surpassed the "acceptable" threshold of 0.50 but fall below the more stringent benchmark of 0.75, but seasonal R scores for $PM_{2.5}$ did not meet even the less stringent "acceptable" threshold of 0.40 for most forecast system versions or time periods.

Seasonal time series plots of four other statistics in the same format as Fig. 19 are shown in Figs. S215 to S219 for NO_2, O_3, and $PM_{2.5}$. It was found that RAQDPS023 seasonal MB and NMB scores for 2013–2016 were better than the historical operational scores for NO_2, slightly worse for O_3, and worse for $PM_{2.5}$, NME scores were better for NO_2 and O_3 and neutral for $PM_{2.5}$, RMSE scores were better for NO_2 and O_3 and worse for $PM_{2.5}$, and FAC2 scores were better for NO_2, neutral for O_3, and worse for $PM_{2.5}$. The RAQDPS023 seasonal MB and NMB scores for 2021/22 compared to 2013–2016 were worse for NO_2 and $PM_{2.5}$ but slightly better for O_3, NME scores were slightly better for all three species, RMSE scores were comparable for all three species, and FAC2 scores were worse for NO_2 and comparable for O_3 and $PM_{2.5}$.

4.2 Impact of biomass burning emissions

As noted in Sect. 2.2, BB emissions were included in the 2021/22 RAQDPS-FW023 forecasts but not in the 2021/22 RAQDPS023 forecasts or the four annual hindcasts considered in this paper. Such episodic but often large emissions can have a significant impact on both local and regional air quality (e.g., Jaffe et al., 2004; Liu et al., 2015; Rappold et al., 2017; Hand et al., 2024), and discussions in Sect. 3 of a number of figures (Figs. 4, 5, 8, 12) noted a drop in RAQDPS023 skill for predicting $PM_{2.5}$ concentration in the summer, the time of year when BB emissions are generally highest (e.g., Munoz-Alpizar et al., 2017; Chen et al., 2019). Table A4 summarizes annual wildfire statistics for Canada and the U.S. for the 2013–2022 period. Large year-to-year variations can be seen for both countries in the number of wildfires and the land area burned each year. At the continental scale, 2015, 2017, and 2021 stand out as high wildfire years in terms of land area burned whereas 2016, 2019, and 2020 were lower wildfire years. The 2021/22 RAQDPS023 forecasts thus corresponded to a high BB emissions year in 2021, in fact a record-breaking year (Jain et al., 2024). Note, however, that years with high levels of land area burned in the U.S. (e.g., 2015, 2017, 2018, 2020) may have a greater impact in terms of overall North American population exposure to wildfire smoke since many Canadian wildfires occur far from population centres. Note too that summer 2015 was identified as an outlier in the discussion of Fig. S201 in Sect. S3.3.2 due to the elevated $PM_{2.5}$-EC and TOM concentrations that were observed.

To assess the impact of the inclusion of BB emissions on forecast scores we can compare the RAQDPS023 and RAQDPS-FW023 forecasts for the 12-month 2021/22 period. The only difference between these two forecast system versions was the inclusion of BB emissions in the RAQDPS-FW023 runs. Table 7 compares seasonal evaluation statistics for hourly NO_2, O_3, and $PM_{2.5}$ forecasts for the two versions. For NO_2 only the summer statistics were slightly different (e.g., MB, FAC2). For O_3 there were slight differences in a few scores for the winter and spring and larger differences in the summer and autumn, including a reduction in seasonal MB in the summer of 0.8 ppbv for the RAQDPS-FW023. Differences in O_3 seasonal CRMSE, FAC2, and R scores in the summer and autumn were also small improvements. And for $PM_{2.5}$ there were larger differences in all four seasons, especially in summer and autumn. The summer mean $PM_{2.5}$ concentration predicted by the RAQDPS-FW023 was slightly more than double that of the RAQDPS023, which improved the summer NMB score from -0.51 to 0.04. Summer NME, FAC2, and R scores were also considerably better for the wildfire version (R increased from 0.09 to 0.55), although the RMSE and CRMSE values for the wildfire version were worse and the NSD value is now too high (1.87) vs. too low (0.35). In addition, for the $PM_{2.5}$ benchmarks recommended by Emery et al. (2017) for acceptable performance (NMB=±0.30, R=0.40), seasonal scores for the RAQDPS023 did not meet the thresholds for summer NMB or R and autumn R but RAQDPS-FW023 scores did meet these thresholds. These large improvements in most evaluation metrics suggest that BB is an emissions source that is important for part of the year, but at the same time there is a deterioration in a few scores, suggesting that there is room to improve the estimation of wildfire emissions.

Some additional analyses to understand the impact of BB emissions on 2021/22 RAQDPS-FW023 forecasts are presented in Sect. S4.2. These include plots of spatial distributions of annual and seasonal mean NO_2, O_3, and $PM_{2.5}$ abundance fields and station-specific annual statistics, time series of $\overline{O}$, median O, $\overline{M}$, median M, MB, NMB, CRMSE, FAC2, R, and NSD scores, and diurnal time series of five statistics by season and by region. These analyses confirmed that the inclusion of BB emissions improved some summer, autumn, and also annual scores for $PM_{2.5}$ in 2021/22 along with minor improvements for O_3 scores and no impact for NO_2 scores. Winter and spring scores, on the other hand, were virtually unchanged, which meant that $PM_{2.5}$ mass was underpredicted in these seasons by both system versions. In addition, Fig. S237 showed that the 2021 wildfire season was an outlier even relative to 2015. Monthly time series of predicted median $PM_{2.5}$ suggested that anthropogenic $PM_{2.5}$ emissions may have been too low in the cold-season months (Fig. S234). And regional analyses suggested that wildfire smoke affected all of North America in 2021/22 (Fig. S231 vs. Fig. 12).

4.3 Comparison with other AQ forecast systems

It is also of interest to compare RAQDPS023 forecast performance with that of peer AQ forecast systems operated by other agencies since such peer systems face similar constraints and limitations, including lack of access to year-specific input emissions and to meteorological and/or chemical data assimilation. To perform such a comparison, however, Simon et al. (2012) noted the challenge posed by the use of different evaluation metrics, sampling periods, sampling durations, measurement networks, and spatial domains in publications by different forecasting teams.

One such peer AQ forecast system is the U.S. National Air Quality Forecast Capability (NAQFC), which went operational in 2004 followed by many upgrades (see https://www.emc.ncep.noaa.gov/mmb/aq/AQChangelog.html; accessed 6 Aug. 2025). A number of papers that discuss NAQFC performance evaluations have been published, including Eder et al. (2006, 2009) Saylor and Stein (2012), Chai et al. (2013), Pan et al. (2014), Huang et al. (2017), Lee et al. (2017), Chen et al. (2021), Campbell et al. (2022), an Li et al. (2025). A limited number of evaluation results for surface $PM_{2.5}$ predictions from these papers, including those summarized in Table A5, can be used to compare NAQFC and RAQDPS performance over the parallel evolution of these two North American regional operational forecast systems. The overall conclusion based on the limited published comparisons is that performance was comparable for the two systems over the past decade (see Sect. S4.3 for details).

An ongoing North American AQ forecast intercomparison being conducted under the WMO GAFIS initiative (see Sect. S4.2) allows more recent NAQFC and RAQDPS023 performance to be compared directly for 2021/22 (and subsequent years). The 2022 second-quarter GAFIS report (Manseau et al., 2022) compared monthly MFB, FAC2, and R scores for daily maximum forecasts of NO_2, O_3, and $PM_{2.5}$ abundances for three regional AQ forecast systems, the RAQDPS023, RAQDPS-FW023, and NAQFC, and three global AQ forecast systems for the 12-month 2021/22 forecast period, with a focus on spring 2022. For the monthly mean diurnal time series of daily maximum surface concentrations of NO_2, O_3, and $PM_{2.5}$ presented in this report, there were marked variations evident between the six different AQ forecast systems. There was also considerable variation between the systems in monthly combined performance scores. Although no one system dominated, the NAQFC tended to be the top performer for O_3 forecasts, the RAQDPS023 and RAQDPS-FW023 for NO_2 forecasts, and IFS-CAMS for $PM_{2.5}$ forecasts, and the combined performance scores tended to be highest overall for O_3 forecasts and lowest for $PM_{2.5}$ forecasts. The monthly R scores can also be compared to available benchmarks. Most of the NO_2 R scores for five of the systems (NO_2 forecasts were not available for NAQFC) were above the acceptable benchmark of 0.50 recommended by Zhai et al. (2024) and some RAQDPS023, RAQDPS-FW023, and IFS-SILAM scores also exceeded their benchmark goal of 0.60. R scores for O_3 for all six systems exceeded the acceptable benchmark of 0.50 recommended by Emery et al. (2017) and in a few cases also exceeded their benchmark goal of 0.75. And for $PM_{2.5}$ R scores none of the systems attained the acceptable benchmark of 0.60 recommended by Huang et al. (2021), but in July and August 2021 the lower benchmark of 0.40 recommended by Emery et al. (2017) was reached by most of the models, including the RAQDPS-FW023 but not the RAQDPS023. **Section S4.3** also describes an ongoing evaluation and intercomparison of 11 operational regional AQ forecast models for Europe that is similar to the GAFIS regional intercomparison for North America and compares a few evaluation scores from this intercomparison with the RAQDPS023.

4.4 Forecast system shortcomings, opportunities, and priorities for further development

Many of the evaluation results for RAQDPS023 forecasts and hindcasts versus AQ measurements, previous system versions, and peer forecast systems discussed in previous sections were positive. For example, Figs. 6 and 7 (and S140 and S141) show model skill for predicted NO_2 and O_3 surface VMR monthly means at the continental scale as does Fig. 12 for four North American regions. Good agreement for all-station, annual-mean diurnal time series for NO_2 and O_3 is evident in Figs. 9 and 10 and for seasonal-mean diurnal time series in Figs. S165 and S166. And Fig. 15 shows good model performance in predicting all-station, seasonal-mean $PM_{2.5}$ chemical composition and gravimetric $PM_{2.5}$ total mass. Evaluation results presented in Sect. 3 also suggested that the year-specific 2013–2016 annual emissions that were used for the RAQDPS023 retrospective runs broadly represented emissions changes over that period (Table 1) given the year-ordered agreement between measurements and model predictions (e.g., Figs. S144, S146, S151, and S158 for HNO_3, SO_2, $PM_{2.5}$-SO_4, and $SO_4^=$ concentration in precipitation, respectively). Results presented in Sect. 4.1 showed some improvements in forecast skill over a 10-year period due to a series of operational RAQDPS upgrades, including increasing seasonal R scores for hourly NO_2 and O_3 forecasts (Fig. 19), decreasing seasonal NME scores for hourly O_3 and $PM_{2.5}$ forecasts (Fig. S217), decreasing seasonal RMSE scores for hourly $PM_{2.5}$ forecasts (Fig. S218), and increasing seasonal FAC2 scores for hourly NO_2 and O_3 forecasts (Fig. S219). Relative to suggested benchmark values, seasonal NMB scores for NO_2, O_3, and $PM_{2.5}$ for most RAQDPS operational versions met the acceptable benchmark (Fig. S216) as did seasonal R scores for NO_2 and O_3 (Fig. 19). Lastly, the ongoing GAFIS quarterly evaluation of RAQDPS and RAQDPS-FW forecasts for North America along with those made by four operational AQ forecast peer models described in Sect. 4.3 found that RAQDPS and RAQDPS-FW performance in 2021/22 was competitive with their peer models for surface NO_2 and O_3 VMRs, though less so for surface $PM_{2.5}$ total mass (for which all of the models performed least well).

It is also evident that many evaluation scores were similar for the five years simulated by the RAQDPS023, and some scores point to systematic errors in the model itself rather than model inputs since the anthropogenic input emissions used were tailored to be year-specific and the similarities in results are present despite interannual variations in meteorology. The most concerning systematic error may be consistent underpredictions of hourly $PM_{2.5}$ surface concentrations. These underpredictions were evident visually in Figs. 1 and 5, in which observed annual and seasonal $PM_{2.5}$ mean values at station locations stood out from the predicted mean $PM_{2.5}$ concentration fields across the continent due to their higher values, and in Table 3 where all-station annual NMB values for hourly $PM_{2.5}$ concentrations were negative for all five years. However, these underpredictions also varied strongly by season and were smallest in winter and largest in summer (Table S3S). In addition, Figure 11 showed a large underprediction of all-station annual-mean diurnal values for hourly $PM_{2.5}$ concentration at all hours but especially in early afternoon, and Fig. S167 showed corresponding underpredictions for seasonal-mean diurnal values, with the largest differences in summer.

Some other $PM_{2.5}$ evaluation results, however, tell a more nuanced story. First, the results just summarized are largely for all-station aggregated statistics. When monthly statistics were calculated separately for urban stations only and for rural stations only, as shown in Figs. S213 and S214, a key difference was that hourly $PM_{2.5}$ underpredictions for 2013–2016 were limited to rural stations in all months and to urban stations in the summer. For most of the year, however, hourly $PM_{2.5}$ mass was overpredicted at urban stations. Model predictions also showed better agreement with daily gravimetric $PM_{2.5}$ total mass measurements for 2013–2016 than with hourly continuous $PM_{2.5}$ measurements for those years (cf. Fig. S198 vs. Fig. S142). Results from the evaluation of $PM_{2.5}$ chemical composition and $PM_{2.5}$ reconstructed mass for 2013–2016 add further complications. Two $PM_{2.5}$ chemical components (EC, SS) were found to be consistently overpredicted in all months (Figs. S154, S157), one component (SO_4) was consistently underpredicted in all months (Fig. S151), and another component (CM) had large overpredictions for most months (Fig. S156). As a consequence, predicted annual mean $PM_{2.5}$ composition was incorrect on a rank-ordered basis since annual mean $PM_{2.5}$-

CM was predicted to be the second-largest $PM_{2.5}$ component after annual mean $PM_{2.5}$-TOM but was observed to be the fourth-largest annual mean $PM_{2.5}$ component after $PM_{2.5}$-TOM, SO_4, and NO_3 (Tables 5 and S5A). Despite these errors in predicting $PM_{2.5}$ composition, however, predictions of $PM_{2.5}$ reconstructed mass were surprisingly good because of partial cancellation of overpredictions and underpredictions of the mass of individual chemical components. In fact, Fig. 15 showed that observed seasonal reconstructed $PM_{2.5}$ dry mass for the combined CSN, IMPROVE, and NAPS $PM_{2.5}$ speciation networks for 2013–2016 was lower than predicted seasonal $PM_{2.5}$ dry mass for three of the four seasons in all years, summer being the exception. For the corresponding analysis based only on CSN measurements, however, observed seasonal reconstructed $PM_{2.5}$ dry mass was lower than predicted seasonal $PM_{2.5}$ dry mass for 15 out of 16 seasons (summer 2015, a high wildfire period, was the exception), with the largest model overpredictions occurring in the winter (Fig. S200). But for the same analysis based only on IMPROVE measurements, observed seasonal reconstructed $PM_{2.5}$ mass was higher than predicted seasonal $PM_{2.5}$ mass for all 16 seasons, with the largest model underpredictions occurring in the summer (Fig. S201). Since the CSN network consists largely of urban stations while the IMPROVE network consists largely of rural stations, these differences are consistent with Figs. S213 and S214.

Taken together the above findings offer five clues on possible improvements to RAQDPS023 hourly $PM_{2.5}$ predictions:
(i) model underpredictions were largest for the summer months and are consistent across multiple years;
(ii) underpredictions were larger in western North America than eastern North America;
(iii) underpredictions were strongly associated with rural stations while urban stations were sometimes associated with overpredictions;
(iv) predictions of some $PM_{2.5}$ chemical components had systematic biases, both negative and positive; and
(v) underpredictions were greater for the combined continuous hourly $PM_{2.5}$ measurement networks than the combined daily gravimetric $PM_{2.5}$ measurement networks.

The maximum summertime bias and western bias, which are both consistent with the timing and location of the majority of wildfires in North America (e.g., Holden et al., 2011; Mao et al., 2011; Hand et al., 2013; Schichtel et al., 2017), strongly suggest the importance of including wildfire emissions. However, Figs. S245 and S246 showed that while mean $PM_{2.5}$ underpredictions were much improved in the summer months at both urban and rural stations, this was not true for the rest of the year, especially at rural stations, so that other improvements are needed.

It is also clear that multiple actions will be required to address the shortcomings identified for predictions of the multiple chemical components that make up $PM_{2.5}$. For example, the overprediction of monthly mean $PM_{2.5}$ total mass in urban areas vs. underprediction in rural areas suggests that the spatial allocation of anthropogenic primary $PM_{2.5}$ emissions might need to be modified to re-allocate some of these emissions from urban to rural areas. This change might also help to address persistent $PM_{2.5}$-EC overpredictions (Fig. S154), whose emissions vary directly with population (e.g., Fig. 14). Another approach to reduce urban overpredictions might be to add an urban heat island mixing parameterization (e.g., Ren et al., 2020) or a SGS on-road mobile mixing parameterization (Makar et al., 2021). Figures 12 and S231 showed that $PM_{2.5}$ total mass was underpredicted in spring 2022 in the eastern U.S., where biogenic emissions are high (e.g., Fig. S21), and Fig. S207 showed that $PM_{2.5}$-TOM was underpredicted at IMPROVE measurement sites in the eastern U.S. from 2013 to 2016. These results suggest that biogenic SOA levels might be underpredicted, so the parameterization of biogenic SOA should be reviewed to see whether its contribution to $PM_{2.5}$-TOM in rural areas might be increased (e.g., Schichtel et al., 2017; Zhang et al., 2018a). Process representations related to $PM_{2.5}$-SO_4 production and $PM_{2.5}$-CM emissions should also be reviewed to see whether the contributions of these processes in rural areas could be increased (and the contribution of $PM_{2.5}$-CM emissions in urban areas reduced and the poor monthly representation of $PM_{2.5}$-CM concentration evident in Fig. S156 addressed). One possible way to increase rural $PM_{2.5}$-CM levels would be to add a parameterization for wind-blown dust emissions from natural sources, which is not part of the RAQDPS023 system, as another source of $PM_{2.5}$ emissions (e.g., Park et al., 2010; Appel et al., 2013; Foroutan et al., 2017). Another avenue to investigate given the overpredictions of $PM_{2.5}$-CM in the winter is the meteorological modulation scheme that was used by the RAQDPS023 to reduce fugitive dust emissions when the ground was predicted to be wet or snow-covered (Moran et al., 2025). In addition, the $PM_{2.5}$-SS component, which should be included in the calculation of predicted $PM_{2.5}$ total mass, was used in the calculation of predicted $PM_{2.5}$ total mass in Fig. 15, where it made a non-negligible contribution, but was not included in the disseminated RAQDPS023 operational forecasts of $PM_{2.5}$ concentration due to extreme $PM_{2.5}$-SS overpredictions that were encountered in the earliest RAQDPS versions. This omission should be addressed, but at the same time an effort should be made to reduce the remaining $PM_{2.5}$-SS overprediction (e.g., Table 5, Figs. S157, S105, S186) while keeping in mind the higher uncertainties of these scores due to the use of a measured proxy species to estimate $PM_{2.5}$-SS mass (see Sect. S2.4).

The RAQDPS023 predictions of hourly $PM_{2.5}$ total mass only considered $PM_{2.5}$ dry mass. The better agreement of predicted $PM_{2.5}$ total mass with gravimetric $PM_{2.5}$ total mass measurements, which are analyzed under low-humidity laboratory conditions and only contain particle-bound water, than with non-FRM continuous $PM_{2.5}$ total mass measurements, which are made regardless of humidity levels and include both include semi-volatile and particle-bound water, suggests that predicted hourly $PM_{2.5}$ total mass should include an aerosol water component (e.g., Frank, 2006; Malm et al., 2011; Nguyen et al., 2016; Pye et al., 2017; Widziewicz-Rzońca and Tytła, 2020). The gravimetric $PM_{2.5}$ analysis also results in the loss of some $PM_{2.5}$-NO_3 and NH_4 as well as some aerosol water (e.g., Frank, 2006; Malm et al., 2011; Chow et al., 2015; Nguyen et al., 2016; Hand et al., 2019). In addition, predicted $PM_{2.5}$-NO_3 is biased low when compared to laboratory-analyzed speciation measurements (e.g., Table S5S-mr, Fig. S152). The representation of inorganic heterogeneous chemistry used by the RAQDPS023 (Moran et al., 2025) should be examined critically. For example, one process missing from the RAQDPS023 inorganic hetereogeneous chemistry parameterization was an explicit treatment of the role of base cations (e.g., Miller et al., 2024).

One shortcoming revealed by the evaluation of O_3 forecasts was the underprediction of the well-known Northern Hemisphere spring O_3 peak (e.g., Penkett and Brice, 1986; Monks, 2000; Liudchik et al., 2015). All-station spring NMB values for O_3 surface VMR were the lowest of the four seasons for 2013–2016 and ranged from -0.16 to -0.20 (Table S3S). The largest negative all-station monthly NMB values for O_3 occurred in April and May in 2013–2016 (Fig. S141). Other examples of negative O_3 bias in spring and early summer months are evident in Table S7 and in Figs. 3, 12, S42, and S166. Different explanations for this systematic springtime and western underprediction are possible, but one obvious candidate for investigation is the model's treatment of O_3 chemical lateral boundary conditions (e.g., Pendlebury et al., 2018; Moran et al., 2025). Two more aspects of the O_3 predictions that could be improved are also worth noting. First, annual mean O_3 surface VMRs showed a latitudinal step jump over the southeastern U.S. for all five years (Fig. 13). This artefact was likely caused by the parameterization of O_3 dry deposition used by the RAQDPS023, which considered five coarse "seasons" that were defined based only on broad latitudinal bands with no dependence on longitude or elevation (e.g., Zhang et al., 2002; Makar et al., 2018; Moran et al., 2025). And second, Fig. S42 showed that many coastal stations along the U.S. Gulf of Mexico and Florida peninsula had positive annual NMB values for 2013–2016. These O_3 overpredictions at coastal stations are consistent with the model's lack of any treatment of marine halogen chemistry, which can reduce surface O_3 concentrations (e.g., Sarwar et al., 2015; Li et al., 2019).

The evaluation of predictions of other gas-phase species besides NO_2 and O_3 also pointed to shortcomings in modelling some $PM_{2.5}$ precursors. For example, annual-mean SO_2, HNO_3, and ISOP VMRs were consistently overpredicted and annual-mean NH_3 VMRs were underpredicted for 2013–2016 (Table 4). All-station seasonal-mean SO_2 VMRs were also overpredicted for all 16 seasons in 2013–2016 (Table S4S), which was consistent with the underprediction of mean $PM_{2.5}$-SO_4 air concentrations for the same seasons (Table S5S). Figure S146 showed the largest monthly NMB values to be associated with the winter and summer seasons. Table S8 showed that the highest annual-mean regional SO_2 VMRs were predicted to occur in eastern Canada whereas measurements put the highest annual-mean regional SO_2 VMRs in either the eastern U.S. or western Canada. The majority of measurement stations with annual SO_2 overpredictions were located in eastern North America or Alberta, whereas many stations in the western U.S. exhibited underpredictions (Figs. S61, S62). Two SO_2 removal processes were found to be missing from the RAQDPS023: the soil-wetness and cuticle-wetness gas-phase dry deposition pathways (Moran et al., 2025). Adding treatments for these two pathways is an obvious first step to reduce SO_2 overpredictions. Note that other gas-phase species such as HNO_3 and NH_3, which use SO_2 as an archetype for modelling dry deposition (Zhang et al., 2002), would also be impacted by this change: monthly mean HNO_3 VMR was also consistently overpredicted for all months (Fig. S144) although monthly mean NH_3 VMR was not (Fig. S145). In addition, the fact that monthly mean SO_2 VMR was consistently overpredicted while monthly mean $PM_{2.5}$-SO_4 air concentration was consistently underpredicted (Fig. S151) suggests that predicted SO_2-to-SO_4 conversion was too low. The two chemical pathways for SO_2-to-SO_4 conversion considered by the RAQDPS023 were gas-phase oxidation and aqueous-phase oxidation. It is shown in Fig. S158 that monthly NMB values for weekly $SO_4^{=}$ concentration in precipitation for 2013–2016 were positive for most of the year. This result makes it more likely that the RAQDPS023 representation of gas-phase oxidation of SO_2 may be the main reason for the year-round underprediction of $PM_{2.5}$-SO_4 air concentration.

All-station seasonal-mean NH_3 VMR was underpredicted for all seasons in 2013–2016, with the seasonal NMB values in the winter (Table S4S). In addition, Puchalski et al. (2011) found the Radiello passive samplers used by the AMoN network, which provided most NH_3 measurements, to be biased low, thus suggesting an even larger model underprediction. This consistent underprediction for NH_3 VMR occurred at the same time as overpredictions for most of the year of (i) $PM_{2.5}$-NH_4 air concentration, (ii) NH_4^+ concentration in precipitation, and (iii) NH_4^+ wet deposition. Table S8 showed that the highest annual mean NH_3 VMRs were predicted to occur in eastern Canada whereas the western U.S. had the highest observed annual mean VMRs. Figure S145 shows time series of monthly mean NH_3 VMRs and related evaluation statistics for 2013–2016, for which monthly NMB values were negative for all months and years but were largest in the winter. The accuracy of emissions is always a potential factor to explain either model over- or underpredictions, and the large negative NMB values for NH_3 VMR occurring in the west and in the winter could point to an underestimate of western and wintertime NH_3 emissions (e.g., Momeni et al., 2025). The level of autumn NH_3 emissions could also be underestimated. In addition, Figs. S160 and S164 show peak overpredictions of monthly NH_4^+ wet concentration and wet deposition in the summer at the same time as monthly ambient NH_3 was underpredicted (Fig. S145) but ambient $PM_{2.5}$-NH_4 was overpredicted (Fig. S153). This could mean that too much NH_3 gas was being removed in the model by wet deposition in the summer when inorganic aerosol thermodynamics favours ambient NH_3 over $PM_{2.5}$-NH_4 and ambient HNO_3 over $PM_{2.5}$-NO_3 (e.g., Ansari and Pandis, 1998).

Improvements to the prediction of HNO_3 and $PM_{2.5}$-NO_3 could also improve the prediction of $PM_{2.5}$-NH_4. For example, monthly NMB scores for $PM_{2.5}$-NH_4 peak in September and October (Fig. S153) when monthly NMB scores for HNO_3 also peak (Fig. S144), while monthly $PM_{2.5}$-NO_3 was underpredicted in all months except September and October (Fig. S152), and monthly NO_3^- wet concentration was also overpredicted in September and October (Fig. S159). HNO_3 precursors NO_2 and NO were overpredicted in almost all months (Figs. S140 and S143), so reducing their levels by (if justifiable) decreasing NO_x emissions would decrease HNO_3 levels. Increases to available NH_3 levels in the cold season through temporal reallocation of NH_3 emissions would result in increased $PM_{2.5}$-NO_3 levels and decreased HNO_3 levels in those months. And implementation of the missing dry deposition pathways for SO_2 to wet surfaces noted above will also increase HNO_3 dry deposition . It is clear, though, from this discussion how intertwined the sulfur, oxidized nitrogen, and reduced nitrogen budgets are via emissions, chemistry, gas-phase partitioning, and wet and dry removal.

Lastly, seasonal-mean ISOP was overpredicted for all seasons in the 2013–2016 period (Table S4S). Annual overpredictions also occurred at all available PAMS measurement stations (Figs. S77, S78). Figure S150 showed time series of monthly ISOP prediction errors, which included very high NMB scores, poor FAC2 values, and near-zero R scores in the cold season. Although the ISOP measurements considered here were obtained from a small number of U.S. stations, an earlier study by Stroud et al. (2008) that compared Canadian ISOP measurements against predictions by the AURAMS model, which used the same gas-phase chemistry mechanism and a similar treatment of biogenic emissions as the RAQDPS023, also found overpredictions, especially in eastern Canada. When considered together these findings suggest that further examination of the RAQDPS023 biogenic emissions scheme (Moran et al., 2025) is warranted for both magnitude and timing, including the spatiotemporal specification of vegetation phenology.

5 Summary and conclusions

This paper presents results from a comprehensive, five-year performance evaluation and analysis of both prospective and retrospective annual simulations made with version 023 of the ECCC Regional Air Quality Deterministic Prediction System (RAQDPS023), an operational online chemical weather forecast system for North America. A companion paper by Moran et al. (2025) provides a comprehensive and detailed description of this forecast system. The performance evaluation consists of three parts. In the first part, near-real-time (NRT) hourly measurements of three pollutant species, NO_2, O_3, and $PM_{2.5}$ total mass, were used to perform an operational evaluation of the first year of RAQDPS023 forecasts from July 2021 to June 2022. The Canadian Air Quality Health Index is based on these three species, so they are considered to be the key forecast species out of the dozens that were predicted by the RAQDPS023. The anthropogenic input emission files used for these 2021/22 forecasts were based on a projected 2020 Canadian national emission inventory and projected 2023 U.S. and Mexican national emission inventories (i.e., emission inventories forecasted from retrospective base-year inventories). The performance of the RAQDPS-FW023, a second version of the ECCC operational AQ forecast system that is a duplicate of the RAQDPS023 except for the addition of NRT biomass burning (BB) emissions, was also evaluated for 2021/22 using the same measurement data set.

The 2021/22 measurement data set spanned much of North America and included roughly 200 surface sites in Canada and 1100 sites in the U.S., although only one or two of the three pollutants were measured at some sites. Before being used for the evaluation, the AQ measurements underwent a two-step screening process. The first step was to apply validity checks to discard negative concentrations and above-threshold concentration values flagged as suspicious or invalid. Each valid measurement (or scaled or combination of measurements for some $PM_{2.5}$ chemical components) was then paired with a forecast value, after which the second screening step, period-specific completeness checking, was performed to ensure that at least a minimum number of valid hourly measurements were available at a site for the evaluation period being considered to ensure temporal representativeness. No calculation was performed if there were not enough valid measurements. Annual, seasonal, monthly, and hour-of-day values of 10 statistical metrics were then calculated for both individual measurement networks and combined networks, for both the entire continent and four continental quadrants, and for urban sites only and rural sites only. In addition to summary tables, some of these statistical scores were presented visually in multiple ways, including site-specific "dot" statistics maps, monthly and diurnal time series, and density scatterplots,.

In the second part of the performance evaluation, an expanded and more detailed analysis was performed on four annual hindcasts for 2013–2016 made with the RAQDPS024 (algorithmically equivalent to the RAQDPS023 but run on a newer computer system; see Moran et al., 2025). The anthropogenic input emission files used for each of these four annual simulations were year-specific since they were generated for each year based on multi-year retrospective data sets of Canadian, U.S., and Mexican annual emission inventories. The evaluation of retrospective simulations makes it possible to access a much larger set of AQ measurement data for North America, including more AQ measurement networks, more chemical species (23 vs. 3), and more measurement sites (roughly 2000 vs. 1300). These historical data sets have also undergone quality assurance and quality control checks by the individual networks before being released. The consideration of multiple simulation years reduces the confounding influence of interannual meteorological variability, which helps with the identification of systematic prediction errors. However, differences between measurement network infrastructure and procedures, in particular, sampling methodology and duration but also sampling frequency and instrument type, mean that pre-processing of both measurements and model predictions is often required before measured values and model predictions can be compared across networks. Most importantly, after station-level screening for validity and completeness, the measurements from some networks had to be averaged or accumulated in time to match the longest sampling duration employed by any network for that same species in order to calculate consistent all-station (i.e., multiple-network or combined) statistics. In a few cases, predicted model species also had to be summed before being paired with measurements.

Of the three key AQHI forecast species, evaluation scores for O_3 hourly forecasts made by the RAQDPS023 were generally the highest for all five years, followed by NO_2 scores and then $PM_{2.5}$ scores. The finding that $PM_{2.5}$ total mass was the most difficult AQHI pollutant to predict was not surprising given its inherent complexity as a hybrid, primary-secondary multi-pollutant that spans a wide size range and has numerous sources. Another important finding was that monthly mean hourly $PM_{2.5}$ total mass predicted by the RAQDPS023 was biased low in all months in 2021/22 and in most months, especially in the summer, in 2013–2016. Relative to the NMB, NME, and R "acceptability" benchmarks recommended by Zhai et al. (2024) for NO_2 VMR predictions, all-station seasonal RAQDPS023 predictions met the NMB benchmark for 19 out of 20 seasons for these five years, met the NME benchmark for the five winter seasons, and met the R benchmark for all 20 seasons. Similarly, for the NMB, NME, and R acceptability benchmarks recommended by Emery et al. (2017) for O_3 VMR predictions, all-station seasonal RAQDPS023 predictions met the NMB benchmark for 15 seasons (but none of the springs), met the NME benchmark for only one season, but met the R benchmark for all 20 seasons. And for the NMB, NME, and R acceptability benchmarks recommended by Emery et al. (2017) for $PM_{2.5}$ concentration predictions, all-station seasonal RAQDPS023 predictions met the NMB benchmark for 16 seasons (but not four summers), did not meet the NME benchmark for any season, but met the R benchmark for all 20 seasons. The ongoing WMO GAFIS multi-model comparison of operational AQ forecast systems for North America also found RAQDPS023 forecasts to be competitive with four peer forecast systems for NO_2 and O_3 for all months in 2021/22 and for $PM_{2.5}$ total mass for cold-season months.

Biomass burning is an important seasonal source of emissions of both primary $PM_{2.5}$ mass and its gas-phase precursors. A comparison of RAQDPS-FW023 and RAQDPS023 performance for 2021/22 found much improved $PM_{2.5}$ evaluation scores for the RAQDPS-FW023 for the summer months, when BB emissions peak. It was also shown that the inclusion of BB emissions that occur mainly in the summer and early autumn affected annual evaluation statistics significantly. This comparison has quantified the impact of BB emissions on AQ forecasts and has provided strong evidence for the importance of including BB emissions, which mainly affect $PM_{2.5}$ levels in the summer and, to a lesser degree, autumn. One further insight came from separate evaluations with hourly $PM_{2.5}$ measurements that had been divided into urban and rural subsets, which was that once BB emissions were included, remaining model underpredictions of hourly $PM_{2.5}$

concentration were found to occur mainly in rural areas from October to June, whereas model overpredictions of $PM_{2.5}$ concentration were sometimes a second issue in urban areas. Other explanations are thus still needed for hourly $PM_{2.5}$ underpredictions outside of the wildfire season, especially in rural areas and in the eastern U.S.

The evaluation of $PM_{2.5}$ predictions for the 2013–2016 annual hindcasts was expanded by considering two additional types of $PM_{2.5}$ measurements available from multiple networks, namely (i) daily gravimetric $PM_{2.5}$ total mass measurements and (ii) daily $PM_{2.5}$ chemical composition measurements. These additional $PM_{2.5}$ data sets augmented the ourly $PM_{2.5}$ total mass data sets and provided insights into causes of the hourly $PM_{2.5}$ mass underpredictions. One finding was that RAQDPS024 predictions of daily gravimetric $PM_{2.5}$ total mass were less negatively biased than those for hourly $PM_{2.5}$ mass, which draws attention to instrument and analysis differences between the two types of measurements. A second finding was that daily $PM_{2.5}$ mass reconstruction results based on observed daily $PM_{2.5}$ speciation measurements agreed well both with observed gravimetric $PM_{2.5}$ total mass measurements for 2013–2016 and with RAQDPS024 predictions of gravimetric $PM_{2.5}$ total mass, which was again surprising in view of the consistent model underpredictions of hourly $PM_{2.5}$ total mass. The RAQDPS024 calculation of hourly $PM_{2.5}$ total mass, however, did not include either aerosol water or SS components, but comparisons with observed gravimetric $PM_{2.5}$ total mass suggested that both chemical components should be included in the model calculation of $PM_{2.5}$ total mass. And while RAQDPS024 predictions of all-station $PM_{2.5}$ chemical composition were reasonably good, it was also shown that the good agreement between observed and predicted gravimetric $PM_{2.5}$ total mass was partly due to compensating model errors in the prediction of individual $PM_{2.5}$ chemical components because the model was found to overpredict EC and SS but underpredict SO_4 in all seasons and to overpredict TOM and CM at urban stations but underpredict these components at rural stations. Lastly, both observed and predicted seasonal $PM_{2.5}$ residual mass were found to be largest in the summer while annual $PM_{2.5}$ residual mass was largest in eastern North America. Some possible explanations include incorrect partitioning of primary $PM_{2.5}$ emissions between urban and rural areas and underpredictions of ammonium nitrate and biogenic SOA in the summer.

In the third part of the evaluation, trends in seasonal performance were shown for the first decade of operational forecasts by the GEM-MACH-based version of the RAQDPS from 2010 to 2019. Overall skill in predicting hourly NO_2 and O_3 VMRs improved modestly over this period due to a series of modelling system upgrades, including updated input emissions files (see Moran et al., 2025). Predictions of hourly $PM_{2.5}$ total mass, on the other hand, improved initially but then declined after 2017. These seasonal forecast scores for earlier operational RAQDPS versions were then compared to seasonal scores for the RAQDPS023 hindcast simulations for 2013–2016 and 2021/22 forecasts. RAQDPS023 seasonal scores were better overall for NO_2 and O_3 but showed little improvement for $PM_{2.5}$.

Meteorology and climate can also affect model performance through the direct influence of some meteorological variables, such as wind speed, PBL height, temperature, precipitation, solar radiation, and cloud cover, on pollutant abundances and removal, and indirectly through related factors such as vegetation phenology, snow cover, and emissions affected by meteorology. As a consequence, there were seasonal or diurnal cycles in objective scores for many pollutants that were as large or larger than year-to-year fluctuations or trends in model performance. These temporal variations in model performance can provide clues and guidance to identify those components of the modelling system where further improvements may be needed. Examples include underprediction of surface O_3 VMR in the spring, peak underprediction of NH_3 VMR in winter, and consistent underpredictions of $PM_{2.5}$-SO_4 for all seasons. Some model errors are also correlated in time for chemically coupled species, such as those for SO_2 and $PM_{2.5}$-SO_4, for HNO_3 and $PM_{2.5}$-NO_3, and for NH_3 and $PM_{2.5}$-NH_4, and may be in phase or out of phase.

One confounding factor faced by this study was the changing chemical environment in North America caused by significant changes in regional and national emissions that have occurred over the past two decades, including the four-year period from 2013 to 2016 and the six-year period from 2016 to 2021. For example, the impact of large monotonic reductions in Canadian and U.S. SO_2 annual emissions of 16% and 50%, respectively, from 2013 to 2016 can be seen in both observed and predicted monthly time series of SO_2 VMR, $PM_{2.5}$-SO_4 concentration, and $SO_4^=$ concentration in precipitation and wet deposition. This agreement in temporal trends between measurements and model predictions affirms the representativeness of the retrospective, multi-year emission inventory data sets used for this study. However, having represented these emissions changes in the 2013–2016 model simulations, monthly and seasonal scores were found to vary in a similar manner from year to year, which points to systematic errors in the model itself (although 2016 seemed to be an outlier for some species). In addition, these results clearly show the value of using year-specific emissions, although this is not possible in the case of AQ forecast models because current emissions cannot be known beforehand.

This study has demonstrated the value of a comprehensive, quasi-diagnostic model performance evaluation. Consideration of a wide range of pollutant species allows some pollutant mass budgets to be examined and consistency to be checked for coupled pollutant species. Different statistical analyses can also provide different information. For example, spatial plots of site-level statistics can show regional clustering of similar station scores and regional differences in station scores, including compensating site-level errors between regions or between urban and rural stations that can be hidden in and improve network-level aggregated statistics. Density scatterplots can show different levels of scatter between species or seasons and reveal low measurement precision by making quantization of values visible. Time series of monthly or hourly mean values are more aggregated but can help to understand temporal variations in performance and, when collated, mass budgets. Stratified analyses (e.g., by individual network, season, monthly, hour of day, region, or urban/rural) have also revealed model behaviour, including compensating errors, that was not visible in highly aggregated (e.g., all-station, annual) "headline" statistics. This comprehensive, quasi-diagnostic model performance evaluation has also provided a baseline set of scores against which future system versions can be compared.

Results from the performance evaluation described in this paper have pointed to aspects of the RAQDPS023 that warrant further investigation and improvement. First, the anthropogenic input emissions files and the parameterizations of natural emissions used by the forecast system should be one important focus, especially as anthropogenic emissions continue to change. In addition to strong evidence to support the inclusion of BB emissions, other results point to the likely overallocation of $PM_{2.5}$ primary emissions to urban areas, overallocation of NH_3 emissions to summer and underallocation to winter, overallocation of fugitive dust emissions to winter, excessive sea-salt emissions in all seasons, and several missing natural emissions sources, including wind-blown dust from natural surfaces and lightning NO emissions. Evaluation results for ISOP VMR also showed large positive biases, arguing for the biogenic emissions scheme to be revisited. Second, O_3 forecasts might be improved by revising O_3 lateral boundary conditions for the spring and introducing a parameterization of halogen chemistry impacts near oceans on O_3. Third, the RAQDPS023 was missing two gas-phase dry deposition processes, and the addition of these processes should decrease SO_2 and HNO_3 levels, which were both overpredicted. Seasonal surface properties in the gas-phase dry deposition scheme also need to have a more detailed representation of seasonal variations and to include dependence on longitude and elevation, which was lacking in the RAQDPS023. Fourth, precipitation-chemistry measurements suggested that too much NH_3 gas was being removed by wet deposition. Fifth, the RAQDPS023 was found to underpredict $PM_{2.5}$-SO_4 and overpredict $PM_{2.5}$-EC and $PM_{2.5}$-SS concentrations in all seasons, and the treatments of the lifecycles of each of these components will need to be examined. Particulate organic matter, especially $PM_{2.5}$-SOM from biogenic emissions, may be underpredicted in the summer, which will require examination of the representation of biogenic emissions and of SOA formation. $PM_{2.5}$-NH_4 and $PM_{2.5}$-NO_3 may also be underpredicted in the summer, which will require the representation of inorganic heterogeneous thermodynamics to be assessed, including the addition of an explicit treatment of base cations, a known gap in the RAQDPS023. Lastly, the calculation of hourly $PM_{2.5}$ total mass should include sea salt and both volatile and particle-bound aerosol water.

Appendix

Table A1. List of acronyms and abbreviations.

ADOM	Acid Deposition and Oxidant Model (Canada)
AirNow	Aerometric Information Retrieval Now (U.S.)
ALD2	acetaldehyde and higher aldehydes (ADOM-2 lumped VOC species)
AMoN	Ammonia Monitoring Network (U.S.)
APEI	Air Pollutant Emissions Inventory (Canada)
AQ	air quality
AQHI	Air Quality Health Index (Canada)
AQS	Air Quality System (U.S.)
BB	biomass burning
BDL	below detection limit
BEIS	Biogenic Emission Inventory System (U.S.)
CAMS	Copernicus Atmosphere Monitoring Service (EU)
CAPMoN	Canadian Acid Precipitation Monitoring Network
CASTNET	Clean Air Status and Trends Network (U.S.)
CEDS	Community Emissions Data System
CFFEPS	Canadian Forest Fire Emissions Prediction System
CFR	Code of Federal Regulations (U.S.)
CM	crustal material
CMC	Canadian Meteorological Centre
CRES	cresols and phenols (ADOM-2 lumped VOC species)
CRMSE	centered root mean square error
CSN	Chemical Speciation Network (U.S.)
CV	coefficient of variation (or relative standard deviation)
EC	elemental carbon
ECA	eastern Canada
ECCC	Environment and Climate Change Canada
ECMWF	European Centre for Medium-range Weather Forecasts
EEA	European Environment Agency
EI	emission inventory
EMEP	European Monitoring and Evaluation Programme
EPA	Environmental Protection Agency (U.S.)
EQUATES	EPA's Air QUAlity TimE Series
EST	Eastern Standard Time
ETHE	ethene and some isoprene oxidation products (ADOM-2 lumped VOC species)
EUS	eastern U.S.
FAC2	factor-of-two metric
FB	fractional bias
FE	fractional error
FEM	Federal Equivalent Method (U.S.)
FRM	Federal Reference Method (U.S.)
FW	FireWork

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

GAFIS	Global Air quality Forecasting and Information System (WMO)
GAW	Global Atmospheric Watch (WMO)
GEM	Global Environmental Multiscale (model) (ECCC)
GEM-MACH	Global Environmental Multiscale–Modelling Atmospheric CHemistry (model) (ECCC)
GEMS	Global and regional Earth-system Monitoring using Satellite and in-situ data (EU)
GEOS-CF	Goddard Earth Observing System - Composition Forecasting (NASA)
GIS	geographic information system
HETV	HETerogeneous Vectorized scheme
IAU	incremental analysis update
IAY	Instantaneous secondary organic Aerosol Yield
IFS-CAMS	Integrated Forecasting System – Copernicus Atmosphere Monitoring Service (ECMWF)
IFS-SILAM	Integrated Forecasting System – SILAM (ECMWF/Finnish Meteorological Institute)
IOA	index of agreement
IMPROVE	Interagency Monitoring of Protected Visual Environments (U.S.)
ISOP	isoprene
LBC	lateral boundary condition
LT	local time
MAE	mean absolute error
MB	mean bias
MDA8	maximum daily 8-hr average
MDL	minimum detection limit
MFB	mean fractional bias
MFE	mean fractional error
MMR	mass mixing ratio
MOVES	MOtor Vehicle Emission Simulator (U.S.)
NAAQS	National Ambient Air Quality Standards (U.S.)
NACC	NOAA-EPA Atmosphere-Chemistry Coupler
NADP	National Atmospheric Deposition Program (U.S.)
NAPS	National Air Pollution Surveillance system (Canada)
NAQFC	National Air Quality Forecast Capability (U.S.)
NASA	National Aeronautics and Space Administration (U.S.)
NAtChem	National Atmospheric Chemistry database (Canada)
NATTS	National Air Toxics Trends Sites (U.S.)
NEI	National Emission Inventory (U.S.)
NH4	particle ammonium
NMAE	normalized MAE
NMB	normalized mean bias
NME	normalized mean absolute error
NMSE	normalized mean square error
NO3	particle nitrate
NOAA	National Oceanic and Atmospheric Administration (U.S.)
NPRI	National Pollutant Release Inventory (Canada)
NRT	near real time
NSD	normalized standard deviation

NTN	National Trends Network (U.S. NADP)
NWP	numerical weather prediction
OC	organic carbon
OM	organic matter
PAMS	Photochemical Assessment Monitoring Stations (U.S.)
PBL	planetary boundary layer
PCL	Precipitation Coverage Length
PM	particulate matter
$PM_{2.5}$	particulate matter with aerodynamic diameter smaller than 2.5 µm
POM	primary organic matter
PR	precipitation
QA/QC	quality assurance/quality control
RAQDPS	Regional Air Quality Deterministic Prediction System (ECCC)
RAQDPS-FW	RAQDPS-FireWork
RDPS	Regional Deterministic Prediction System (ECCC)
RH	relative humidity
RMSE	root mean square error
SCC	Source Classification Code
SDM	standard deviation of model predictions
SDO	standard deviation of observations
SEMARNAT	Secretariat of Environment and Natural Resources (Mexico)
SGS	subgrid-scale
SILAM	System for Integrated modeLling of Atmospheric composition (Finnish Meteorological Institute)
S/L	state/local
SMOKE	Sparse Matrix Operator Kernel Emissions (modeling system)
SO4	particle sulfate
SOA	secondary organic aerosol
SOM	secondary organic matter
SS	sea salt
STP	standard temperature and pressure
TEOM	tapered element oscillating microbalance
TF	transportable fraction
TOC	total organic carbon
TOM	total organic matter
TP	total precipitation
UTC	Coordinated Universal Time
VMR	volume mixing ratio
VOC	volatile organic compound
WCA	western Canada
WMO	World Meteorological Organization
WRF	Weather Research and Forecasting NWP model (U.S.)
WUS	western U.S.

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

Table A2. Statistical measures used in this study for model performance evaluation plus coefficient of variation (or relative standard deviation). MFB and MFE have been used in some related studies (e.g., Manseau et al., 2022)

Metric Name	Abbreviation	Definition
Observed mean	$\overline{\mathrm{O}}$	$\frac{1}{N}\sum_1^N O_i$
Model mean	$\overline{\mathrm{M}}$	$\frac{1}{N}\sum_1^N M_i$
Mean bias	MB	$\frac{1}{N}\sum_1^N (M_i - O_i)$
Root mean square error	RMSE	$\left(\frac{1}{N}\sum_1^N (M_i - O_i)^2\right)^{1/2}$
Normalized mean bias	NMB	$\frac{\sum_1^N (M_i - O_i)}{\sum_1^N O_i}$
Normalized mean absolute error	NME	$\frac{\sum_1^N \lvert M_i - O_i \rvert}{\sum_1^N \lvert O_i \rvert}$
Pearson correlation coefficient	R	$\frac{\sum[(M_i - \bar{M}) \times (O_i - \bar{O})]}{\sqrt{\sum (M_i - \bar{M})^2 \times \sum (O_i - \bar{O})^2}}$
Centred RMSE	CRMSE	$\left(\frac{1}{N}\sum_1^N [(M_i - \bar{M}) - (O_i - \bar{O})]^2\right)^{1/2}$
Standard deviation (observations)	σ_O (or SDO)	$\left(\frac{1}{N}\sum_1^N (O_i - \bar{O})^2\right)^{1/2}$
Standard deviation (model)	σ_M (or SDM)	$\left(\frac{1}{N}\sum_1^N (M_i - \bar{M})^2\right)^{1/2}$
Normalized standard deviation	NSD	σ_M / σ_O
Coefficient of variation (observations)	CVO	$\sigma_O / \overline{O}$
Coefficient of variation (model)	CVM	$\sigma_M / \overline{M}$
Mean fractional bias	MFB	$\frac{2}{N}\sum_1^N \left(\frac{M_i - O_i}{M_i + O_i}\right)$
Mean fractional error	MFE	$\frac{2}{N}\sum_1^N \left\lvert\frac{M_i - O_i}{M_i + O_i}\right\rvert$

Table A3. Major upgrades to the operational RAQDPS from 2009-2021. See Moran et al. (2025) for more details.

Version	Release Date	Short Description
001	Nov. 2009	First version (Emission Inventories: 2006 CA, 2005 US, 1999 MX)
004	Oct. 2011	New emissions (EIs: 2006 CA, projected 2012 US, 1999 MX)
007	Oct. 2012	New model code, new grid (15 km → 10 km, 58 → 80 levels)
009	Feb. 2013	New model code with 3 bug fixes, including one to near-surface vertical diffusion
013	Jun. 2015	New emissions (EIs: 2010 CA, 2011 US, 1999 MX)
016	Sep. 2016	New model code, new vertical discretization (non-staggered → staggered)
020	Sep. 2018	New model code, new emissions (EIs: 2013 CA, projected 2017 US, 2008 MX)
021	Jul. 2019	New model code, new vertical discretization (80 → 84 levels), longer forecast (2–>3 days)
023	Nov. 2021	New model code, new emissions (EIs: projected 2020 CA, projected 2023 US & 2023 MX)

Table A4. Summary of Canadian and U.S. national annual wildfire statistics for 2013–2022 period. Data sources: Canadian Interagency Forest Fire Centre (2023); NOAA National Centers for Environmental Information (2023).

	Number of Fires			Hectares Burned		
Year	Canada	U.S.	Total	Canada	U.S.	Total
2013	6,246	46,615	52,861	4,203,867	1,743,054	5,946,921
2014	5,126	63,345	68,471	4,563,847	1,451,836	6,015,683
2015	7,068	61,922	68,990	3,903,277	4,097,506	8,000,783
2016	5,173	65,575	70,748	1,532,440	2,204,130	3,736,570
2017	5,611	66,131	71,742	3,371,825	3,958,259	7,330,084
2018	7,068	55,911	62,979	2,272,269	3,473,262	5,745,531
2019	3,933	49,786	53,719	1,787,793	1,873,699	3,661,492
2020	3,916	58,258	62,174	227,389	4,158,019	4,385,408
2021	6,596	58,733	65,329	4,307,520	2,889,342	7,196,862
2022	5,726	66,255	71,981	1,656,504	3,049,067	4,705,571

Table A5. Comparison of selected NAQFC domain-average monthly statistics for daily mean surface $PM_{3.5}$ predictions ($\mu g \cdot m^{-3}$) for 2014 and 2015 (from Lee et al., 2017) with domain-average seasonal statistics for RAQDPS010 forecasts for hourly mean surface $PM_{2.5}$ predictions ($\mu g \cdot m^{-3}$) for spring (MAM) and summer (JJA) 2014 and RAQDPS011 forecasts for winter (DJF) 2015 (from Moran et al., 2021a).

Period	System	O	M	MB	NMB	RMSE	R
May 2014	NAQFC	7.76	6.20	-1.56	-0.20	4.46	0.32
Spring 2014	RAQDPS010			-0.88	-0.12	8.02	0.37
July 2014	NAQFC	9.93	6.62	-2.71	-0.28	5.36	0.23
Summer 2014	RAQDPS010			0.64	0.07	11.21	0.29
January 2015	NAQFC	9.83	11.16	1.33	0.13	6.46	0.38
Winter 2015	RAQDPS011			-1.04	-0.12	10.15	0.38

Code and data availability

Version 5.1 of the GEM numerical weather prediction model code used by the RAQDPS023 is free software which can be redistributed and/or modified under the terms of version 2.1 of the GNU Lesser General Public License as published by the Free Software Foundation. The GEM source code has been developed by the Meteorological Research Division of ECCC. This code is available for download from https://zenodo.org/records/17782580 (Environment and Climate Change Canada, 2025).

MACH, the atmospheric chemistry library for the GEM model (©2007–2021, Air Quality Research Division and National Prediction Operations Division, Environment and Climate Change Canada), is free software that can be redistributed and/or modified under the terms of the GNU Lesser General Public License as published by the Free Software Foundation – either version 2.1 of the license or any later version. The GEM-MACH version 3.1.0.0 code used by the RAQDPS023 can be downloaded from website https://doi.org/10.5281/zenodo.15330612 (Savic-Jovcic et al., 2025). Related documentation is also available on that website, including information about key input and configuration files and copies of several relevant reports. The GEM-MACH v3.1.1.2 source code for the RAQDPS024, an equivalent version to the RAQDPS023 that went into operation after a migration to a new ECCC high-performance computer system in June 2022, is available at https://zenodo.org/records/13952893.

The CFFEPS version 4.1 code that was used by the RAQDPS-FW023 and RAQDPS-FW024 is free software that can be redistributed and/or modified under the terms of the GNU Lesser General Public License, either version 2.1 or any later version, as published by the Free Software Foundation. It is available to download from website https://doi.org/10.5281/zenodo.15305591 (Anderson and Chen, 2021).

Data sources for the AQ measurement data sets used in this study are listed in Table S2a of the Supplement. Multiple sets of files containing (i) the "raw" measurements that were used in this study, (ii) intermediate measurement files after necessary unit conversions and proxy calculations, (iii) filtered measurement files after application of validity and temporal completeness checks, and (iv) final, evaluation-ready, paired model-measurement files after temporal aggregation for data pooling and filtering for reconstructed mass completeness are available from https://doi.org/10.5281/zenodo.16944371 (Lupu and Moran, 2025). This data repository also contains a complete set of the output files of evaluation scores that were the basis for all of the evaluation-related tables and figures presented in this paper.

A package of seasonal and annual model-predicted gridded surface concentration fields and dry, wet, and total acidic deposition fields for the 2013-2016 simulations is available from the website https://doi.org/10.5281/zenodo.16970403 (Moran and Savic-Jovcic, 2025). Some of the archived model hourly output fields used to create this package were also used by Cathcart et al. (2025) in their recent paper.

All other data sets used in this work are available upon request from the authors. Please contact one of the corresponding authors to make a request.

Supplement

The supplement related to this article is available on-line at https://doi.org/10.5281/zenodo.16929949.

Author contributions

MDM was the science lead for the development of the online RAQDPS from the RAQDPS001 up to the RAQDPS023 and was the co-supervisor for all operational deliveries from 2009 to 2021. He conceived the objectives and scope of this study, oversaw the evaluations of the RAQDPS023, and prepared the initial and final versions of this paper. AL assisted with the 2013-2016 simulations, obtained all AQ measurement data sets, developed all measurement data processing and evaluation scripts, and generated all evaluation tables and data-related figures for the 2013–2016 and 2021/22 annual simulations, VSJ developed and maintained RAQDPS code and scripts, performed the 2013–2016 annual simulations, and prepared model-related figures and analyses. JZ performed the 2010–2019 operational evaluation, and JZ, QZ, EIB, and RM generated the 2013–2016 and 2021/22 anthropogenic emissions files and performed the analyses to construct Tables 1 and S1 and Figure S1. CAS led the migration from the RAQDPS023 to the RAQDPS024 on the new CMC supercomputers in June 2022 and the development of the RAQDPS025, which became operational in June 2024. SM has been the co-supervisor for all operational deliveries of the RAQDPS with the assistance of VSJ, JC, KM, RMA, and DK. JC led the development and delivery of the RAQDPS-FW023 and all CFFEPS versions with the assistance of KM and RMA. PMM oversaw operational evaluation of North American NRT AQ forecasts at ECCC for GAFIS. Lastly, AL, VSJ, JZ, RM, CAS, JC, QZ, EIB, SM, and RMA reviewed the manuscript.

Competing interests

The authors declare that they have no conflicts of interest.

Acknowledgements

We would like to thank Ayodeji Akingunola, David Anselmo, Didier Davignon, Annie Duhamel, Samuel Gilbert, Hugo Landry, Jessica Miville, David Niemi, Radenko Pavlovic, Si-Jun Peng, Jacinthe Racine, and Mourad Sassi of ECCC for their contributions to this project. We appreciate the generosity of Kristen Foley, Alison Eyth, and their colleagues at the U.S. EPA and General Dynamics Information Technology for providing early access to the EQUATES emissions data set (Foley et al., 2023). Access to a host of AQ measurement data sets was key to this study (see Table S2a), and we thank (i) ECCC for the use of daily filter pack and precipitation–chemistry observations from the CAPMoN network, (ii) ECCC and the provincial, territorial, and regional government NAPS partners for the use of their NAPS network ambient AQ observations, (iii) the U.S. NADP network for the use of observations from the

AMoN and NTN networks, (iv) the U.S. EPA for use of observations from the CASTNET and CSN networks and the AQS, (v) the U.S. EPA, the U.S. National Park Service, and other federal, state, and tribal partners for use of observations from the IMPROVE network, and (vi) the U.S. EPA and its state and local, tribal, federal, and Canadian and Mexican partners (https://www.airnow.gov/partners/) for the use of near-real-time observations from the AirNow metanetwork. AL thanks Kulbir Banwait, Amanda Cole, Amy Hou, and Bill Sukloff of the ECCC CAPMoN and NAtChem groups for insightful discussions about various AQ measurement datasets. We also appreciate the provision of operational AQ forecasts by NOAA, NASA, ECMWF, and the Finnish Meteorological Institute to ECCC in NRT under the WMO GAFIS initiative. MDM would also like to thank Rosa Wu for her continuing support of this paper over multiple years. Lastly, we have chosen to use colour-vision-deficiency-friendly, perceptually-uniform colour maps designed by Crameri et al. (2020) for some figures in this paper.

References

Ames, R. B. and Malm, W. C.: Comparison of sulfate and nitrate particle mass concentrations measured by IMPROVE and the CDN, Atmos. Environ., 35, 905–916, https://doi.org/10.1016/S1352-2310(00)00369-1, 2001.

Anderson, K. and Chen, J.: Canadian Fire Emissions Prediction System (CFFEPS) v4.1, Zenodo [software], https://doi.org/10.5281/ZENODO.2579382, 2021.

Ansari, A. S. and Pandis, S. N.: Response of inorganic PM to precursor concentrations, Environ. Sci. Technol., 32, 2706–2714, https://doi.org/10.1021/es971130j, 1998.

Appel, K. W., Bhave, P. V., Gilliland, A. B., Sarwar, G., and Roselle, S. J.: Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance; Part II—particulate matter, Atmos. Environ., 42, 6057–6066, https://doi.org/10.1016/j.atmosenv.2008.03.036, 2008.

Appel, K. W., Foley, K. M., Bash, J. O., Pinder, R. W., Dennis, R. L., Allen, D. J., and Pickering, K.: A multi-resolution assessment of the Community Multiscale Air Quality (CMAQ) model v4.7 wet deposition estimates for 2002–2006, Geosci. Model Dev., 4, 357–371, https://doi.org/10.5194/gmd-4-357-2011, 2011.

Appel, K. W., Pouliot, G. A., Simon, H., Sarwar, G., Pye, H. O. T., Napelenok, S. L., Akhtar, F., and Roselle, S. J.: Evaluation of dust and trace metal estimates from the Community Multiscale Air Quality (CMAQ) model version 5.0, Geosci. Model Dev., 6, 883–899, https://doi.org/10.5194/gmd-6-883-2013, 2013.

Appel, K. W., Napelenok, S. L., Foley, K. M., Pye, H. O. T., Hogrefe, C., Luecken, D. J., Bash, J. O., Roselle, S. J., Pleim, J. E., Foroutan, H., Hutzell, W. T., Pouliot, G. A., Sarwar, G., Fahey, K. M., Gantt, B., Gilliam, R. C., Heath, N. K., Kang, D., Mathur, R., Schwede, D. B., Spero, T. L., Wong, D. C., and Young, J. O.: Description and evaluation of the Community Multiscale Air Quality (CMAQ) modeling system version 5.1, Geosci. Model Dev., 10, 1703–1732, https://doi.org/10.5194/gmd-10-1703-2017, 2017.

Appel, K. W., Bash, J. O., Fahey, K. M., Foley, K. M., Gilliam, R. C., Hogrefe, C., Hutzell, W. T., Kang, D., Mathur, R., Murphy, B. N., Napelenok, S. L., Nolte, C. G., Pleim, J. E., Pouliot, G. A., Pye, H. O. T., Ran, L., Roselle, S. J., Sarwar, G., Schwede, D. B., Sidi, F. I., Spero, T. L., and Wong, D. C.: The Community Multiscale Air Quality (CMAQ) model versions 5.3 and 5.3.1: system updates and evaluation, Geosci. Model Dev., 14, 2867–2897, https://doi.org/10.5194/gmd-14-2867-2021, 2021.

Bachmann, J. D.: Tackling multi-pollutant particles, EM Air Waste Manag. Assoc. Mag. Environ. Manag., 6–9, https://www.researchgate.net/publication/268280224_Tackling_Multi-pollutant_Particles, 2013.

Bencala, K. E. and Seinfeld, J. H.: An air quality model performance assessment package, Atmospheric Environ. 1967, 13, 1181–1185, https://doi.org/10.1016/0004-6981(79)90043-X, 1979.

Benish, S. E., Bash, J. O., Foley, K. M., Appel, K. W., Hogrefe, C., Gilliam, R., and Pouliot, G.: Long-term regional trends of nitrogen and sulfur deposition in the United States from 2002 to 2017, Atmospheric Chem. Phys., 22, 12749–12767, https://doi.org/10.5194/acp-22-12749-2022, 2022.

Bermejo, R. and Conde, J.: A conservative quasi-monotone semi-Lagrangian scheme, Mon. Weather Rev., 130, 423–430, https://doi.org/10.1175/1520-0493(2002)130<0423:ACQMSL>2.0.CO;2, 2002.

Biswas, J., Hogrefe, C., Rao, S. T., Hao, W., and Sistla, G.: Evaluating the performance of regional-scale photochemical modeling systems. Part III—Precursor predictions, Atmos. Environ., 35, 6129–6149, https://doi.org/10.1016/S1352-2310(01)00401-0, 2001.

Bloom, S. C., Takacs, L. L., da Silva, A. M., and Ledvina, D.: Data assimilation using incremental analysis updates, Mon. Weather Rev., 124, 1256–1271, https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2, 1996.

Borrego, C., Monteiro, A., Ferreira, J., Miranda, A. I., Costa, A. M., Carvalho, A. C., and Lopes, M.: Procedures for estimation of modelling uncertainty in air quality assessment, Environ. Int., 34, 613–620, https://doi.org/10.1016/j.envint.2007.12.005, 2008.

Brasseur, G. P. and Kumar, R.: Chemical weather and chemical climate, AGU Adv., 2, e2021AV000399, https://doi.org/10.1029/2021AV000399, 2021.

Cai, C., Hogrefe, C., Katsafados, P., Kallos, G., Beauharnois, M., Schwab, J. J., Ren, X., Brune, W. H., Zhou, X., He, Y., and Demerjian, K. L.: 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, https://doi.org/10.1016/j.atmosenv.2008.08.029, 2008.

Campbell, P. C., Tang, Y., Lee, P., Baker, B., Tong, D., Saylor, R., Stein, A., Huang, J., Huang, H.-C., Strobach, E., McQueen, J., Pan, L., Stajner, I., Sims, J., Tirado-Delgado, J., Jung, Y., Yang, F., Spero, T. L., and Gilliam, R. C.: Development and evaluation of an advanced National Air Quality Forecasting Capability using the NOAA Global Forecast System version 16, Geosci. Model Dev., 15, 3281–3313, https://doi.org/10.5194/gmd-15-3281-2022, 2022.

Canadian Interagency Forest Fire Centre: Canada Report 2022, 15 pp., https://ciffc.ca/sites/default/files/2023-02/Canada_Report_2022_Final.pdf, 2023.

Caron, J.-F., Milewski, T., Buehner, M., Fillion, L., Reszka, M., Macpherson, S., and St-James, J.: Implementation of deterministic weather forecasting systems based on ensemble–variational data assimilation at Environment Canada. Part II: The regional system, Mon. Weather Rev., 143, 2560–2580, https://doi.org/10.1175/MWR-D-14-00353.1, 2015.

Cathcart, H., Aherne, J., Moran, M. D., Savic-Jovcic, V., Makar, P. A., and Cole, A.: Estimates of critical loads and exceedances of acidity and nutrient nitrogen for mineral soils in Canada for 2014–2016 average annual sulfur and nitrogen atmospheric deposition, Biogeosciences, 22, 535–554, https://doi.org/10.5194/bg-22-535-2025, 2025.

Chai, T., Kim, H.-C., Lee, P., Tong, D., Pan, L., Tang, Y., Huang, J., McQueen, J., Tsidulko, M., and Stajner, I.: Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and NO_2 measurements, Geosci. Model Dev., 6, 1831–1850, https://doi.org/10.5194/gmd-6-1831-2013, 2013.

Chan, E. A. W., Gantt, B., and McDow, S.: The reduction of summer sulfate and switch from summertime to wintertime $PM_{2.5}$ concentration maxima in the United States, Atmos. Environ., 175, 25–32, https://doi.org/10.1016/j.atmosenv.2017.11.055, 2018.

Chang, J. C. and Hanna, S. R.: Air quality model performance evaluation, Meteorol. Atmospheric Phys., 87, https://doi.org/10.1007/s00703-003-0070-7, 2004.

Chen, J. and Menelaou, K.: Regional Air Quality Deterministic Prediction System with near-real-time wildfire emissions (RAQDPSFW): Upgrade to version 023, Technical note, November, Canadian Centre for Meteorological and Environmental Prediction, Montreal, 31 pp., https://collaboration.cmc.ec.gc.ca/cmc/CMOI/product_guide/docs/tech_notes/technote_raqdpsfw_e.pdf, 2021.

Chen, J., Anderson, K., Pavlovic, R., Moran, M. D., Englefield, P., Thompson, D. K., Munoz-Alpizar, R., and Landry, H.: The FireWork v2.0 air quality forecast system with biomass burning emissions from the Canadian Forest Fire Emissions Prediction System v2.03, Geosci. Model Dev., 12, 3283–3310, https://doi.org/10.5194/gmd-12-3283-2019, 2019.

Chen, X., Zhang, Y., Wang, K., Tong, D., Lee, P., Tang, Y., Huang, J., Campbell, P. C., Mcqueen, J., Pye, H. O. T., Murphy, B. N., and Kang, D.: Evaluation of the offline-coupled GFSv15–FV3–CMAQv5.0.2 in support of the next-generation National Air Quality Forecast Capability over the contiguous United States, Geosci. Model Dev., 14, 3969–3993, https://doi.org/10.5194/gmd-14-3969-2021, 2021.

Chow, J. C.: Measurement methods to determine compliance with ambient air quality standards for suspended particles, J. Air Waste Manag. Assoc., 45, 320–382, https://doi.org/10.1080/10473289.1995.10467369, 1995.

Chow, J. C., Lowenthal, D. H., Chen, L.-W. A., Wang, X., and Watson, J. G.: Mass reconstruction methods for $PM_{2.5}$: a review, Air Qual. Atmosphere Health, 8, 243–263, https://doi.org/10.1007/s11869-015-0338-3, 2015.

Chuang, M.-T., Zhang, Y., and Kang, D.: Application of WRF/Chem-MADRID for real-time air quality forecasting over the southeastern United States, Atmos. Environ., 45, 6241–6250, https://doi.org/10.1016/j.atmosenv.2011.06.071, 2011.

Clifton, O. E., Schwede, D., Hogrefe, C., Bash, J. O., Bland, S., Cheung, P., Coyle, M., Emberson, L., Flemming, J., Fredj, E., Galmarini, S., Ganzeveld, L., Gazetas, O., Goded, I., Holmes, C. D., Horváth, L., Huijnen, V., Li, Q., Makar, P. A., Mammarella, I., Manca, G., Munger, J. W., Pérez-Camanyo, J. L., Pleim, J., Ran, L., San Jose, R., Silva, S. J., Staebler, R., Sun, S., Tai, A. P. K., Tas, E., Vesala, T., Weidinger, T., Wu, Z., and Zhang, L.: A single-point modeling approach for the intercomparison and evaluation of ozone dry deposition across chemical transport models (Activity 2 of AQMEII4), Atmospheric Chem. Phys., 23, 9911–9961, https://doi.org/10.5194/acp-23-9911-2023, 2023.

CMC-RAQDPS-023: The Regional Air Quality Deterministic Prediction System (RAQDPS) version 023 and the Regional Air Quality Deterministic Prediction System with Near-Real-Time Wildfire Emissions (RAQDPSFW) version 023 of the Meteorological Service of Canada (MSC): Technical Specifications Document, November, Canadian Centre for Meteorological and Environmental Prediction, Montreal, 19 pp., https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_specifications/tech_specifications_RAQDPS_023_e.pdf, 2021.

CMC-RAQDPS-025: The Regional Air Quality Deterministic Prediction System (RAQDPS): Upgrade from version 024 to version 025, June, Canadian Centre for Meteorological and Environmental Prediction, Montreal, 89 pp., https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_notes/technote_raqdps-v25_e.pdf, 2024.

CMC-RDPS-8.0.0: The Regional Deterministic Prediction System (RDPS) version 8.0.0 of the Meteorological Service of Canada (MSC): Technical specifications document, December, Canadian Centre for Meteorological and Environmental Prediction, Montreal, 10 pp., https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_specifications/tech_specifications_RDPS_8.0.0_e.pdf, 2021.

Colette, A., Andersson, C., Manders, A., Mar, K., Mircea, M., Pay, M.-T., Raffort, V., Tsyro, S., Cuvelier, C., Adani, M., Bessagnet, B., Bergström, R., Briganti, G., Butler, T., Cappelletti, A., Couvidat, F., D'Isidoro, M., Doumbia, T., Fagerli, H., Granier, C., Heyes, C., Klimont, Z., Ojha, N., Otero, N., Schaap, M., Sindelarova, K., Stegehuis, A. I., Roustan, Y., Vautard, R., Van Meijgaard, E., Vivanco, M. G., and Wind, P.: EURODELTA-Trends, a multi-model experiment of air quality hindcast in Europe over 1990–2010, Geosci. Model Dev., 10, 3255–3276, https://doi.org/10.5194/gmd-10-3255-2017, 2017.

Crameri, F., Shephard, G. E., and Heron, P. J.: The misuse of colour in science communication, Nat. Commun., 11, 5444, https://doi.org/10.1038/s41467-020-19160-7, 2020.

Dabek-Zlotorzynska, E., Dann, T. F., Kalyani Martinelango, P., Celo, V., Brook, J. R., Mathieu, D., Ding, L., and Austin, C. C.: Canadian National Air Pollution Surveillance (NAPS) $PM_{2.5}$ speciation program: Methodology and $PM_{2.5}$ chemical composition for the years 2003–2008, Atmos. Environ., 45, 673–686, https://doi.org/10.1016/j.atmosenv.2010.10.024, 2011.

Demerjian, K.: A review of national monitoring networks in North America, Atmos. Environ., 34, 1861–1884, https://doi.org/10.1016/S1352-2310(99)00452-5, 2000.

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S. T., Scheffe, R., Schere, K., Steyn, D., and Venkatram, A.: A framework for evaluating regional-scale numerical photochemical modeling systems, Environ. Fluid Mech., 10, 471–489, https://doi.org/10.1007/s10652-009-9163-2, 2010.

Dennis, R. L. and Downton, M. W.: Evaluation of urban photochemical models for regulatory use, Atmospheric Environ. 1967, 18, 2055–2069, https://doi.org/10.1016/0004-6981(84)90192-6, 1984.

Dennis, R. L., McHenry, J. N., Barchet, W. R., Binkowski, F. S., and Byun, D. W.: Correcting RADM's sulfate underprediction: Discovery and correction of model errors and testing the corrections through comparisons against field data, Atmospheric Environ. Part Gen. Top., 27, 975–997, https://doi.org/10.1016/0960-1686(93)90012-N, 1993.

Derwent, R., Fraser, A., Abbot, J., Jenkin, M., Willis, P., and Murrells, T.: Evaluating the Performance of Air Quality Models, Department for Environment Food and Rural Affairs, United Kingdom, https://uk-air.defra.gov.uk/assets/documents/reports/cat05/1006241607_100608_MIP_Final_Version.pdf, 2010.

Dickson, R. J. and Oliver, W. R.: Emissions models for regional air quality studies, Environ. Sci. Technol., 25, 1533–1535, https://doi.org/10.1021/es00021a003, 1991.

Dunlea, E. J., Herndon, S. C., Nelson, D. D., Volkamer, R. M., San Martini, F., Sheehy, P. M., Zahniser, M. S., Shorter, J. H., Wormhoudt, J. C., Lamb, B. K., Allwine, E. J., Gaffney, J. S., Marley, N. A., Grutter, M., Marquez, C., Blanco, S., Cardenas, B., Retama, A., Ramos Villegas, C. R., Kolb, C. E., Molina, L. T., and Molina, M. J.: Evaluation of nitrogen dioxide chemiluminescence monitors in a polluted urban environment, Atmospheric Chem. Phys., 7, 2691–2704, https://doi.org/10.5194/acp-7-2691-2007, 2007.

Dye, T. S., Chan, A. C., Anderson, C. B., D.E.B. Strohm, Wayland, R. A., and White, J. E.: From raw air quality data to the nightly news: an overview of how EPA's AirNow program operates, Sixth Conference on Atmospheric Chemistry, 12-14 January, Seattle, Washington, American Meteorological Society, https://ams.confex.com/ams/pdfpapers/72477.pdf, 2004.

ECCC: 1990–2016 Air Pollutant Emission Inventory Report, Environment and Climate Change Canada, Gatineau, Quebec, 92 pp., https://publications.gc.ca/collections/collection_2018/eccc/En81-26-2016-eng.pdf, 2018.

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

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

Eder, B., Kang, D., Rao, S. T., Mathur, R., Yu, S., Otte, T., Schere, K., Wayland, R., Jackson, S., Davidson, P., McQueen, J., and Bridgers, G.: Using National Air Quality Forecast Guidance to develop local Air Quality Index forecasts, Bull. Am. Meteorol. Soc., 91, 313–326, https://doi.org/10.1175/2009BAMS2734.1, 2010.

Emery, C., Liu, Z., Russell, A. G., Odman, M. T., Yarwood, G., and Kumar, N.: Recommendations on statistics and benchmarks to assess photochemical model performance, J. Air Waste Manag. Assoc., 67, 582–598, https://doi.org/10.1080/10962247.2016.1265027, 2017.

Entekhabi, D., Reichle, R. H., Koster, R. D., and Crow, W. T.: Performance metrics for soil moisture retrievals and application requirements, J. Hydrometeorol., 11, 832–840, https://doi.org/10.1175/2010JHM1223.1, 2010.

Environment and Climate Change Canada: Version 5.1 package for the Global Environmental Multiscale (GEM) model (ECCC-ASTD-MRD/gem: 5.1.0), , Zenodo [software], https://doi.org/10.5281/zenodo.17782580, 2025.

Feng, J., Chan, E., and Vet, R.: Air quality in the eastern United States and eastern Canada for 1990–2015: 25 years of change in response to emission reductions of SO_2 and NO_x in the region, Atmospheric Chem. Phys., 20, 3107–3134, https://doi.org/10.5194/acp-20-3107-2020, 2020.

Feng, J., Cole, A., Wetherbee, G. A., and Banwait, K.: Inter-comparison of measurements of inorganic chemical components in precipitation from NADP and CAPMoN at collocated sites in the USA and Canada during 1986–2019, Environ. Monit. Assess., 195, 1333, https://doi.org/10.1007/s10661-023-11771-z, 2023.

Fillion, L., Mitchell, H. L., Ritchie, H., and Staniforth, A.: The impact of a digital filter finalization technique in a global data assimilation system, Tellus A, 47, 304–323, https://doi.org/10.1034/j.1600-0870.1995.t01-2-00002.x, 1995.

Fillion, L., Tanguay, M., Lapalme, E., Denis, B., Desgagné, M., Lee, V., Ek, N., Liu, Z., Lajoie, M., Caron, J.-F., and Pagé, C.: The Canadian Regional Data Assimilation and Forecasting System, Weather Forecast., 25, 1645–1669, https://doi.org/10.1175/2010WAF2222401.1, 2010.

Foley, K. M., Hogrefe, C., Pouliot, G., Possiel, N., Roselle, S. J., Simon, H., and Timin, B.: Dynamic evaluation of CMAQ part I: Separating the effects of changing emissions and changing meteorology on ozone levels between 2002 and 2005 in the eastern US, Atmos. Environ., 103, 247–255, https://doi.org/10.1016/j.atmosenv.2014.12.038, 2015.

Foley, K. M., Pouliot, G. A., Eyth, A., Aldridge, M. F., Allen, C., Appel, K. W., Bash, J. O., Beardsley, M., Beidler, J., Choi, D., Farkas, C., Gilliam, R. C., Godfrey, J., Henderson, B. H., Hogrefe, C., Koplitz, S. N., Mason, R., Mathur, R., Misenis, C., Possiel, N., Pye, H. O. T., Reynolds, L., Roark, M., Roberts, S., Schwede, D. B., Seltzer, K. M., Sonntag, D., Talgo, K., Toro, C., Vukovich, J., Xing, J., and Adams, E.: 2002–2017 anthropogenic emissions data for air quality modeling over the United States, Data Brief, 47, 109022, 33 pp., https://doi.org/10.1016/j.dib.2023.109022, 2023.

Foroutan, H., Young, J., Napelenok, S., Ran, L., Appel, K. W., Gilliam, R. C., and Pleim, J. E.: Development and evaluation of a physics-based windblown dust emission scheme implemented in the CMAQ modeling system, J. Adv. Model. Earth Syst., 9, 585–608, https://doi.org/10.1002/2016MS000823, 2017.

Frank, N. H.: Retained nitrate, hydrated sulfates, and carbonaceous mass in Federal Reference Method fine particulate matter for six eastern U.S. cities, J. Air Waste Manag. Assoc., 56, 500–511, https://doi.org/10.1080/10473289.2006.10464517, 2006.

Fruin, S., Urman, R., Lurmann, F., McConnell, R., Gauderman, J., Rappaport, E., Franklin, M., Gilliland, F. D., Shafer, M., Gorski, P., and Avol, E.: Spatial variation in particulate matter components over a large urban area, Atmos. Environ., 83, 211–219, https://doi.org/10.1016/j.atmosenv.2013.10.063, 2014.

Galmarini, S., Makar, P., Clifton, O. E., Hogrefe, C., Bash, J. O., Bellasio, R., Bianconi, R., Bieser, J., Butler, T., Ducker, J., Flemming, J., Hodzic, A., Holmes, C. D., Kioutsioukis, I., Kranenburg, R., Lupascu, A., Perez-Camanyo, J. L., Pleim, J., Ryu, Y.-H., San Jose, R., Schwede, D., Silva, S., and Wolke, R.: Technical note: AQMEII4 Activity 1: evaluation of wet and dry deposition schemes as an integral part of regional-scale air quality models, Atmospheric Chem. Phys., 21, 15663–15697, https://doi.org/10.5194/acp-21-15663-2021, 2021.

Gan, C. M., Binkowski, F., Pleim, J., Xing, J., Wong, D., Mathur, R., and Gilliam, R.: Assessment of the aerosol optics component of the coupled WRF–CMAQ model using CARES field campaign data and a single column model, Atmos. Environ., 115, 670–682, https://doi.org/10.1016/j.atmosenv.2014.11.028, 2015.

Gantt, B.: 10 Years (2011-2020) of the NCore Network: FEMs vs FRMs, 2022 National Ambient Air Monitoring Conference, Aug. 22-25, Pittsburgh, Pennsylvania, https://www.epa.gov/system/files/documents/2022-10/Gantt_Brett_Wed1030.pdf, 2022.

Gaudel, A., Cooper, O. R., Ancellet, G., Barret, B., Boynard, A., Burrows, J. P., Clerbaux, C., Coheur, P.-F., Cuesta, J., Cuevas, E., Doniki, S., Dufour, G., Ebojie, F., Foret, G., Garcia, O., Granados-Muñoz, M. J., Hannigan, J. W., Hase, F., Hassler, B., Huang, G., Hurtmans, D., Jaffe, D., Jones, N., Kalabokas, P., Kerridge, B., Kulawik, S., Latter, B., Leblanc, T., Le Flochmoën, E., Lin, W., Liu, J., Liu, X., Mahieu, E., McClure-Begley, A., Neu, J. L., Osman, M., Palm, M., Petetin, H., Petropavlovskikh, I., Querel, R., Rahpoe, N., Rozanov, A., Schultz, M. G., Schwab, J., Siddans, R., Smale, D., Steinbacher, M., Tanimoto, H., Tarasick, D. W., Thouret, V., Thompson, A. M., Trickl, T., Weatherhead, E., Wespes, C., Worden, H. M., Vigouroux, C., Xu, X., Zeng, G., and Ziemke, J.: Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation, Elem. Sci. Anthr., 6, 39, https://doi.org/10.1525/elementa.291, 2018.

Gilliam, R. C., Hogrefe, C., Godowitch, J. M., Napelenok, S., Mathur, R., and Rao, S. T.: Impact of inherent meteorology uncertainty on air quality model predictions, J. Geophys. Res. Atmospheres, 120, https://doi.org/10.1002/2015JD023674, 2015.

Gilliam, R. C., Herwehe, J. A., Bullock, O. R., Pleim, J. E., Ran, L., Campbell, P. C., and Foroutan, H.: Establishing the suitability of the Model for Prediction Across Scales for global retrospective air quality modeling, J. Geophys. Res. Atmospheres, 126, e2020JD033588, https://doi.org/10.1029/2020JD033588, 2021.

Gilliland, A. B., Hogrefe, C., Pinder, R. W., Godowitch, J. M., Foley, K. L., and Rao, S. T.: Dynamic evaluation of regional air quality models: Assessing changes in O_3 stemming from changes in emissions and meteorology, Atmos. Environ., 42, 5110–5123, https://doi.org/10.1016/j.atmosenv.2008.02.018, 2008.

Godowitch, J. M., Pouliot, G. A., and Trivikrama Rao, S.: Assessing multi-year changes in modeled and observed urban NO_x concentrations from a dynamic model evaluation perspective, Atmos. Environ., 44, 2894–2901, https://doi.org/10.1016/j.atmosenv.2010.04.040, 2010.

Godowitch, J. M., Gilliam, R. C., and Rao, S. T.: Diagnostic evaluation of ozone production and horizontal transport in a regional photochemical air quality modeling system, Atmos. Environ., 45, 3977–3987, https://doi.org/10.1016/j.atmosenv.2011.04.062, 2011.

Hand, J. L., Schichtel, B. A., Pitchford, M., Malm, W. C., and Frank, N. H.: Seasonal composition of remote and urban fine particulate matter in the United States, J. Geophys. Res. Atmospheres, 117, 22 pp., https://doi.org/10.1029/2011JD017122, 2012.

Hand, J. L., Schichtel, B. A., Malm, W. C., and Frank, N. H.: Spatial and temporal trends in PM $_{2.5}$ organic and elemental carbon across the United States, Adv. Meteorol., 2013, 1–13, https://doi.org/10.1155/2013/367674, 2013.

Hand, J. L., Prenni, A. J., Schichtel, B. A., Malm, W. C., and Chow, J. C.: Trends in remote $PM_{2.5}$ residual mass across the United States: Implications for aerosol mass reconstruction in the IMPROVE network, Atmos. Environ., 203, 141–152, https://doi.org/10.1016/j.atmosenv.2019.01.049, 2019.

Hand, J. L., Prenni, A. J., and Schichtel, B. A.: Trends in seasonal mean speciated aerosol composition in remote areas of the United States from 2000 through 2021, J. Geophys. Res. Atmospheres, 129, e2023JD039902, https://doi.org/10.1029/2023JD039902, 2024.

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., Seibert, J. J., Vu, L., Andres, R. J., Bolt, R. M., Bond, T. C., Dawidowski, L., Kholod, N., Kurokawa, J., Li, M., Liu, L., Lu, Z., Moura, M. C. P., O'Rourke, P. R., and Zhang, Q.: Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS), Geosci. Model Dev., 11, 369–408, https://doi.org/10.5194/gmd-11-369-2018, 2018.

Hogrefe, C., Rao, S. T., Kasibhatla, P., Hao, W., Sistla, G., Mathur, R., and McHenry, J.: Evaluating the performance of regional-scale photochemical modeling systems: Part II—ozone predictions, Atmos. Environ., 35, 4175–4188, https://doi.org/10.1016/S1352-2310(01)00183-2, 2001a.

Hogrefe, C., Rao, S. T., Kasibhatla, P., Kallos, G., Tremback, C. J., Hao, W., Olerud, D., Xiu, A., McHenry, J., and Alapaty, K.: Evaluating the performance of regional-scale photochemical modeling systems: Part I—meteorological predictions, Atmos. Environ., 35, 4159–4174, https://doi.org/10.1016/S1352-2310(01)00182-0, 2001b.

Hogrefe, C., Pouliot, G., Wong, D., Torian, A., Roselle, S., Pleim, J., and Mathur, R.: Annual application and evaluation of the online coupled WRF–CMAQ system over North America under AQMEII phase 2, Atmos. Environ., 115, 683–694, https://doi.org/10.1016/j.atmosenv.2014.12.034, 2015.

Holden, A. S., Sullivan, A. P., Munchak, L. A., Kreidenweis, S. M., Schichtel, B. A., Malm, W. C., and Collett, J. L.: Determining contributions of biomass burning and other sources to fine particle contemporary carbon in the western United States, Atmos. Environ., 45, 1986–1993, https://doi.org/10.1016/j.atmosenv.2011.01.021, 2011.

Houyoux, M. R., Vukovich, J. M., Coats, C. J., Wheeler, N. J. M., and Kasibhatla, P. S.: Emission inventory development and processing for the Seasonal Model for Regional Air Quality (SMRAQ) project, J. Geophys. Res. Atmospheres, 105, 9079–9090, https://doi.org/10.1029/1999JD900975, 2000.

Huang, J., McQueen, J., Wilczak, J., Djalalova, I., Stajner, I., Shafran, P., Allured, D., Lee, P., Pan, L., Tong, D., Huang, H.-C., DiMego, G., Upadhayay, S., and Delle Monache, L.: Improving NOAA NAQFC $PM_{2.5}$ predictions with a bias correction approach, Weather Forecast., 32, 407–421, https://doi.org/10.1175/WAF-D-16-0118.1, 2017.

Huang, L., Zhu, Y., Zhai, H., Xue, S., Zhu, T., Shao, Y., Liu, Z., Emery, C., Yarwood, G., Wang, Y., Fu, J., Zhang, K., and Li, L.: Recommendations on benchmarks for numerical air quality model applications in China – Part 1: $PM_{2.5}$ and chemical species, Atmospheric Chem. Phys., 21, 2725–2743, https://doi.org/10.5194/acp-21-2725-2021, 2021.

Hystad, P., Setton, E., Cervantes, A., Poplawski, K., Deschenes, S., Brauer, M., Van Donkelaar, A., Lamsal, L., Martin, R., Jerrett, M., and Demers, P.: Creating national air pollution models for population exposure assessment in Canada, Environ. Health Perspect., 119, 1123–1129, https://doi.org/10.1289/ehp.1002976, 2011.

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baró, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., Flemming, J., Forkel, R., Giordano, L., Jiménez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Kuenen, J. J. P., Makar, P. A., Manders-Groot, A., Neal, L., Pérez, J. L., Pirovano, G., Pouliot, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R. S., Syrakov, D., Torian, A., Tuccella, P., Werhahn, J., Wolke, R., Yahya, K., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., and Galmarini, S.: Evaluation of operational on-line-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part I: Ozone, Atmos. Environ., 115, 404–420, https://doi.org/10.1016/j.atmosenv.2014.09.042, 2015a.

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baró, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., Denier Van Der Gon, H., Flemming, J., Forkel, R., Giordano, L., Jiménez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Makar, P. A., Manders-Groot, A., Neal, L., Pérez, J. L., Pirovano, G., Pouliot, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R. S., Syrakov, D., Torian, A., Tuccella, P., Wang, K., Werhahn, J., Wolke, R., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., and Galmarini, S.: Evaluation of operational online-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part II: Particulate matter, Atmos. Environ., 115, 421–441, https://doi.org/10.1016/j.atmosenv.2014.08.072, 2015b.

Jaffe, D., Bertschi, I., Jaegle, L., Novelli, P., Reid, J. S., Tanimoto, H., Vingarzan, R., and Westphal, D. L.: Long-range transport of Siberian biomass burning emissions and impact on surface ozone in western North America, Geophys. Res. Lett., 31, L16106, 4 pp., https://doi.org/10.1029/2004GL020093, 2004.

Jain, P., Sharma, A. R., Acuna, D. C., Abatzoglou, J. T., and Flannigan, M.: Record-breaking fire weather in North America in 2021 was initiated by the Pacific northwest heat dome, Commun. Earth Environ., 5, 202, https://doi.org/10.1038/s43247-024-01346-2, 2024.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., and Giovis, C.: A review and evaluation of intraurban air pollution exposure models, J. Expo. Sci. Environ. Epidemiol., 15, 185–204, https://doi.org/10.1038/sj.jea.7500388, 2005.

Kelly, J. T., Koplitz, S. N., Baker, K. R., Holder, A. L., Pye, H. O. T., Murphy, B. N., Bash, J. O., Henderson, B. H., Possiel, N. C., Simon, H., Eyth, A. M., Jang, C., Phillips, S., and Timin, B.: Assessing PM2.5 model performance for the conterminous U.S. with comparison to model performance statistics from 2007-2015, Atmos. Environ., 214, 116872, https://doi.org/10.1016/j.atmosenv.2019.116872, 2019.

Knote, C., Tuccella, P., Curci, G., Emmons, L., Orlando, J. J., Madronich, S., Baró, R., Jiménez-Guerrero, P., Luecken, D., Hogrefe, C., Forkel, R., Werhahn, J., Hirtl, M., Pérez, J. L., San José, R., Giordano, L., Brunner, D., Yahya, K., and Zhang, Y.: Influence of the choice of gas-phase mechanism on predictions of key gaseous pollutants during the AQMEII phase-2 intercomparison, Atmos. Environ., 115, 553–568, https://doi.org/10.1016/j.atmosenv.2014.11.066, 2015.

Koo, B., Kumar, N., Knipping, E., Nopmongcol, U., Sakulyanontvittaya, T., Odman, M. T., Russell, A. G., and Yarwood, G.: Chemical transport model consistency in simulating regulatory outcomes and the relationship to model performance, Atmos. Environ., 116, 159–171, https://doi.org/10.1016/j.atmosenv.2015.06.036, 2015.

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

Lamsal, L. N., Duncan, B. N., Yoshida, Y., Krotkov, N. A., Pickering, K. E., Streets, D. G., and Lu, Z.: U.S. NO_2 trends (2005–2013): EPA Air Quality System (AQS) data versus improved observations from the Ozone Monitoring Instrument (OMI), Atmos. Environ., 110, 130–143, https://doi.org/10.1016/j.atmosenv.2015.03.055, 2015.

Lavery, T. F., Rogers, C. M., Baumgardner, R., and Mishoe, K. P.: Intercomparison of Clean Air Status and Trends Network nitrate and nitric acid measurements with data from other monitoring programs, J. Air Waste Manag. Assoc., 59, 214–226, https://doi.org/10.3155/1047-3289.59.2.214, 2009.

Lee, H. J., Chatfield, R. B., and Bell, M. L.: Spatial analysis of concentrations of multiple air pollutants using NASA DISCOVER-AQ aircraft measurements: Implications for exposure assessment, Environ. Res., 160, 487–498, https://doi.org/10.1016/j.envres.2017.10.017, 2018.

Lee, P., McQueen, J., Stajner, I., Huang, J., Pan, L., Tong, D., Kim, H., Tang, Y., Kondragunta, S., Ruminski, M., Lu, S., Rogers, E., Saylor, R., Shafran, P., Huang, H.-C., Gorline, J., Upadhayay, S., and Artz, R.: NAQFC developmental forecast guidance for fine particulate matter ($PM_{2.5}$), Weather Forecast., 32, 343–360, https://doi.org/10.1175/WAF-D-15-0163.1, 2017.

Li, Q., Borge, R., Sarwar, G., de la Paz, D., Gantt, B., Domingo, J., Cuevas, C. A., and Saiz-Lopez, A.: Impact of halogen chemistry on summertime air quality in coastal and continental Europe: application of the CMAQ model and implications for regulation, Atmospheric Chem. Phys., 19, 15321–15337, https://doi.org/10.5194/acp-19-15321-2019, 2019.

Li, W., Tang, B., Campbell, P. C., Tang, Y., Baker, B., Moon, Z., Tong, D., Huang, J., Wang, K., Stajner, I., and Montuoro, R.: Updates and evaluation of NOAA's online-coupled air quality model version 7 (AQMv7) within the Unified Forecast System, Geosci. Model Dev., 18, 1635–1660, https://doi.org/10.5194/gmd-18-1635-2025, 2025.

Liu, J. C., Pereira, G., Uhl, S. A., Bravo, M. A., and Bell, M. L.: A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke, Environ. Res., 136, 120–132, https://doi.org/10.1016/j.envres.2014.10.015, 2015.

Liudchik, A., Pakatashkin, V., Umreika, S., and Girgzdiene, R.: Role of ozone deposition in the occurrence of the spring maximum, Atmosphere-Ocean, 53, 42–49, https://doi.org/10.1080/07055900.2013.853284, 2015.

Lupu, A. and Moran, M. D.: Operational GEM-MACH model evaluation against air quality surface observation networks across Canada and the United States for 2013-16 and 2021/22, Zenodo [dataset], https://doi.org/10.5281/zenodo.16944371, 2025.

Ma, S., Tong, D., Lamsal, L., Wang, J., Zhang, X., Tang, Y., Saylor, R., Chai, T., Lee, P., Campbell, P., Baker, B., Kondragunta, S., Judd, L., Berkoff, T. A., Janz, S. J., and Stajner, I.: Improving predictability of high-ozone episodes through dynamic boundary conditions, emission refresh and chemical data assimilation during the Long Island Sound Tropospheric Ozone Study (LISTOS) field campaign, Atmospheric Chem. Phys., 21, 16531–16553, https://doi.org/10.5194/acp-21-16531-2021, 2021.

Makar, P. A., Nissen, R., Teakles, A., Zhang, J., Zheng, Q., Moran, M. D., Yau, H., and diCenzo, C.: Turbulent transport, emissions and the role of compensating errors in chemical transport models, Geosci. Model Dev., 7, 1001–1024, https://doi.org/10.5194/gmd-7-1001-2014, 2014.

Makar, P. A., Akingunola, A., Aherne, J., Cole, A. S., Aklilu, Y., Zhang, J., Wong, I., Hayden, K., Li, S.-M., Kirk, J., Scott, K., Moran, M. D., Robichaud, A., Cathcart, H., Baratzedah, P., Pabla, B., Cheung, P., Zheng, Q., and Jeffries, D. S.: Estimates of exceedances of critical loads for acidifying deposition in Alberta and Saskatchewan, Atmospheric Chem. Phys., 18, 9897–9927, https://doi.org/10.5194/acp-18-9897-2018, 2018.

Makar, P. A., Stroud, C., Akingunola, A., Zhang, J., Ren, S., Cheung, P., and Zheng, Q.: Vehicle-induced turbulence and atmospheric pollution, Atmospheric Chem. Phys., 21, 12291–12316, https://doi.org/10.5194/acp-21-12291-2021, 2021.

Malm, W. C., Schichtel, B. A., Pitchford, M. L., Ashbaugh, L. L., and Eldred, R. A.: Spatial and monthly trends in speciated fine particle concentration in the United States, J. Geophys. Res. Atmospheres, 109, D03306, 22 pp., https://doi.org/10.1029/2003JD003739, 2004.

Malm, W. C., Schichtel, B. A., and Pitchford, M. L.: Uncertainties in $PM_{2.5}$ gravimetric and speciation measurements and what we can learn from them, J. Air Waste Manag. Assoc., 61, 1131–1149, https://doi.org/10.1080/10473289.2011.603998, 2011.

Manseau, P. M., Peng, S. J., Stroud, C., Savic-Jovcic, V., and Lupu, A.: AQ Multi Model Verification for North America: 2022/04 - 2022/06, Environment and Climate Change Canada, July, 11 pp., https://hpfx.collab.science.gc.ca/~svfs000/na-aq-mm-fe/reports/2022/na-aq-mm-fe_reports_2022_Q2.pdf, 2022.

Mao, Y. H., Li, Q. B., Zhang, L., Chen, Y., Randerson, J. T., Chen, D., and Liou, K. N.: Biomass burning contribution to black carbon in the western United States mountain ranges, Atmospheric Chem. Phys., 11, 11253–11266, https://doi.org/10.5194/acp-11-11253-2011, 2011.

Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Denier Van Der Gon, H. A. C., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., Eskes, H. J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadygrov, N., Kaiser, J. W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouïl, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, Á., Van Velthoven, P., Van Versendaal, R., Vira, J., and Ung, A.: A regional air quality forecasting system over Europe: the MACC-II daily ensemble production, Geosci. Model Dev., 8, 2777–2813, https://doi.org/10.5194/gmd-8-2777-2015, 2015.

Mashayekhi, R., Pavlovic, R., Racine, J., Moran, M. D., Manseau, P. M., Duhamel, A., Katal, A., Miville, J., Niemi, D., Peng, S. J., Sassi, M., Griffin, D., and McLinden, C. A.: Isolating the impact of COVID-19 lockdown measures on urban air quality in Canada, Air Qual. Atmosphere Health, 14, 1549–1570, https://doi.org/10.1007/s11869-021-01039-1, 2021.

Mathur, R., Yu, S., Kang, D., and Schere, K. L.: Assessment of the wintertime performance of developmental particulate matter forecasts with the Eta-Community Multiscale Air Quality modeling system, J. Geophys. Res. Atmospheres, 113, 2007JD008580, https://doi.org/10.1029/2007JD008580, 2008.

Matthias, V., Arndt, J. A., Aulinger, A., Bieser, J., Denier Van Der Gon, H., Kranenburg, R., Kuenen, J., Neumann, D., Pouliot, G., and Quante, M.: Modeling emissions for three-dimensional atmospheric chemistry transport models, J. Air Waste Manag. Assoc., 68, 763–800, https://doi.org/10.1080/10962247.2018.1424057, 2018.

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

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

McKeen, S., Grell, G., Peckham, S., Wilczak, J., Djalalova, I., Hsie, E. -Y., Frost, G., Peischl, J., Schwarz, J., Spackman, R., Holloway, J., De Gouw, J., Warneke, C., Gong, W., Bouchet, V., Gaudreault, S., Racine, J., McHenry, J., McQueen, J., Lee, P., Tang, Y., Carmichael, G. R., and Mathur, R.: 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. Atmospheres, 114, 2008JD011697, https://doi.org/10.1029/2008JD011697, 2009.

McNair, L. A., Harley, R. A., and Russell, A. G.: Spatial inhomogeneity in pollutant concentrations, and their implications for air quality model evaluation, Atmos. Environ., 30, 4291–4301, https://doi.org/10.1016/1352-2310(96)00098-2, 1996.

McNider, R. T. and Pour-Biazar, A.: Meteorological modeling relevant to mesoscale and regional air quality applications: a review, J. Air Waste Manag. Assoc., 70, 2–43, https://doi.org/10.1080/10962247.2019.1694602, 2020.

McTaggart-Cowan, R., Vaillancourt, P. A., Zadra, A., Chamberland, S., Charron, M., Corvec, S., Milbrandt, J. A., Paquin-Ricard, D., Patoine, A., Roch, M., Separovic, L., and Yang, J.: Modernization of atmospheric physics parameterization in Canadian NWP, J. Adv. Model. Earth Syst., 11, 3593–3635, https://doi.org/10.1029/2019MS001781, 2019.

Meng, Z., Dabdub, D., and Seinfeld, J. H.: Chemical coupling between atmospheric ozone and particulate matter, Science, 277, 116–119, https://doi.org/10.1126/science.277.5322.116, 1997.

Miller, S. J., Makar, P. A., and Lee, C. J.: HETerogeneous vectorized or Parallel (HETPv1.0): an updated inorganic heterogeneous chemistry solver for the metastable-state NH_4^+ –Na^+ –Ca^{2+} –K^+ –Mg^{2+} –SO_4^{2-} –NO_3^- –Cl^- –H_2 O system based on ISORROPIA II, Geosci. Model Dev., 17, 2197–2219, https://doi.org/10.5194/gmd-17-2197-2024, 2024.

Momeni, M., Kashfi Yeganeh, A., Zanganeh Kia, H., Ghahremanloo, M., Mousavinezhad, S., De Guzman, H. J., Shephard, M. W., Jacobson, M. Z., and Choi, Y.: Using multi-satellite observations to constrain ammonia emissions and unlock their potential over open water, Sci. Rep., 15, https://doi.org/10.1038/s41598-025-09933-9, 2025.

Monks, P. S.: A review of the observations and origins of the spring ozone maximum, Atmos. Environ., 34, 3545–3561, https://doi.org/10.1016/S1352-2310(00)00129-1, 2000.

Moran, M. D. and Savic-Jovcic, V.: RAQDPS023 Predicted 2013-2016 and 2021/22 Seasonal and Annual Dry, Wet, and Total Acidic Deposition Fields and Related Concentration Fields for North America, Zenodo [dataset], https://doi.org/10.5281/zenodo.16970403, 2025.

Moran, M., Zhang, J., Pavlovic, R., Savic-Jovcic, V., Ménard, S., Landry, H., Zheng, Q., Lupu, A., Gilbert, S., Peng, S. J., and Manseau, P. M.: Evolution of the performance of the Canadian operational Regional Air Quality Deterministic Prediction System from 2010 to 2019, in: Air Pollution Modeling and its Application XXVII, edited by: Mensink, C. and Matthias, V., Springer Berlin Heidelberg, Berlin, Heidelberg, 157–166, https://doi.org/10.1007/978-3-662-63760-9_24, 2021a.

Moran, M. D., Ménard, S., and Kornic, D.: Regional Air Quality Deterministic Prediction System (RAQDPS): Upgrade from version 022 to version 023, Technical note, December, Canadian Centre for Meteorological and Environmental Prediction, Montreal, 48 pp., https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/technote_raqdps023_20211130_e.pdf, 2021b.

Moran, M. D., Savic-Jovcic, V., Stroud, C. A., Ménard, S., Gong, W., Zhang, J., Zheng, Q., Chen, J., Akingunola, A., Lupu, A., Menelaou, K., and Munoz-Alpizar, R.: Operational chemical weather forecasting with the ECCC online Regional Air Quality Deterministic Prediction System version 023 (RAQDPS023) - Part 1: System description, Geosci. Model Dev. (under review), https://doi.org/10.5194/egusphere-2025-4323, 2025.

Nappo, C. J., Caneill, J. Y., Furman, R. W., Gifford, F. W., Kaimal, J. C., Kramer, M. L., Lockhart, T. J., Pendergast, M. M., Pielke, R. A., Randerson, D., Shreffler, J. H., and Wyngaard, J. C.: The workshop on the representativeness of meteorological observations, June 1981, Boulder, Colo., Bull. Amer. Meteor. Soc., 63, 761–764, 1982.

National Atmospheric Deposition Program: National Atmospheric Deposition Program 2013 Annual Summary, Illinois State Water Survey, University of Illinois at Urbana-Champaign, Illinois, 28 pp., https://nadp.slh.wisc.edu/wp-content/uploads/2021/05/2013as.pdf, 2014.

National Atmospheric Deposition Program: National Atmospheric Deposition Program 2016 Annual Summary, Illinois State Water Survey, University of Illinois at Urbana-Champaign, Illinois, 28 pp., https://nadp.slh.wisc.edu/wp-content/uploads/2021/05/2016as.pdf, 2017.

Nguyen, T. K. V., Zhang, Q., Jimenez, J. L., Pike, M., and Carlton, A. G.: Liquid water: ubiquitous contributor to aerosol mass, Environ. Sci. Technol. Lett., 3, 257–263, https://doi.org/10.1021/acs.estlett.6b00167, 2016.

NOAA National Centers for Environmental Information: Monthly Wildfires Report for Annual 2022, WWW Document, https://www.ncei.noaa.gov/access/monitoring/monthly-report/fire/202213, 2023.

Noble, C. A., Vanderpool, R. W., Peters, T. M., McElroy, F. F., Gemmill, D. B., and Wiener, R. W.: Federal reference and equivalent methods for measuring fine particulate matter, Aerosol Sci. Technol., 34, 457–464, https://doi.org/10.1080/02786820121582, 2001.

Pagowski, M., Grell, G. A., McKeen, S. A., Peckham, S. E., and Devenyi, D.: Three-dimensional variational data assimilation of ozone and fine particulate matter observations: some results using the Weather Research and Forecasting—Chemistry model and Grid-Point Statistical Interpolation, Q. J. R. Meteorol. Soc., 136, 2013–2024, https://doi.org/10.1002/qj.700, 2010.

Pan, L., Tong, D., Lee, P., Kim, H.-C., and Chai, T.: 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, https://doi.org/10.1016/j.atmosenv.2014.06.020, 2014.

Park, S. H., Gong, S. L., Gong, W., Makar, P. A., Moran, M. D., Zhang, J., and Stroud, C. A.: Relative impact of windblown dust versus anthropogenic fugitive dust in $PM_{2.5}$ on air quality in North America, J. Geophys. Res. Atmospheres, 115, 2009JD013144, https://doi.org/10.1029/2009JD013144, 2010.

Parrish, D. D. and Fehsenfeld, F. C.: Methods for gas-phase measurements of ozone, ozone precursors and aerosol precursors, Atmos. Environ., 34, 1921–1957, https://doi.org/10.1016/S1352-2310(99)00454-9, 2000.

Pavlovic, R., Chen, J., Anderson, K., Moran, M. D., Beaulieu, P.-A., Davignon, D., and Cousineau, S.: The FireWork air quality forecast system with near-real-time biomass burning emissions: Recent developments and evaluation of performance for the 2015 North American wildfire season, J. Air Waste Manag. Assoc., 66, 819–841, https://doi.org/10.1080/10962247.2016.1158214, 2016.

Pendlebury, D., Gravel, S., Moran, M. D., and Lupu, A.: Impact of chemical lateral boundary conditions in a regional air quality forecast model on surface ozone predictions during stratospheric intrusions, Atmos. Environ., 174, 148–170, https://doi.org/10.1016/j.atmosenv.2017.10.052, 2018.

Penkett, S. A. and Brice, K. A.: The spring maximum in photo-oxidants in the Northern Hemisphere troposphere, Nature, 319, 655–657, https://doi.org/10.1038/319655a0, 1986.

Puchalski, M. A., Sather, M. E., Walker, J. T., Lehmann, C. M. B., Gay, D. A., Mathew, J., and Robarge, W. P.: Passive ammonia monitoring in the United States: Comparing three different sampling devices, J. Environ. Monit., 13, 3156, https://doi.org/10.1039/c1em10553a, 2011.

Pun, B. K., Seigneur, C., Bailey, E. M., Gautney, L. L., Douglas, S. G., Haney, J. L., and Kumar, N.: Response of atmospheric particulate matter to changes in precursor emissions: A comparison of three air quality models, Environ. Sci. Technol., 42, 831–837, https://doi.org/10.1021/es702333d, 2008.

Pye, H. O. T., Murphy, B. N., Xu, L., Ng, N. L., Carlton, A. G., Guo, H., Weber, R., Vasilakos, P., Appel, K. W., Budisulistiorini, S. H., Surratt, J. D., Nenes, A., Hu, W., Jimenez, J. L., Isaacman-VanWertz, G., Misztal, P. K., and Goldstein, A. H.: On the implications of aerosol liquid water and phase separation for organic aerosol mass, Atmospheric Chem. Phys., 17, 343–369, https://doi.org/10.5194/acp-17-343-2017, 2017.

Rappold, A. G., Reyes, J., Pouliot, G., Cascio, W. E., and Diaz-Sanchez, D.: Community vulnerability to health impacts of wildland fire smoke exposure, Environ. Sci. Technol., 51, 6674–6682, https://doi.org/10.1021/acs.est.6b06200, 2017.

Ren, S., Stroud, C., Belair, S., Leroyer, S., Munoz-Alpizar, R., Moran, M., Zhang, J., Akingunola, A., and Makar, P.: Impact of urbanization on the predictions of urban meteorology and air pollutants over four major North American cities, Atmosphere, 11, 969, https://doi.org/10.3390/atmos11090969, 2020.

Robichaud, A. and Ménard, R.: Multi-year objective analyses of warm season ground-level ozone and $PM_{2.5}$ over North America using real-time observations and Canadian operational air quality models, Atmospheric Chem. Phys., 14, 1769–1800, https://doi.org/10.5194/acp-14-1769-2014, 2014.

Robichaud, A., Ménard, R., Zaïtseva, Y., and Anselmo, D.: Multi-pollutant surface objective analyses and mapping of air quality health index over North America, Air Qual. Atmosphere Health, 9, 743–759, https://doi.org/10.1007/s11869-015-0385-9, 2016.

Sakaguchi, K., Zeng, X., and Brunke, M. A.: The hindcast skill of the CMIP ensembles for the surface air temperature trend, J. Geophys. Res. Atmospheres, 117, 2012JD017765, https://doi.org/10.1029/2012JD017765, 2012.

Sarwar, G., Gantt, B., Schwede, D., Foley, K., Mathur, R., and Saiz-Lopez, A.: Impact of enhanced ozone deposition and halogen chemistry on tropospheric ozone over the Northern Hemisphere, Environ. Sci. Technol., 49, 9203–9211, https://doi.org/10.1021/acs.est.5b01657, 2015.

Savage, N. H., Agnew, P., Davis, L. S., Ordóñez, C., Thorpe, R., Johnson, C. E., O'Connor, F. M., and Dalvi, M.: Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation, Geosci. Model Dev., 6, 353–372, https://doi.org/10.5194/gmd-6-353-2013, 2013.

Savic-Jovcic, V., Moran, M. D., and GEM-MACH Development Team: Global Environmental Multiscale model–Modelling Atmospheric CHemistry (GEM-MACH) version 3.1.0.0, Zenodo [software] [dataset], https://doi.org/10.5281/zenodo.15330612, 2025.

Saylor, R. D. and Stein, A. F.: Identifying the causes of differences in ozone production from the CB05 and CBMIV chemical mechanisms, Geosci. Model Dev., 5, 257–268, https://doi.org/10.5194/gmd-5-257-2012, 2012.

Schichtel, B. A., Hand, J. L., Barna, M. G., Gebhart, K. A., Copeland, S., Vimont, J., and Malm, W. C.: Origin of fine particulate carbon in the rural United States, Environ. Sci. Technol., 51, 9846–9855, https://doi.org/10.1021/acs.est.7b00645, 2017.

Schutgens, N. A. J., Gryspeerdt, E., Weigum, N., Tsyro, S., Goto, D., Schulz, M., and Stier, P.: Will a perfect model agree with perfect observations? The impact of spatial sampling, Atmospheric Chem. Phys., 16, 6335–6353, https://doi.org/10.5194/acp-16-6335-2016, 2016.

Schwede, D., Zhang, L., Vet, R., and Lear, G.: An intercomparison of the deposition models used in the CASTNET and CAPMoN networks, Atmos. Environ., 45, 1337–1346, https://doi.org/10.1016/j.atmosenv.2010.11.050, 2011.

Seigneur, C. and Moran, M. D.: Chemical transport models, in: Particulate Matter Science for Policy Makers: A NARSTO Assessment, Cambridge University Press, Cambridge, 283–323, 2004.

Sillman, S.: The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments, Atmos. Environ., 33, 1821–1845, https://doi.org/10.1016/S1352-2310(98)00345-8, 1999.

Simon, H., Baker, K. R., and Phillips, S.: Compilation and interpretation of photochemical model performance statistics published between 2006 and 2012, Atmos. Environ., 61, 124–139, https://doi.org/10.1016/j.atmosenv.2012.07.012, 2012.

Sirois, A., Vet, R., and Lamb, D.: A comparison of the precipitation chemistry measurements obtained by the CAPMoN and NADP/NTN networks, Environ. Monit. Assess., 62, 273–303, https://doi.org/10.1023/A:1006272609744, 2000.

Smyth, S. C., Jiang, W., Roth, H., Moran, M. D., Makar, P. A., Yang, F., Bouchet, V. S., and Landry, H.: A comparative performance evaluation of the AURAMS and CMAQ air-quality modelling systems, Atmos. Environ., 43, 1059–1070, https://doi.org/10.1016/j.atmosenv.2008.11.027, 2009.

Solazzo, E., Bianconi, R., Vautard, R., Appel, K. W., Moran, M. D., Hogrefe, C., Bessagnet, B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Denier Van Der Gon, H., Ferreira, J., Forkel, R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Jeričević, A., Kraljević, L., Miranda, A. I., Nopmongcol, U., Pirovano, G., Prank, M., Riccio, A., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S. T., and Galmarini, S.: Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII, Atmos. Environ., 53, 60–74, https://doi.org/10.1016/j.atmosenv.2012.01.003, 2012a.

Solazzo, E., Bianconi, R., Pirovano, G., Matthias, V., Vautard, R., Moran, M. D., Wyat Appel, K., Bessagnet, B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Ferreira, J., Forkel, R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Miranda, A. I., Nopmongcol, U., Prank, M., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S. T., and Galmarini, S.: Operational model evaluation for particulate matter in Europe and North America in the context of AQMEII, Atmos. Environ., 53, 75–92, https://doi.org/10.1016/j.atmosenv.2012.02.045, 2012b.

Solomon, P. A., Crumpler, D., Flanagan, J. B., Jayanty, R. K. M., Rickman, E. E., and McDade, C. E.: U.S. national $PM_{2.5}$ chemical speciation monitoring networks—CSN and IMPROVE: Description of networks, J. Air Waste Manag. Assoc., 64, 1410–1438, https://doi.org/10.1080/10962247.2014.956904, 2014.

Spicer, C. W., Buxton, B. E., Holdren, M. W., Smith, D. L., Kelly, T. J., Rust, S. W., Pate, A. D., Sverdrup, G. M., and Chuang, J. C.: Variability of hazardous air pollutants in an urban area, Atmos. Environ., 30, 3443–3456, https://doi.org/10.1016/1352-2310(95)00200-6, 1996.

Stanski, H. R., Wilson, L. J., and Burrows, W. R.: Survey of Common Verification Methods in Meteorology, World Weather Watch Technical Report No. 8, World Meteorological Organization, Geneva, 81 pp., https://elioscloud.wmo.int/share/s/KyB8rBXjSWeTXaRVUdPGqQ, 1989.

Steyn, D. G. and Galmarini, S.: Evaluating the predictive and explanatory value of atmospheric numerical models: Between relativism and objectivism, Open Atmospheric Sci. J., 2, 38–45, https://doi.org/10.2174/1874282300802010038, 2008.

Stroud, C. A., Morneau, G., Makar, P. A., Moran, M. D., Gong, W., Pabla, B., Zhang, J., Bouchet, V. S., Fox, D., Venkatesh, S., Wang, D., and Dann, T.: OH-reactivity of volatile organic compounds at urban and rural sites across Canada: Evaluation of air quality model predictions using speciated VOC measurements, Atmos. Environ., 42, 7746–7756, https://doi.org/10.1016/j.atmosenv.2008.05.054, 2008.

Stroud, C. A., Makar, P. A., Moran, M. D., Gong, W., Gong, S., Zhang, J., Hayden, K., Mihele, C., Brook, J. R., Abbatt, J. P. D., and Slowik, J. G.: Impact of model grid spacing on regional- and urban- scale air quality predictions of organic aerosol, Atmospheric Chem. Phys., 11, 3107–3118, https://doi.org/10.5194/acp-11-3107-2011, 2011.

Su, Y., Sofowote, U., Debosz, J., White, L., and Munoz, A.: Multi-year continuous $PM_{2.5}$ measurements with the Federal Equivalent Method SHARP 5030 and comparisons to filter-based and TEOM measurements in Ontario, Canada, Atmosphere, 9, 191, https://doi.org/10.3390/atmos9050191, 2018.

Swall, J. L. and Foley, K. M.: The impact of spatial correlation and incommensurability on model evaluation, Atmos. Environ., 43, 1204–1217, https://doi.org/10.1016/j.atmosenv.2008.10.057, 2009.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res. Atmospheres, 106, 7183–7192, https://doi.org/10.1029/2000JD900719, 2001.

Tesche, T. W., Morris, R., Tonnesen, G., McNally, D., Boylan, J., and Brewer, P.: CMAQ/CAMx annual 2002 performance evaluation over the eastern US, Atmos. Environ., 40, 4906–4919, https://doi.org/10.1016/j.atmosenv.2005.08.046, 2006.

Tessum, C. W., Hill, J. D., and Marshall, J. D.: Twelve-month, 12 km resolution North American WRF-Chem v3.4 air quality simulation: performance evaluation, Geosci. Model Dev., 8, 957–973, https://doi.org/10.5194/gmd-8-957-2015, 2015.

Thunis, P., Pederzoli, A., and Pernigotti, D.: Performance criteria to evaluate air quality modeling applications, Atmos. Environ., 59, 476–482, https://doi.org/10.1016/j.atmosenv.2012.05.043, 2012.

Toro, C., Foley, K., Simon, H., Henderson, B., Baker, K. R., Eyth, A., Timin, B., Appel, W., Luecken, D., Beardsley, M., Sonntag, D., Possiel, N., and Roberts, S.: Evaluation of 15 years of modeled atmospheric oxidized nitrogen compounds across the contiguous United States, Elem. Sci. Anthr., 9, 00158, https://doi.org/10.1525/elementa.2020.00158, 2021.

Toro, C., Sonntag, D., Bash, J., Burke, G., Murphy, B., Seltzer, K. M., Simon, H., Shephard, M., and Cady-Pereira, K. E.: Sensitivity of air quality to vehicle ammonia emissions in the United States, Atmos. Environ., 120484, https://doi.org/10.1016/j.atmosenv.2024.120484, 2024.

Van Loon, M., Vautard, R., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., Builtjes, P. J. H., Christensen, J. H., Cuvelier, C., Graff, A., Jonson, J. E., Krol, M., Langner, J., Roberts, P., Rouil, L., Stern, R., Tarrasón, L., Thunis, P., Vignati, E., White, L., and Wind, P.: Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble, Atmos. Environ., 41, 2083–2097, https://doi.org/10.1016/j.atmosenv.2006.10.073, 2007.

Vasilakos, P., Russell, A., Weber, R., and Nenes, A.: Understanding nitrate formation in a world with less sulfate, Atmospheric Chem. Phys., 18, 12765–12775, https://doi.org/10.5194/acp-18-12765-2018, 2018.

Vautard, R., Moran, M. D., Solazzo, E., Gilliam, R. C., Matthias, V., Bianconi, R., Chemel, C., Ferreira, J., Geyer, B., Hansen, A. B., Jericevic, A., Prank, M., Segers, A., Silver, J. D., Werhahn, J., Wolke, R., Rao, S. T., and Galmarini, S.: Evaluation of the meteorological forcing used for the Air Quality Model Evaluation International Initiative (AQMEII) air quality simulations, Atmos. Environ., 53, 15–37, https://doi.org/10.1016/j.atmosenv.2011.10.065, 2012.

Venkatram, A.: Inherent uncertainty in air quality modeling, Atmospheric Environ. 1967, 22, 1221–1227, https://doi.org/10.1016/0004-6981(88)90352-6, 1988.

Venkatram, A., Karamchandani, P. K., and Misra, P. K.: Testing a comprehensive acid deposition model, Atmospheric Environ. 1967, 22, 737–747, https://doi.org/10.1016/0004-6981(88)90011-X, 1988.

Vitali, L., Cuvelier, K., Piersanti, A., Monteiro, A., Adani, M., Amorati, R., Bartocha, A., D'Ausilio, A., Durka, P., Gama, C., Giovannini, G., Janssen, S., Przybyła, T., Stortini, M., Vranckx, S., and Thunis, P.: A standardized methodology for the validation of air quality forecast applications (F-MQO): lessons learnt from its application across Europe, Geosci. Model Dev., 16, 6029–6047, https://doi.org/10.5194/gmd-16-6029-2023, 2023.

Wagner, A., Blechschmidt, A.-M., Bouarar, I., Brunke, E.-G., Clerbaux, C., Cupeiro, M., Cristofanelli, P., Eskes, H., Flemming, J., Flentje, H., George, M., Gilge, S., Hilboll, A., Inness, A., Kapsomenakis, J., Richter, A., Ries, L., Spangl, W., Stein, O., Weller, R., and Zerefos, C.: Evaluation of the MACC operational forecast system – potential and challenges of global near-real-time modelling with respect to reactive gases in the troposphere, Atmospheric Chem. Phys., 15, 14005–14030, https://doi.org/10.5194/acp-15-14005-2015, 2015.

Wang, K., Zhang, Y., and Yahya, K.: Decadal application of WRF/Chem over the continental U.S.: Simulation design, sensitivity simulations, and climatological model evaluation, Atmos. Environ., 253, 118331, https://doi.org/10.1016/j.atmosenv.2021.118331, 2021.

Watson, J. G., Chow, J. C., Chen, L.-W. A., and Frank, N. H.: Methods to assess carbonaceous aerosol sampling artifacts for IMPROVE and other long-term networks, J. Air Waste Manag. Assoc., 59, 898–911, https://doi.org/10.3155/1047-3289.59.8.898, 2009.

Wayland, R. A., White, J. E., Dye, T. S., Anderson, C. B., Chan, A. C., and D.E.B. Strohm: Future of AirNow and the Air Quality Index: beyond ozone mapping and forecasting, Sixth Conference on Atmospheric Chemistry, 12-14 January, Seattle, Washington, American Meteorological Society, https://ams.confex.com/ams/pdfpapers/72556.pdf, 2004.

Wetherbee, G. A., Shaw, M. J., Latysh, N. E., Lehmann, C. M. B., and Rothert, J. E.: Comparison of precipitation chemistry measurements obtained by the Canadian Air and Precipitation Monitoring Network and National Atmospheric Deposition Program for the period 1995–2004, Environ. Monit. Assess., 164, 111–132, https://doi.org/10.1007/s10661-009-0879-8, 2010.

Widziewicz-Rzońca, K. and Tytła, M.: First systematic review on PM-bound water: exploring the existing knowledge domain using the CiteSpace software, Scientometrics, 124, 1945–2008, https://doi.org/10.1007/s11192-020-03547-w, 2020.

Williams, J. E. , Huijnen, V., Bouarar, I., Meziane, M., Schreurs, T., Pelletier, S., Marécal, V., Josse, B., and Flemming, J.: Regional evaluation of the performance of the global CAMS chemical modeling system over the United States (IFS cycle 47r1), Geosci. Model Dev., 15, 4657–4687, https://doi.org/10.5194/gmd-15-4657-2022, 2022.

Willmott, C. J.: On the validation of models, Phys. Geogr., 2, 184–194, https://doi.org/10.1080/02723646.1981.10642213, 1981.

WMO: Training Materials and Best Practices for Chemical Weather/Air Quality Forecasting, Report no. ETR-26, World Meteorological Organization, Geneva, 576 pp., https://library.wmo.int/doc_num.php?explnum_id=10439, 2020.

Yahya, K., Zhang, Y., and Vukovich, J. M.: 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, https://doi.org/10.1016/j.atmosenv.2014.04.024, 2014.

Yahya, K., Wang, K., Gudoshava, M., Glotfelty, T., and Zhang, Y.: Application of WRF/Chem over North America under the AQMEII Phase 2: Part I. Comprehensive evaluation of 2006 simulation, Atmos. Environ., 115, 733–755, https://doi.org/10.1016/j.atmosenv.2014.08.063, 2015.

Yu, S., Dennis, R., Roselle, S., Nenes, A., Walker, J., Eder, B., Schere, K., Swall, J., and Robarge, W.: An assessment of the ability of three-dimensional air quality models with current thermodynamic equilibrium models to predict aerosol NO_3^-, J. Geophys. Res., 110, D07S13, https://doi.org/10.1029/2004JD004718, 2005.

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

Zhai, H., Huang, L., Emery, C., Zhang, X., Wang, Y., Yarwood, G., Fu, J. S., and Li, L.: Recommendations on benchmarks for photochemical air quality model applications in China — NO_2, SO_2, CO and PM_{10}, Atmos. Environ., 319, 120290, https://doi.org/10.1016/j.atmosenv.2023.120290, 2024.

Zhang, H., Yee, L. D., Lee, B. H., Curtis, M. P., Worton, D. R., Isaacman-VanWertz, G., Offenberg, J. H., Lewandowski, M., Kleindienst, T. E., Beaver, M. R., Holder, A. L., Lonneman, W. A., Docherty, K. S., Jaoui, M., Pye, H. O. T., Hu, W., Day, D. A., Campuzano-Jost, P., Jimenez, J. L., Guo, H., Weber, R. J., De Gouw, J., Koss, A. R., Edgerton, E. S., Brune, W., Mohr, C., Lopez-Hilfiker, F. D., Lutz, A., Kreisberg, N. M., Spielman, S. R., Hering, S. V., Wilson, K. R., Thornton, J. A., and Goldstein, A. H.: Monoterpenes are the largest source of summertime organic aerosol in the southeastern United States, Proc. Natl. Acad. Sci., 115, 2038–2043, https://doi.org/10.1073/pnas.1717513115, 2018a.

Zhang, J., Moran, M. D., Zheng, Q., Makar, P. A., Baratzadeh, P., Marson, G., Liu, P., and Li, S.-M.: Emissions preparation and analysis for multiscale air quality modeling over the Athabasca Oil Sands Region of Alberta, Canada, Atmospheric Chem. Phys., 18, 10459–10481, https://doi.org/10.5194/acp-18-10459-2018, 2018b.

Zhang, L., Moran, M. D., Makar, P. A., Brook, J. R., and Gong, S.: Modelling gaseous dry deposition in AURAMS: a unified regional air-quality modelling system, Atmos. Environ., 36, 537–560, https://doi.org/10.1016/S1352-2310(01)00447-2, 2002.

Zhang, Y., Liu, P., Pun, B., and Seigneur, C.: A comprehensive performance evaluation of MM5-CMAQ for the summer 1999 southern oxidants study episode, Part III: Diagnostic and mechanistic evaluations, Atmos. Environ., 40, 4856–4873, https://doi.org/10.1016/j.atmosenv.2005.12.046, 2006a.

Zhang, Y., Liu, P., Pun, B., and Seigneur, C.: 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, https://doi.org/10.1016/j.atmosenv.2005.12.043, 2006b.

Zhang, Y., Liu, P., Queen, A., Misenis, C., Pun, B., Seigneur, C., and Wu, S.-Y.: A comprehensive performance evaluation of MM5-CMAQ for the Summer 1999 Southern Oxidants Study episode—Part II: Gas and aerosol predictions, Atmos. Environ., 40, 4839–4855, https://doi.org/10.1016/j.atmosenv.2005.12.048, 2006c.

Zhang, Y., Wen, X., Wang, K., Vijayaraghavan, K., and Jacobson, M. Z.: Probing into regional O_3 and particulate matter pollution in the United States: 2. An examination of formation mechanisms through a process analysis technique and sensitivity study, J. Geophys. Res. Atmospheres, 114, 2009JD011900, https://doi.org/10.1029/2009JD011900, 2009a.

Zhang, Y., Vijayaraghavan, K., Wen, X., Snell, H. E., and Jacobson, M. Z.: 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. Atmospheres, 114, 2009JD011898, https://doi.org/10.1029/2009JD011898, 2009b.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A.: Real-time air quality forecasting, part I: History, techniques, and current status, Atmos. Environ., 60, 632–655, https://doi.org/10.1016/j.atmosenv.2012.06.031, 2012a.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A.: Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects, Atmos. Environ., 60, 656–676, https://doi.org/10.1016/j.atmosenv.2012.02.041, 2012b.

Zhang, Y., Hong, C., Yahya, K., Li, Q., Zhang, Q., and He, K.: Comprehensive evaluation of multi-year real-time air quality forecasting using an online-coupled meteorology-chemistry model over southeastern United States, Atmos. Environ., 138, 162–182, https://doi.org/10.1016/j.atmosenv.2016.05.006, 2016.

Zhang, Y., Mathur, R., Bash, J. O., Hogrefe, C., Xing, J., and Roselle, S. J.: Long-term trends in total inorganic nitrogen and sulfur deposition in the US from 1990 to 2010, Atmospheric Chem. Phys., 18, 9091–9106, https://doi.org/10.5194/acp-18-9091-2018, 2018c.

Tables

Table 1. Comparison of Canadian, U.S., and Mexican annual anthropogenic and biogenic criteria-air-contaminant inventory emissions (tonnes/year) and model-ready anthropogenic and biogenic emissions for five years: 2013–2016 and 2021/22. Note that the U.S. inventory emissions are for all 50 U.S. states and other territories and the Mexican inventory emissions are for all 32 Mexican states. The "3 Country EIs" rows provide the sum of the three national emissions inventory amounts, while the "Total Anthro" rows correspond to the grid-limited, domain-total emissions that are input by the model after SMOKE emissions processing and TF scaling for fugitive PM emissions. The "Total Anthro" VOC emissions are the sum of 12 model VOC species, including EOTH (all unreactive or low-reactivity VOC species that are not considered by the gas-phase chemistry mechanism; see Moran et al, 2025). The "Total Biogenic" emissions depend on meteorology and are accumulated hourly predicted fields of soil NO emissions (in NO_2 units) and biogenic VOC emissions (as model VOC species) saved during each annual simulation.

		Year					Relative Difference (%)		
Species	**Region**	**2013**	**2014**	**2015**	**2016**	**2021/22**	**2016-to-2013**	**2021/22-to-2016**	**2021/22-to-2013**
SO2	Canada	1,245,629	1,196,925	1,067,133	1,050,048	720,937	-15.7	-31.3	-42.1
	U.S.A.	4,855,796	4,632,986	3,232,955	2,453,837	1,614,783	-49.5	-34.2	-66.7
	Mexico	1,907,119	1,955,040	1,955,789	1,956,538	1,963,766	2.6	0.4	3.0
	3 Country EIs	8,008,544	7,784,951	6,255,877	5,460,422	4,299,487	-31.8	-21.3	-46.3
	Total Anthro	6,788,380	6,566,599	5,008,720	4,248,067	2,683,021	-37.4	-36.8	-60.5
NOx	Canada	1,859,213	1,812,458	1,749,885	1,689,466	1,534,067	-9.1	-9.2	-17.5
	U.S.A.	12,010,496	11,318,521	10,317,943	9,281,137	7,462,553	-22.7	-19.6	-37.9
	Mexico	2,633,018	2,613,843	2,628,180	2,642,516	2,722,692	0.4	3.0	3.4
	3 Country EIs	16,502,727	15,744,822	14,696,007	13,613,119	11,719,312	-17.5	-13.9	-29.0
	Total Anthro	15,286,803	14,565,303	13,489,658	12,425,458	9,759,000	-18.7	-21.5	-36.2
	Total Biogenic	613,443	613,133	637,436	649,609	653,066	5.9	0.5	6.5
	Total Anthr+Bio	15,900,246	15,178,436	14,127,094	13,075,067	10,412,066	-17.8	-20.4	-34.5
VOC	Canada	1,639,307	1,676,389	1,624,563	1,530,449	1,517,407	-6.6	-0.9	-7.4
	U.S.A.	11,060,472	11,016,641	10,803,994	9,906,147	9,693,686	-10.4	-2.1	-12.4
	Mexico	4,223,883	4,246,882	4,247,239	4,247,596	4,619,106	0.6	8.7	9.4
	3 Country EIs	16,923,662	16,939,913	16,675,796	15,684,193	15,830,199	-7.3	0.9	-6.5
	Total Anthro	13,528,480	13,532,330	13,282,897	12,374,426	12,292,800	-8.5	-0.7	-9.1
	Total Biogenic	47,373,454	47,372,280	49,734,073	51,315,450	52,655,466	8.3	2.6	11.1
	Total Anthr+Bio	60,901,934	60,904,610	63,016,970	63,689,876	64,948,266	4.6	2.0	6.6
CO	Canada	5,414,122	5,310,926	5,227,752	5,173,958	4,322,475	-4.4	-16.5	-20.2
	U.S.A.	40,865,160	39,613,006	37,758,763	34,539,607	29,945,990	-15.5	-13.3	-26.7
	Mexico	8,738,655	8,863,635	8,867,142	8,870,649	9,015,675	1.5	1.6	3.2
	3 Country EIs	55,017,936	53,787,568	51,853,657	48,584,214	43,284,139	-11.7	-10.9	-21.3
	Total Anthro	48,394,003	47,088,320	45,142,220	41,929,513	36,118,000	-13.4	-13.9	-25.4
NH3	Canada	496,282	490,436	491,656	490,675	543,154	-1.1	10.7	9.4
	U.S.A.	3,814,960	3,723,552	3,839,459	3,853,365	3,481,215	1.0	-9.7	-8.7
	Mexico	840,629	844,080	844,297	844,514	841,397	0.5	-0.4	0.1
	3 Country EIs	5,151,871	5,058,068	5,175,412	5,188,554	4,865,766	0.7	-6.2	-5.6
	Total Anthro	4,521,443	4,424,069	4,541,247	4,553,502	4,234,533	0.7	-7.0	-6.3
PM2.5	Canada	817,212	814,398	793,598	780,950	804,186	-4.4	3.0	-1.6
	U.S.A.	3,421,551	3,580,536	3,287,999	3,359,590	3,614,926	-1.8	7.6	5.7
	Mexico	603,696	591,783	593,799	595,814	661,387	-1.3	11.0	9.6
	3 Country EIs	4,842,459	4,986,718	4,675,396	4,736,354	5,080,499	-2.2	7.3	4.9
	Total Anthro	3,101,173	3,076,508	2,932,475	2,855,582	2,889,000	-7.9	1.2	-6.8
PM10	Canada	3,548,631	3,543,931	3,515,237	3,495,289	3,746,529	-1.5	7.2	5.6
	U.S.A.	15,427,692	15,426,307	15,438,707	15,532,333	18,729,161	0.7	20.6	21.4
	Mexico	839,386	822,182	824,358	826,535	923,082	-1.5	11.7	10.0
	3 Country EIs	19,815,709	19,792,420	19,778,302	19,854,158	23,398,772	0.2	17.9	18.1
	Total Anthro	10,122,137	10,094,335	9,989,971	10,001,543	10,826,000	-1.2	8.2	7.0

Table 2. Number of U.S. and Canadian measurement stations with complete hourly NO_2, O_3, and $PM_{2.5}$ measurements
2550 for an annual evaluation vs. available measurements by network and year for 2013–2016 (AQS, NAPS) and 2021/22 (AirNow, NAPS).

Year	2013		2014		2015		2016		2021/22	
	Complete	Available	Complete	Available	Complete	Available	Complete	Available	Complete	Available
Variable										
				AQS					AirNow US	
NO2	327	405	338	412	341	412	331	402	162	193
O3	720	1310	733	1298	716	1285	755	1274	580	1085
PM2.5	601	777	623	800	621	832	614	826	603	770
				NAPS					NAPS	
NO2	132	141	131	146	138	157	142	159	143	175
O3	171	187	175	192	174	193	180	200	175	209
PM2.5	158	189	179	196	165	198	174	198	173	203

Table 3. Summary table of all-station annual statistics for hourly NO_2, O_3, and $PM_{2.5}$ surface measurements for 2021/22 and 2013–2016. For dimensional statistics, units are ppbv for NO_2 and O_3 and μg·m^{-3} for $PM_{2.5}$.

		AirNow Combined	AQS + NAPS Combined			
Variable	Statistic	2021/22	2013	2014	2015	2016
NO_2	N	2,509,716	3,776,292	3,856,650	3,933,509	3,910,073
	Obs Mean	6.57	7.90	7.74	7.60	7.18
	Model Mean	5.30	8.80	8.53	8.32	7.74
	MB	-1.27	0.90	0.78	0.73	0.57
	NMB	-0.19	0.11	0.10	0.10	0.08
	RMSE	6.09	7.21	7.11	7.00	6.66
	CRMSE	5.96	7.15	7.07	6.96	6.63
	NMAE	0.56	0.58	0.58	0.58	0.58
	Fac2	0.52	0.59	0.58	0.58	0.57
	R	0.65	0.68	0.68	0.68	0.68
	NSD	0.84	1.05	1.05	1.06	1.05
	Obs SD	7.61	8.63	8.59	8.40	8.08
	Model SD	6.37	9.10	9.00	8.93	8.48
O_3	N	6,175,254	7,468,605	7,617,067	7,472,330	7,871,784
	Obs Mean	29.64	29.50	29.14	28.95	29.42
	Model Mean	27.60	25.87	26.03	26.34	26.68
	MB	-2.04	-3.62	-3.11	-2.61	-2.75
	NMB	-0.07	-0.12	-0.11	-0.09	-0.09
	RMSE	10.77	12.15	11.66	11.60	11.35
	CRMSE	10.57	11.59	11.24	11.30	11.01
	NMAE	0.28	0.31	0.31	0.30	0.29
	Fac2	0.83	0.78	0.79	0.79	0.80
	R	0.72	0.71	0.71	0.71	0.72
	NSD	0.88	0.99	0.99	0.98	0.97
	Obs SD	14.76	15.34	14.97	15.03	14.89
	Model SD	13.06	15.18	14.76	14.80	14.51
$PM_{2.5}$	N	6,374,875	6,280,280	6,635,531	6,471,206	6,471,336
	Obs Mean	8.02	8.67	8.34	8.37	7.44
	Model Mean	5.51	8.17	7.87	7.59	6.98
	MB	-2.51	-0.50	-0.46	-0.78	-0.46
	NMB	-0.31	-0.06	-0.06	-0.09	-0.06
	RMSE	9.73	11.62	11.33	11.39	13.43
	CRMSE	9.41	11.61	11.32	11.36	13.42
	NMAE	0.66	0.71	0.72	0.71	0.73
	Fac2	0.46	0.50	0.49	0.49	0.48
	R	0.24	0.29	0.27	0.26	0.17
	NSD	0.69	1.36	1.35	1.13	0.78
	Obs SD	8.78	8.11	7.83	8.73	11.54
	Model SD	6.06	11.02	10.60	9.89	9.05

Table 4. Summary table of all-station annual statistics for nine other ambient gas-phase chemistry measurements (NO, NO_x, CO, HNO_3, NH_3, SO_2, ETHE, ISOP, HCHO) for 2013–2016. For dimensional statistics, units are ppmv for CO, $\mu g \cdot m^{-3}$ for NH_3, and ppbv for other species. Sample duration used for all networks is hourly for NO, NO_x, CO, ETHE, and ISOP, daily for HCHO, weekly for HNO_3 and SO_2, and biweekly for NH_3.

Variable	Statistic	2013	2014	2015	2016
NO	N	3,562,348	3,721,353	3,680,583	3,564,269
	Obs Mean	4.09	3.89	4.01	3.61
	Model Mean	5.39	4.88	4.68	3.93
	MB	1.30	0.99	0.68	0.32
	NMB	0.32	0.25	0.17	0.09
	RMSE	16.44	15.21	14.65	12.90
	CRMSE	16.39	15.18	14.63	12.90
	NMAE	1.34	1.29	1.26	1.21
	FAC2	0.28	0.29	0.28	0.28
	R	0.41	0.42	0.41	0.41
	NSD	1.23	1.21	1.13	1.07
	Obs SD	13.36	12.61	12.60	11.46
	Model SD	16.49	15.30	14.20	12.26
NO_x	N	3,408,562	3,507,613	3,246,050	3,476,220
	Obs Mean	12.03	11.72	11.86	10.86
	Model Mean	14.38	13.40	12.69	11.67
	MB	2.35	1.68	0.83	0.82
	NMB	0.20	0.14	0.07	0.08
	RMSE	21.12	19.72	18.58	17.13
	CRMSE	20.98	19.65	18.56	17.11
	NMAE	0.79	0.77	0.73	0.74
	FAC2	0.53	0.53	0.53	0.53
	R	0.54	0.54	0.55	0.55
	NSD	1.19	1.15	1.04	1.05
	Obs SD	19.81	19.00	19.16	17.60
	Model SD	23.60	21.84	19.91	18.54
HNO_3	N	4,766	4,946	4,874	4,788
	Obs Mean	0.25	0.24	0.23	0.22
	Model Mean	0.33	0.32	0.29	0.25
	MB	0.08	0.08	0.06	0.04
	NMB	0.33	0.32	0.25	0.16
	RMSE	0.19	0.19	0.18	0.14
	CRMSE	0.18	0.18	0.17	0.14
	NMAE	0.52	0.53	0.48	0.42
	FAC2	0.76	0.75	0.78	0.81
	R	0.69	0.64	0.66	0.69
	NSD	1.28	1.24	1.30	1.12
	Obs SD	0.19	0.18	0.17	0.16
	Model SD	0.24	0.22	0.22	0.18
NH_3	N	1,705	1,731	2,417	2,464
	Obs Mean	1.54	1.71	1.62	1.73
	Model Mean	0.87	0.91	0.94	1.00
	MB	-0.67	-0.81	-0.68	-0.73
	NMB	-0.44	-0.47	-0.42	-0.42
	RMSE	2.54	2.86	3.10	2.66
	CRMSE	2.45	2.75	3.02	2.56
	NMAE	0.60	0.62	0.60	0.57

	FAC2	0.50	0.47	0.54	0.57
	R	0.34	0.51	0.44	0.51
	NSD	0.44	0.41	0.36	0.44
	Obs SD	2.59	3.17	3.36	2.97
	Model SD	1.13	1.30	1.21	1.29
SO_2	N	27,172	28,170	28,311	27,279
	Obs Mean	1.23	1.18	0.99	0.83
	Model Mean	1.96	1.78	1.42	1.22
	MB	0.74	0.60	0.43	0.39
	NMB	0.60	0.51	0.43	0.47
	RMSE	2.60	2.36	1.97	1.98
	CRMSE	2.49	2.28	1.92	1.94
	NMAE	1.11	1.00	0.99	1.10
	FAC2	0.45	0.49	0.47	0.45
	R	0.35	0.39	0.40	0.38
	NSD	1.43	1.35	1.29	1.37
	Obs SD	1.75	1.72	1.51	1.43
	Model SD	2.50	2.31	1.94	1.97
CO	N	2,413,310	2,313,412	2,263,157	2,207,060
	Obs Mean	0.27	0.27	0.27	0.28
	Model Mean	0.31	0.30	0.29	0.27
	MB	0.03	0.03	0.01	-0.01
	NMB	0.12	0.09	0.05	-0.05
	RMSE	0.29	0.27	0.25	0.56
	CRMSE	0.28	0.26	0.25	0.56
	NMAE	0.61	0.57	0.55	0.55
	FAC2	0.69	0.72	0.73	0.74
	R	0.42	0.43	0.43	0.14
	NSD	1.04	1.09	1.04	0.38
	Obs SD	0.26	0.24	0.23	0.55
	Model SD	0.27	0.26	0.24	0.21
ETHE	N	95,968	94,497	73,794	57,832
	Obs Mean	1.05	1.03	1.06	0.95
	Model Mean	1.77	1.80	2.01	2.08
	MB	0.72	0.77	0.95	1.13
	NMB	0.68	0.75	0.89	1.19
	RMSE	2.25	2.34	2.55	2.82
	CRMSE	2.13	2.21	2.36	2.58
	NMAE	1.24	1.27	1.42	1.66
	FAC2	0.41	0.42	0.39	0.35
	R	0.32	0.26	0.20	0.18
	NSD	1.05	1.08	1.12	1.13
	Obs SD	1.78	1.75	1.75	1.89
	Model SD	1.87	1.89	1.97	2.13
HCHO	N	4,909	5,063	4,889	4,826
	Obs Mean	2.38	2.07	2.27	2.37
	Model Mean	2.47	2.32	2.43	2.56
	MB	0.09	0.25	0.16	0.19
	NMB	0.04	0.12	0.07	0.08
	RMSE	3.71	2.13	2.09	2.34
	CRMSE	3.71	2.11	2.08	2.33
	NMAE	0.66	0.61	0.56	0.61

	FAC2	0.65	0.66	0.71	0.66
	R	0.26	0.43	0.46	0.41
	NSD	0.64	1.24	1.23	1.31
	Obs SD	3.59	1.75	1.78	1.83
	Model SD	2.28	2.17	2.19	2.39
ISOP	N	79,679	94,818	66,013	42,845
	Obs Mean	0.12	0.11	0.15	0.13
	Model Mean	0.39	0.38	0.56	0.48
	MB	0.27	0.27	0.41	0.35
	NMB	2.27	2.36	2.71	2.60
	RMSE	0.77	0.78	0.99	0.99
	CRMSE	0.73	0.73	0.90	0.93
	NMAE	2.81	2.85	3.12	3.20
	FAC2	0.18	0.19	0.18	0.19
	R	0.45	0.46	0.50	0.37
	NSD	3.37	3.25	3.13	3.69
	Obs SD	0.24	0.25	0.32	0.27
	Model SD	0.80	0.82	1.02	0.99

Table 5. Summary table of all-station annual statistics for daily gravimetric and speciated $PM_{2.5}$ measurements for 2013–2016. For dimensional statistics, units are μg·m^{-3}.

Variable	Statistic	2013	2014	2015	2016
SO4	N	32,101	32,027	30,103	28,972
	Obs Mean	1.17	1.15	0.97	0.75
	Model Mean	0.84	0.82	0.73	0.63
	MB	-0.32	-0.33	-0.25	-0.12
	NMB	-0.28	-0.29	-0.25	-0.16
	RMSE	0.82	0.83	0.73	0.58
	CRMSE	0.75	0.76	0.69	0.57
	NMAE	0.46	0.46	0.47	0.49
	Fac2	0.61	0.64	0.65	0.64
	R	0.71	0.69	0.65	0.61
	NSD	0.83	0.75	0.74	0.84
	Obs SD	1.05	1.05	0.90	0.69
	Model SD	0.87	0.79	0.67	0.57
NO3	N	31,992	31,983	30,046	29,029
	Obs Mean	0.75	0.79	0.68	0.54
	Model Mean	0.67	0.61	0.58	0.49
	MB	-0.08	-0.19	-0.10	-0.05
	NMB	-0.11	-0.23	-0.15	-0.09
	RMSE	1.40	1.25	1.00	0.91
	CRMSE	1.40	1.23	1.00	0.91
	NMAE	0.67	0.63	0.64	0.71
	Fac2	0.37	0.35	0.34	0.32
	R	0.57	0.66	0.70	0.66
	NSD	0.80	0.71	0.85	0.88
	Obs SD	1.64	1.63	1.36	1.17
	Model SD	1.32	1.16	1.16	1.02
NH4	N	14,828	14,711	13,283	12,757
	Obs Mean	0.64	0.69	0.54	0.33
	Model Mean	0.78	0.73	0.68	0.60
	MB	0.14	0.04	0.14	0.28
	NMB	0.22	0.06	0.26	0.85
	RMSE	0.75	0.69	0.61	0.60
	CRMSE	0.73	0.69	0.60	0.54
	NMAE	0.66	0.59	0.72	1.23
	Fac2	0.58	0.60	0.52	0.34
	R	0.52	0.57	0.58	0.51
	NSD	0.81	0.71	0.82	0.92
	Obs SD	0.81	0.83	0.71	0.56
	Model SD	0.66	0.59	0.58	0.52
EC	N	30,770	31,862	28,866	28,114
	Obs Mean	0.34	0.33	0.32	0.30
	Model Mean	0.52	0.50	0.47	0.40
	MB	0.18	0.17	0.15	0.09
	NMB	0.53	0.51	0.48	0.30
	RMSE	0.74	0.66	0.67	0.55
	CRMSE	0.71	0.64	0.65	0.55
	NMAE	0.91	0.91	0.90	0.77
	Fac2	0.57	0.58	0.56	0.56
	R	0.63	0.60	0.58	0.60
	NSD	2.11	1.94	1.91	1.68
	Obs SD	0.43	0.41	0.42	0.41
	Model SD	0.90	0.80	0.80	0.69

OM	N	30,609	31,747	28,688	27,508
	Obs Mean	2.48	2.43	2.76	2.46
	Model Mean	2.74	2.54	2.61	2.41
	MB	0.26	0.11	-0.15	-0.05
	NMB	0.11	0.04	-0.05	-0.02
	RMSE	6.49	3.54	6.24	10.33
	CRMSE	6.48	3.54	6.24	10.33
	NMAE	0.75	0.68	0.72	0.74
	Fac2	0.54	0.55	0.51	0.50
	R	0.36	0.45	0.25	0.10
	NSD	2.52	1.44	1.54	0.54
	Obs SD	2.76	2.65	3.86	9.50
	Model SD	6.95	3.82	5.95	5.14
CM	N	31,429	31,675	29,871	28,902
	Obs Mean	0.58	0.63	0.61	0.56
	Model Mean	0.83	0.85	0.83	0.83
	MB	0.25	0.22	0.22	0.27
	NMB	0.42	0.35	0.36	0.48
	RMSE	1.60	1.67	1.53	1.46
	CRMSE	1.58	1.65	1.52	1.43
	NMAE	1.40	1.37	1.30	1.36
	Fac2	0.31	0.30	0.32	0.32
	R	0.12	0.13	0.17	0.22
	NSD	1.30	1.19	1.27	1.59
	Obs SD	1.02	1.14	1.03	0.85
	Model SD	1.33	1.36	1.30	1.35
SS	N	31,860	31,852	30,185	29,104
	Obs Mean	0.26	0.25	0.29	0.23
	Model Mean	0.45	0.44	0.49	0.51
	MB	0.19	0.19	0.20	0.29
	NMB	0.74	0.78	0.71	1.25
	RMSE	1.02	1.12	1.09	1.16
	CRMSE	1.00	1.10	1.07	1.12
	NMAE	1.36	1.42	1.35	1.70
	Fac2	0.31	0.32	0.33	0.33
	R	0.62	0.58	0.61	0.60
	NSD	2.55	2.74	2.66	3.03
	Obs SD	0.48	0.48	0.48	0.44
	Model SD	1.23	1.31	1.29	1.33
PM2.5	N	104,944	104,095	111,150	102,120
	Obs Mean	8.19	8.12	7.94	7.24
	Model Mean	8.72	8.10	8.10	7.47
	MB	0.54	-0.02	0.16	0.23
	NMB	0.07	0.00	0.02	0.03
	RMSE	8.24	7.86	7.65	7.15
	CRMSE	8.22	7.86	7.65	7.14
	NMAE	0.54	0.51	0.53	0.54
	Fac2	0.70	0.70	0.70	0.68
	R	0.47	0.44	0.45	0.43
	NSD	1.61	1.50	1.41	1.53
	Obs SD	5.69	5.67	5.81	5.04
	Model SD	9.19	8.49	8.20	7.70

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

Table 6. Summary table of all-station annual statistics for weekly precipitation-chemistry measurements for 2013–2016. For dimensional statistics, units are mg·L^{-1} for concentration in precipitation, mm·week^{-1} for precipitation, and 2575 mg·m^{-2}·week^{-1} for wet deposition. Note that daily CAPMoN measurements have been aggregated to weekly values for consistency with NADP measurements.

Variable	Statistic	2013	2014	2015	2016
SO4 conc	N	9,115	8,997	9,759	9,193
	Obs Mean	0.81	0.74	0.65	0.62
	Model Mean	0.82	0.84	0.74	0.61
	MB	0.02	0.10	0.09	-0.01
	NMB	0.02	0.14	0.14	-0.01
	RMSE	0.91	0.91	0.87	1.19
	CRMSE	0.91	0.90	0.87	1.19
	NMAE	0.59	0.65	0.68	0.61
	Fac2	0.64	0.65	0.63	0.65
	R	0.41	0.38	0.37	0.25
	NSD	0.89	1.00	1.09	0.50
	Obs SD	0.88	0.81	0.74	1.19
	Model SD	0.78	0.81	0.80	0.60
NO3 conc	N	9,115	8,995	9,752	9,190
	Obs Mean	0.96	0.93	0.91	0.90
	Model Mean	0.94	0.91	0.90	0.78
	MB	-0.02	-0.01	-0.01	-0.12
	NMB	-0.02	-0.02	-0.01	-0.14
	RMSE	0.97	0.93	1.02	0.94
	CRMSE	0.97	0.93	1.02	0.93
	NMAE	0.55	0.54	0.57	0.52
	Fac2	0.69	0.70	0.68	0.69
	R	0.49	0.52	0.48	0.46
	NSD	1.06	1.03	1.06	0.80
	Obs SD	0.92	0.94	0.97	0.98
	Model SD	0.98	0.97	1.03	0.78
NH4 conc	N	9,103	8,980	9,727	9,181
	Obs Mean	0.37	0.36	0.39	0.39
	Model Mean	0.46	0.46	0.54	0.52
	MB	0.09	0.09	0.16	0.13
	NMB	0.23	0.25	0.41	0.35
	RMSE	0.65	0.65	0.91	0.79
	CRMSE	0.65	0.65	0.90	0.78
	NMAE	0.78	0.80	0.95	0.89
	Fac2	0.60	0.59	0.57	0.58
	R	0.47	0.41	0.38	0.39
	NSD	1.62	1.54	1.98	1.75
	Obs SD	0.45	0.44	0.49	0.47
	Model SD	0.72	0.68	0.96	0.83
PR	N	13,732	13,874	14,050	13,403
	Obs Mean	19.48	20.27	19.77	19.38
	Model Mean	21.72	21.64	19.63	20.11
	MB	2.25	1.37	-0.14	0.73
	NMB	0.12	0.07	-0.01	0.04
	RMSE	19.62	20.63	17.60	18.91
	CRMSE	19.49	20.58	17.60	18.90
	NMAE	0.54	0.53	0.49	0.52
	Fac2	0.57	0.57	0.57	0.56
	R	0.72	0.71	0.78	0.75
	NSD	1.07	1.06	0.99	1.03

	Obs SD	25.24	26.33	26.65	26.40
	Model SD	26.91	27.90	26.52	27.18
SO4 dep	N	9,115	8,997	9,759	9,193
	Obs Mean	15.65	14.41	11.41	10.71
	Model Mean	15.91	15.18	11.36	10.08
	MB	0.26	0.78	-0.06	-0.63
	NMB	0.02	0.05	0.00	-0.06
	RMSE	16.29	16.85	12.91	12.58
	CRMSE	16.29	16.83	12.91	12.56
	NMAE	0.61	0.65	0.65	0.62
	Fac2	0.55	0.56	0.53	0.56
	R	0.58	0.49	0.53	0.58
	NSD	0.90	0.85	0.83	0.76
	Obs SD	18.55	17.85	14.33	15.04
	Model SD	16.72	15.09	11.84	11.38
NO3 dep	N	9,115	8,995	9,752	9,190
	Obs Mean	16.48	15.64	13.71	14.12
	Model Mean	16.16	14.76	12.35	11.65
	MB	-0.33	-0.88	-1.35	-2.47
	NMB	-0.02	-0.06	-0.10	-0.18
	RMSE	14.46	16.49	12.16	13.00
	CRMSE	14.46	16.47	12.09	12.76
	NMAE	0.54	0.54	0.55	0.52
	Fac2	0.63	0.64	0.62	0.62
	R	0.59	0.46	0.56	0.56
	NSD	0.89	0.72	0.85	0.72
	Obs SD	16.84	17.81	13.76	15.13
	Model SD	14.97	12.86	11.68	10.93
NH4 dep	N	9,103	8,980	9,727	9,181
	Obs Mean	6.60	6.26	5.85	6.14
	Model Mean	7.65	7.28	6.75	6.94
	MB	1.05	1.02	0.90	0.80
	NMB	0.16	0.16	0.15	0.13
	RMSE	8.88	9.54	7.78	8.33
	CRMSE	8.82	9.48	7.73	8.29
	NMAE	0.69	0.72	0.72	0.68
	Fac2	0.56	0.57	0.55	0.57
	R	0.59	0.47	0.55	0.56
	NSD	1.08	1.01	1.05	1.06
	Obs SD	9.32	9.20	7.90	8.53
	Model SD	10.06	9.29	8.33	9.03

Table 7. Comparison of RAQDPS023 and RAQDPS-FW023 all-station seasonal scores for predicted hourly surface NO_2, O_3, and $PM_{2.5}$ abundances for 2021/22. For dimensional statistics, units are ppbv for NO_2 and O_3 and µg·m^{-3} for $PM_{2.5}$.

		Winter		Spring		Summer		Autumn	
Variable	Statistic	OP023	FW023	OP023	FW023	OP023	FW023	OP023	FW023
NO2	N	618,274	618,274	635,685	635,685	633,977	633,977	621,780	621,780
	Obs Mean	9.42	9.42	5.53	5.53	4.57	4.57	6.82	6.82
	Model	7.32	7.32	4.50	4.50	3.98	4.02	5.45	5.46
	MB	-2.10	-2.10	-1.03	-1.03	-0.59	-0.55	-1.38	-1.37
	NMB	-0.22	-0.22	-0.19	-0.19	-0.13	-0.12	-0.20	-0.20
	RMSE	7.69	7.69	5.69	5.69	4.62	4.64	6.00	6.00
	CRMSE	7.40	7.40	5.60	5.60	4.59	4.60	5.84	5.84
	NMAE	0.52	0.52	0.60	0.60	0.60	0.61	0.55	0.55
	Fac2	0.56	0.56	0.48	0.48	0.49	0.50	0.54	0.54
	R	0.66	0.66	0.62	0.62	0.56	0.56	0.64	0.64
	NSD	0.85	0.85	0.87	0.87	0.89	0.90	0.78	0.79
	Obs SD	9.54	9.54	6.76	6.76	5.14	5.14	7.49	7.49
	Model SD	8.14	8.14	5.89	5.89	4.57	4.62	5.88	5.89
O3	N	1,523,37	1,523,37	1,567,17	1,567,17	1,570,21	1,570,21	1,514,48	1,514,48
	Obs Mean	26.37	26.37	34.78	34.78	31.34	31.34	25.83	25.83
	Model	27.56	27.57	30.59	30.61	27.09	27.93	25.07	25.31
	MB	1.19	1.19	-4.19	-4.17	-4.25	-3.41	-0.76	-0.52
	NMB	0.05	0.05	-0.12	-0.12	-0.14	-0.11	-0.03	-0.02
	RMSE	9.65	9.65	11.02	11.01	12.36	11.95	9.76	9.67
	CRMSE	9.57	9.57	10.19	10.19	11.60	11.46	9.73	9.66
	NMAE	0.28	0.28	0.25	0.25	0.31	0.30	0.28	0.28
	Fac2	0.82	0.82	0.88	0.88	0.82	0.83	0.81	0.82
	R	0.68	0.68	0.66	0.66	0.75	0.76	0.75	0.76
	NSD	0.88	0.88	0.93	0.93	0.90	0.93	0.89	0.91
	Obs SD	12.50	12.50	12.82	12.82	16.94	16.94	14.42	14.42
	Model SD	11.06	11.06	11.92	11.93	15.32	15.79	12.88	13.12
PM2.5	N	1,574,82	1,574,82	1,616,88	1,616,88	1,622,57	1,622,57	1,560,59	1,560,59
	Obs Mean	8.03	8.03	6.20	6.20	10.39	10.39	7.41	7.41
	Model	7.00	7.04	4.56	4.69	5.06	10.82	5.47	7.06
	MB	-1.04	-0.99	-1.64	-1.52	-5.34	0.42	-1.94	-0.35
	NMB	-0.13	-0.12	-0.26	-0.24	-0.51	0.04	-0.26	-0.05
	RMSE	8.38	8.37	5.93	5.88	14.14	19.92	8.48	10.26
	CRMSE	8.32	8.31	5.70	5.68	13.09	19.92	8.25	10.25
	NMAE	0.66	0.66	0.63	0.62	0.69	0.65	0.65	0.64
	Fac2	0.49	0.49	0.46	0.47	0.43	0.56	0.48	0.54
	R	0.40	0.41	0.37	0.38	0.09	0.55	0.28	0.49
	NSD	1.16	1.17	0.94	0.96	0.35	1.87	0.76	1.50
	Obs SD	7.00	7.00	5.21	5.21	12.74	12.74	7.68	7.68
	Model SD	8.14	8.17	4.91	5.01	4.46	23.79	5.83	11.54

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

Figures

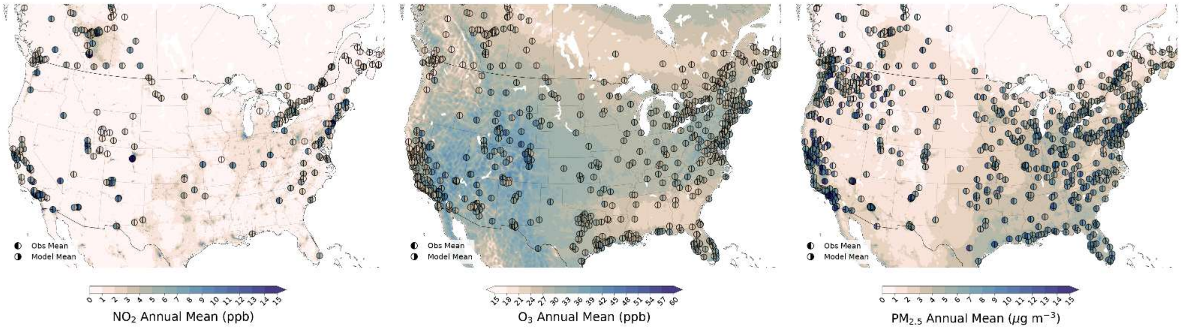
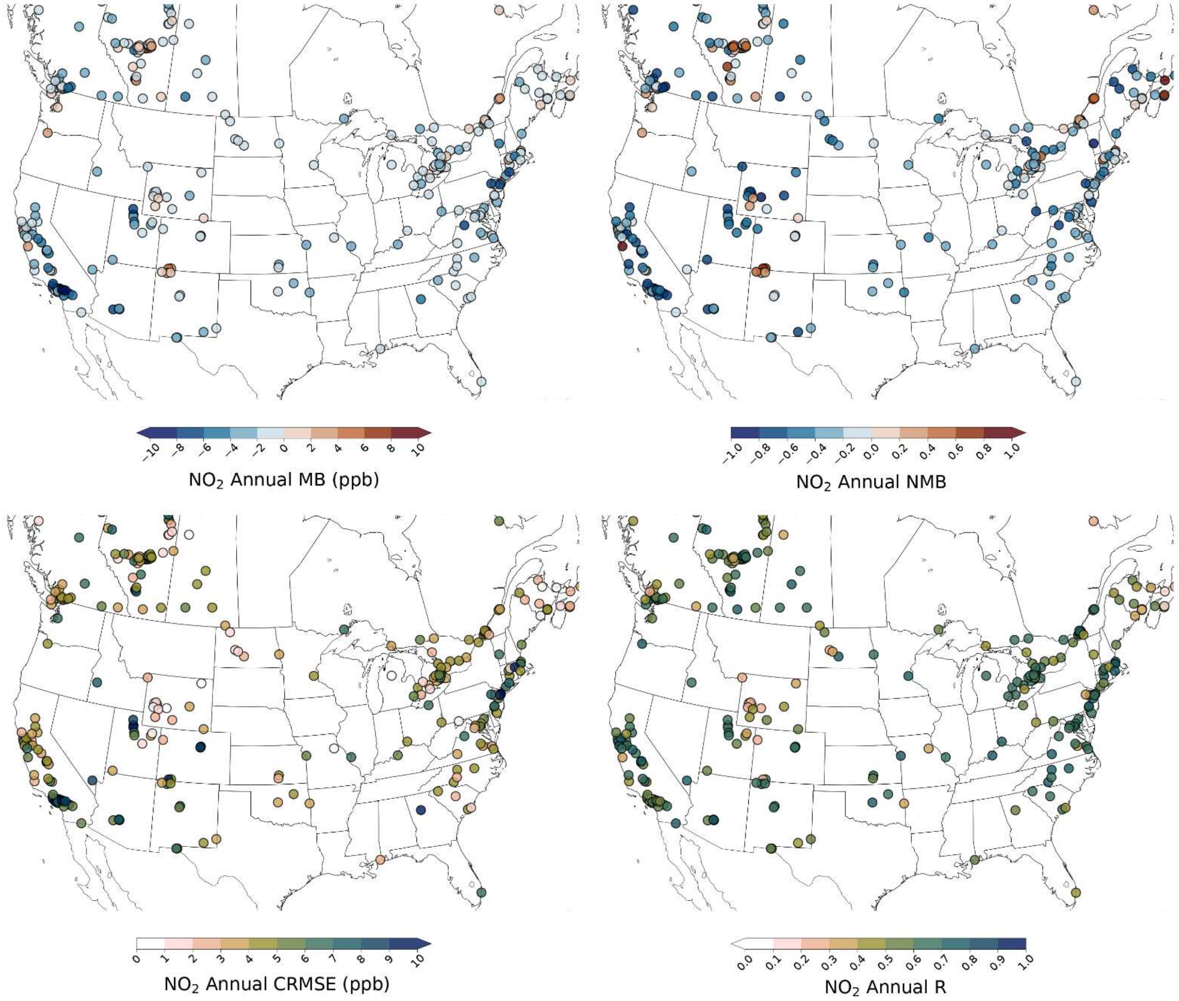
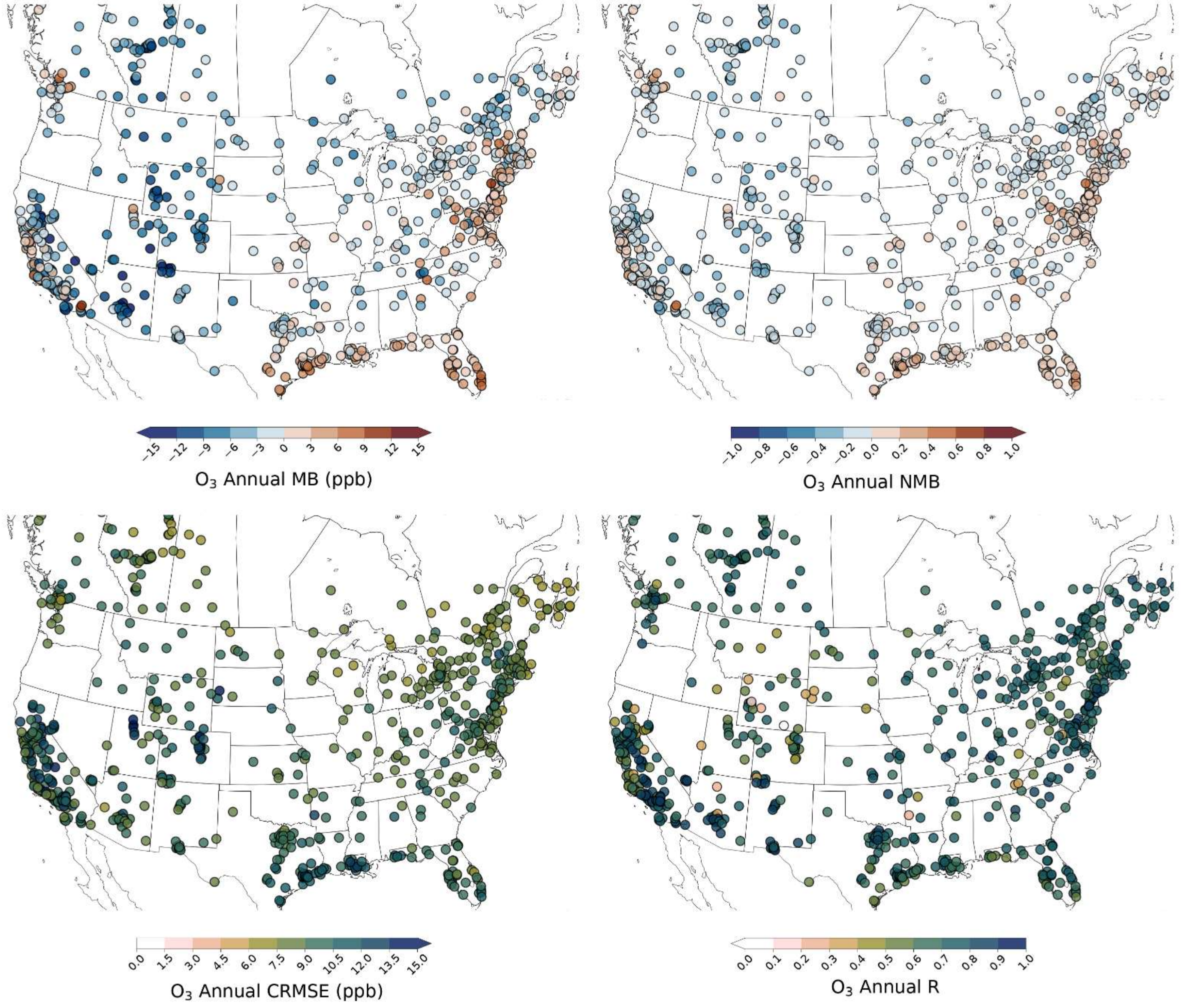


Figure 1. Spatial distribution of predicted 2021/22 mean annual abundance fields of hourly (a) NO_2, (b) O_3, and (c) $PM_{2.5}$-noSS with observed and predicted NRT station annual abundances superimposed (shown as filled-in divided circles, same colour bar). Units are ppbv, ppbv, and $\mu g \cdot m^{-3}$, respectively.

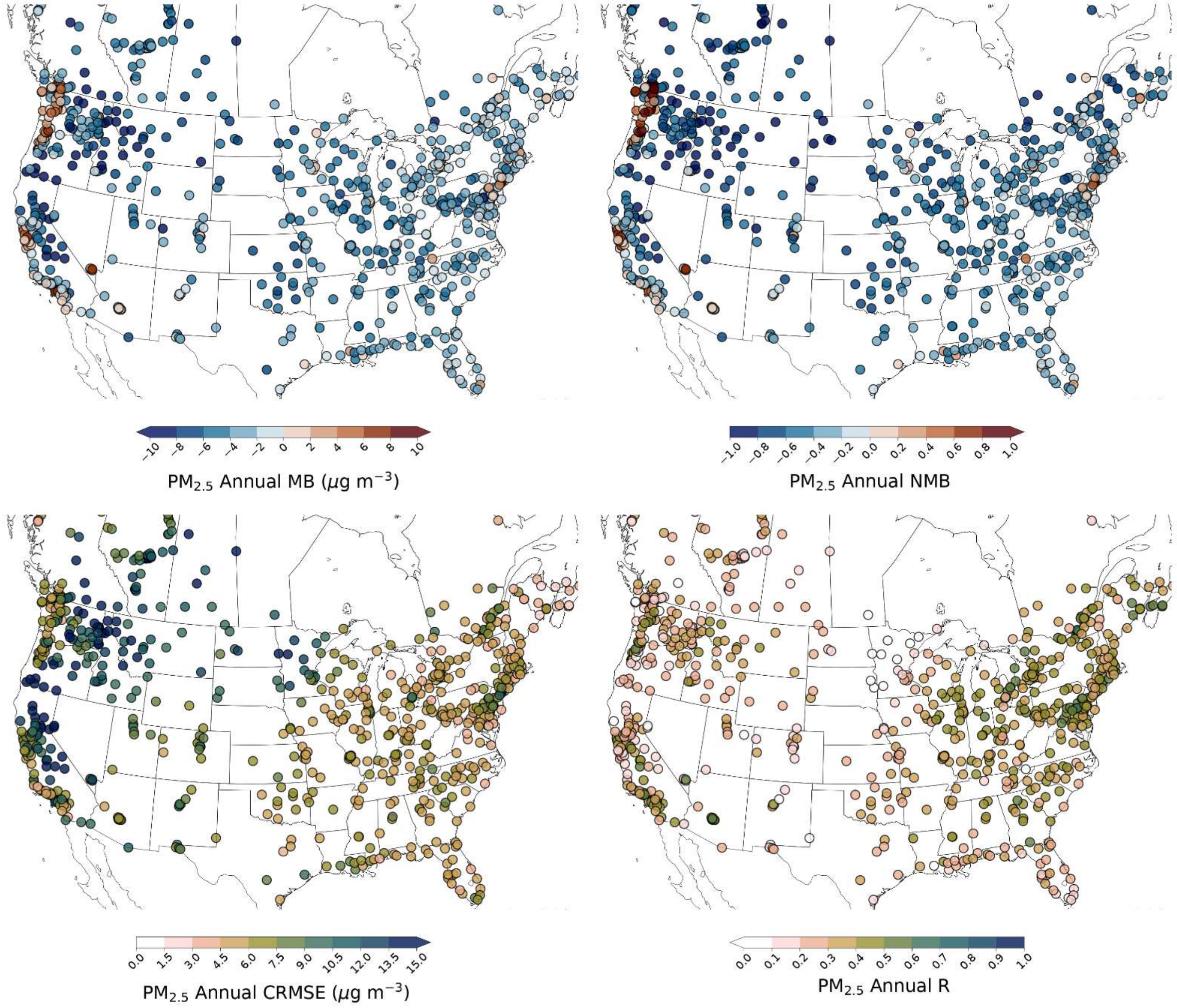


2595 Figure 2. Spatial distribution of 2021/22 annual (a) MB, (b) NMB, (c) CRMSE, and (d) R scores at all NRT stations for hourly NO_2 measurements (ppbv).

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025



2600 Figure 3. Spatial distribution of 2021/22 annual (a) MB, (b) NMB, (c) CRMSE, and (d) R scores at all NRT stations for hourly O_3 measurements (ppbv).



2605 Figure 4. Spatial distribution of 2021/22 annual (a) MB, (b) NMB, (c) CRMSE, and (d) R scores at all NRT stations for hourly PM$_{2.5}$ measurements (µg·m^{-3}).

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

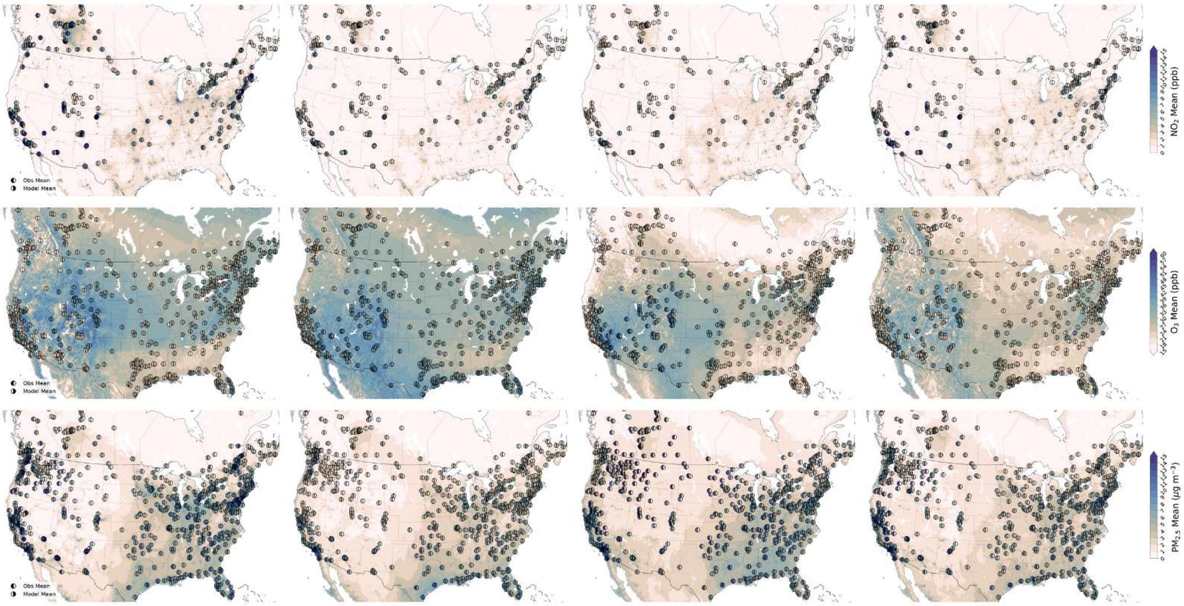


Figure 5. Spatial distribution of predicted 2021/22 mean seasonal abundance fields (from left to right: winter [DJF], spring [MAM], summer [JJA], autumn [SON]) of hourly (top row) NO_2, (middle row) O_3, and (bottom row) $PM_{2.5}$-noSS with observed and predicted station seasonal abundances superimposed (shown as filled-in divided circles, same colour bar). Units are ppbv, ppbv, and μg·m^{-3}, respectively.

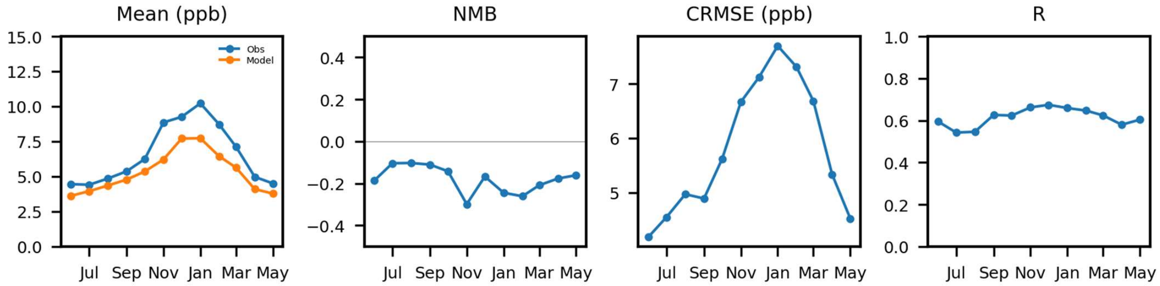
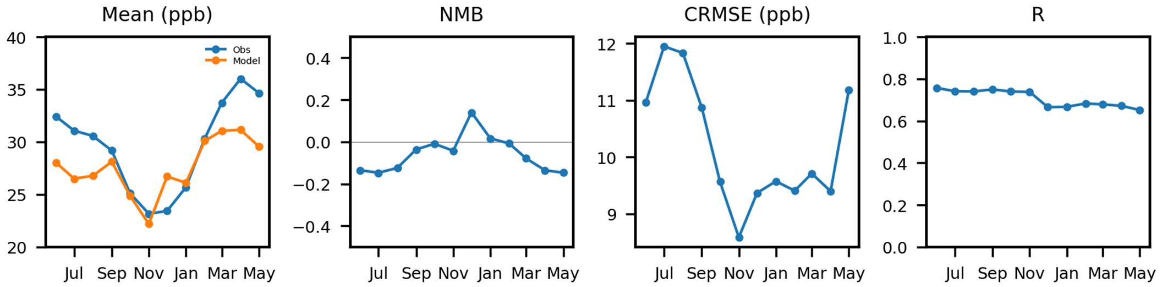


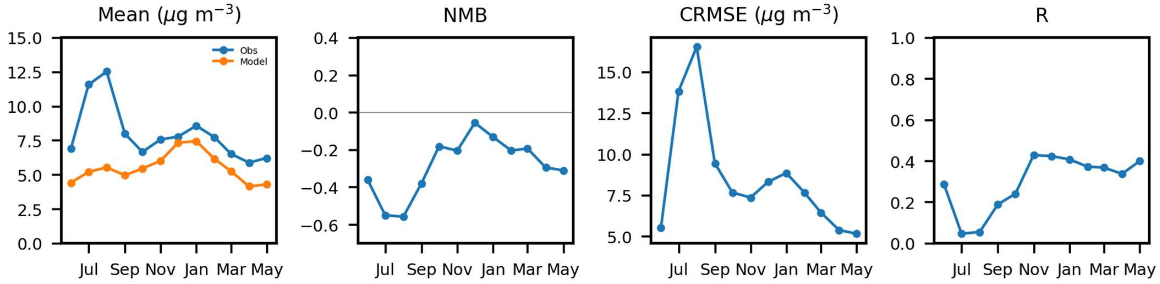
Figure 6. Time series of (a) observed and predicted monthly means of hourly NO_2 volume mixing ratio (ppbv) and monthly (b) NMB, (c) CRMSE, and (d) R scores for 2021/22 for all NRT stations.

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025



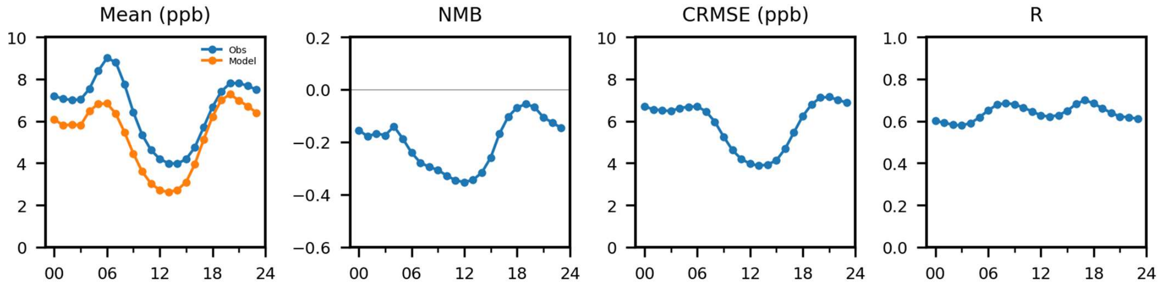
2620

Figure 7. Time series of (a) observed and predicted monthly means of hourly O_3 volume mixing ratio (ppbv) and monthly (b) NMB, (c) CRMSE, and (d) R scores for 2021/22 for all NRT stations.



2625

Figure 8. Time series of monthly (a) observed and predicted mean hourly $PM_{2.5}$ concentration ($\mu g \cdot m^{-3}$) and (b) NMB, (c) CRMSE, and (d) R scores for 2021/22 for all NRT stations.



2630

Figure 9. Time series of diurnal (a) observed and predicted mean hourly NO_2 volume mixing ratio (ppbv) and (b) NMB, (c) CRMSE, and (d) R scores for all of 2021/22 for all NRT stations. Time is in hours LT.

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025

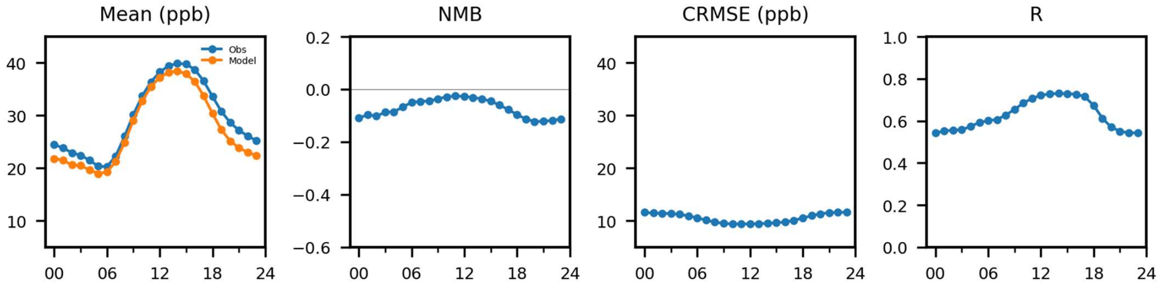


Figure 10. Time series of diurnal (a) observed and predicted mean hourly O_3 volume mixing ratio (ppbv) and (b) NMB, (c) CRMSE, and (d) R scores for all of 2021/22 for all NRT stations. Time is in hours LT.

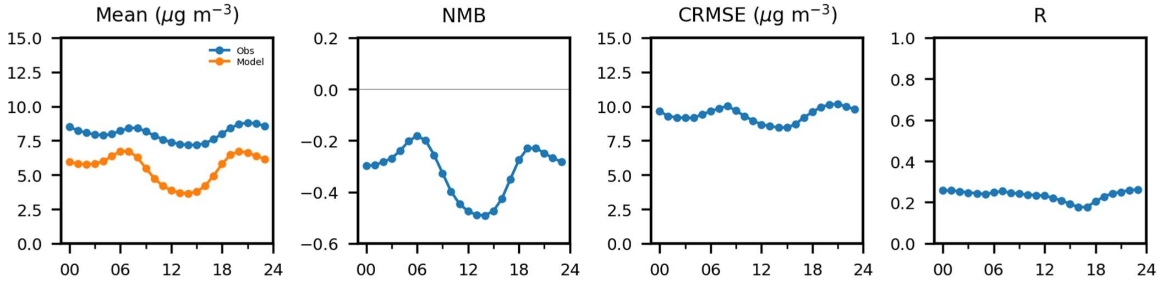


Figure 11. Time series of diurnal (a) observed and predicted mean hourly $PM_{2.5}$ concentration (µg·m^{-3}) and (b) NMB, (c) CRMSE, and (d) R scores for all of 2021/22 for all NRT stations. Time is in hours LT.

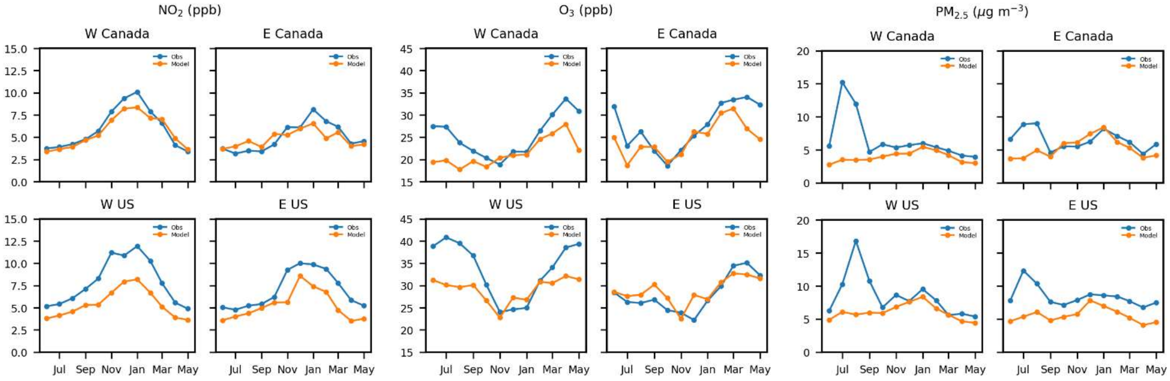


Figure 12. Time series of monthly observed and predicted mean hourly (left) NO_2 volume mixing ratio (ppbv), (centre) O_3 volume mixing ratio (ppbv), and (right) $PM_{2.5}$ concentration (µg·m^{-3}) for all NRT stations in LT but stratified into four regions [see Fig. S7] for 2021/22. Orange curves denote predicted values and blue curves denote observed values.

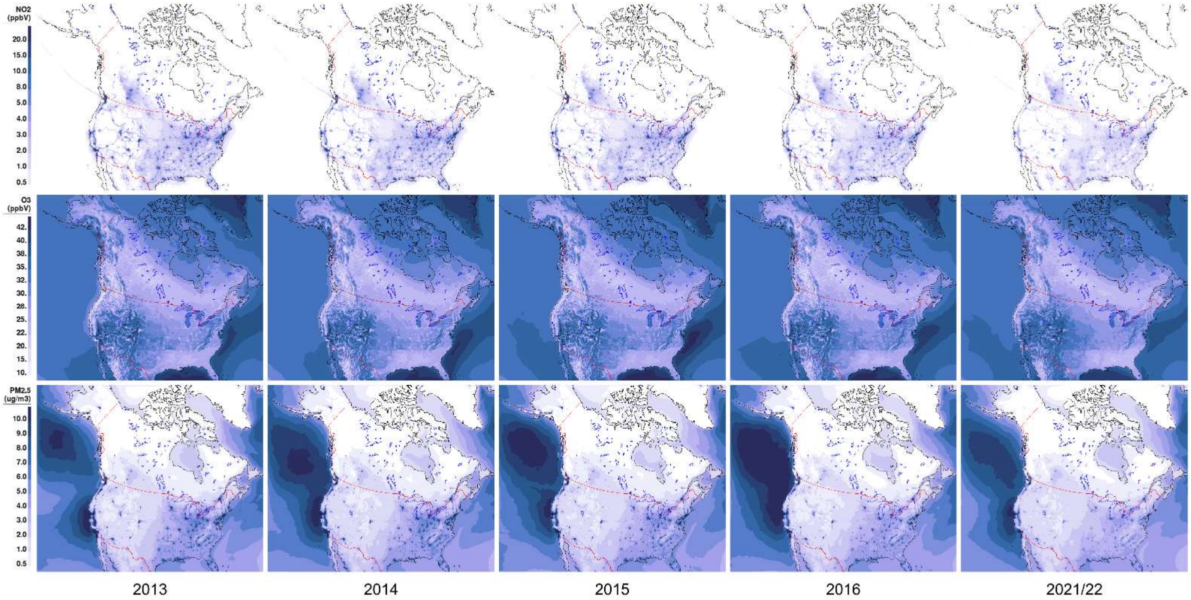
2655 

Figure 13. Spatial distribution of predicted mean annual abundances of hourly (top row) NO_2, (middle row) O_3, and (bottom row) $PM_{2.5}$ total mass for five years (from left to right, 2013–2016 and 2021/22). Units are ppbv, ppbv, and µg·m^{-3}, respectively. 2660

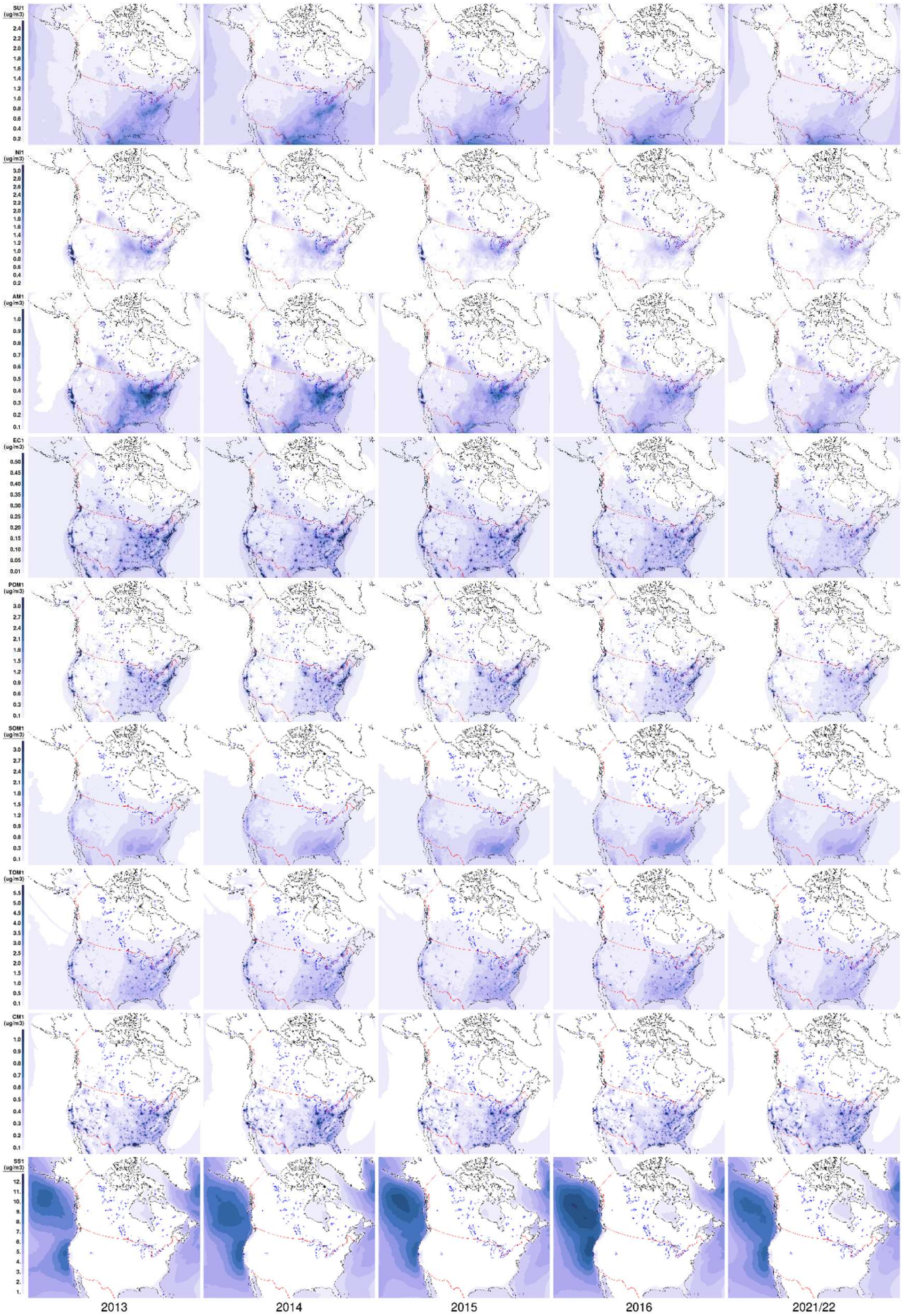
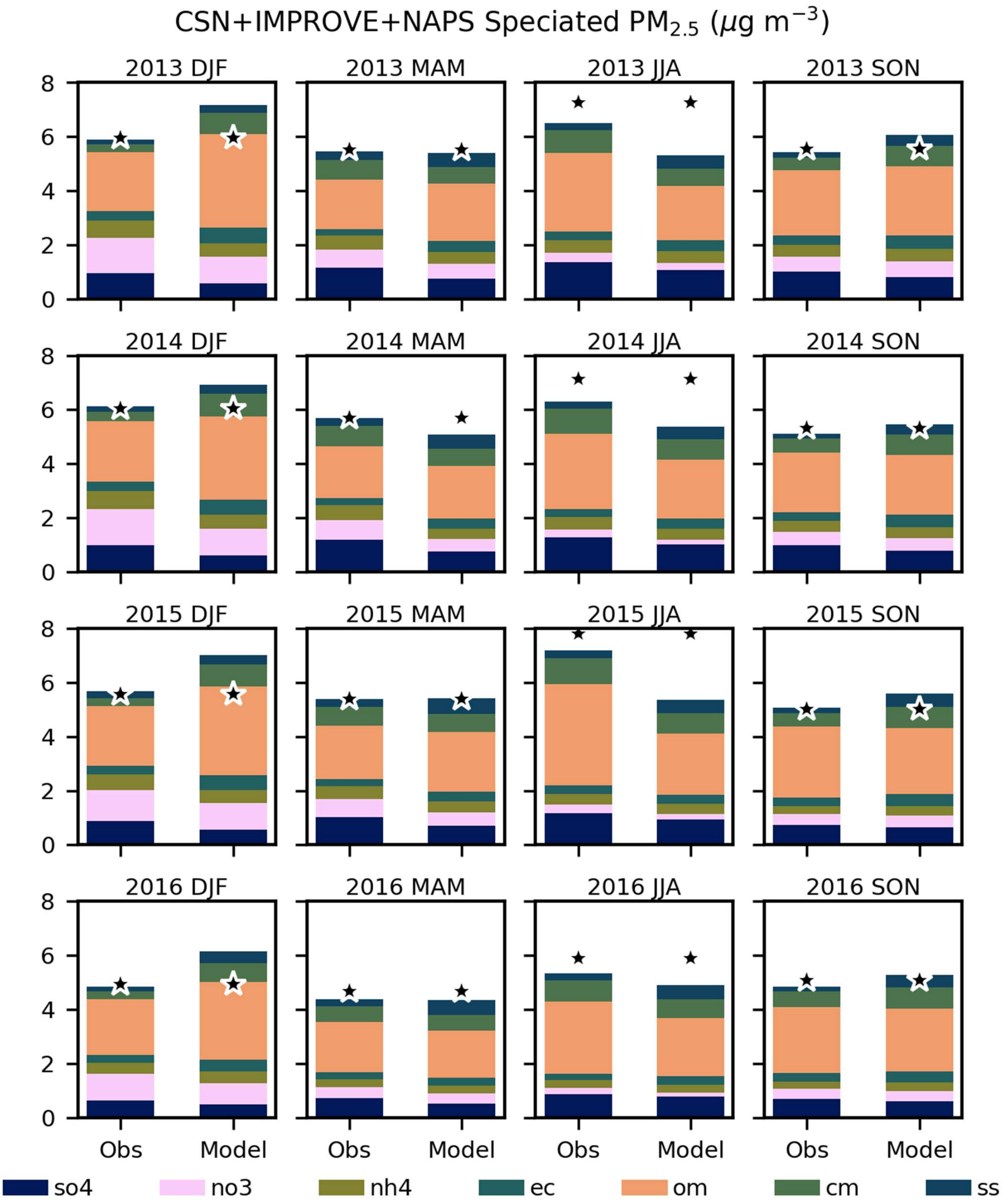


Figure 14. Spatial distribution of predicted mean annual ambient concentrations of nine daily $PM_{2.5}$ chemical components ($\mu g \cdot m^{-3}$) for five years (from left to right, 2013–2016 and 2021/22). These components are SO4, NO3, NH4, EC, POM, SOM, TOM, CM, and SS (in rows ordered from top to bottom).



2670

Figure 15. Stacked bar graphs of observed vs. predicted domain-wide seasonal $PM_{2.5}$ chemical component concentrations (μg·m^{-3}) based on combined CSN, IMPROVE, and NAPS $PM_{2.5}$ speciation daily measurements and hindcasts for four consecutive years. The top row corresponds to 2013 seasons, the next two rows below to 2014 and 2015 seasons, and the bottom row to 2016 seasons. Each row has four seasonal bar-graph pairs (observed and 2675 predicted), starting with winter (DJF) on the left, and then spring (MAM), summer (JJA), and autumn (SON) on the right. The stars mark the measured or predicted seasonal gravimetric mass.

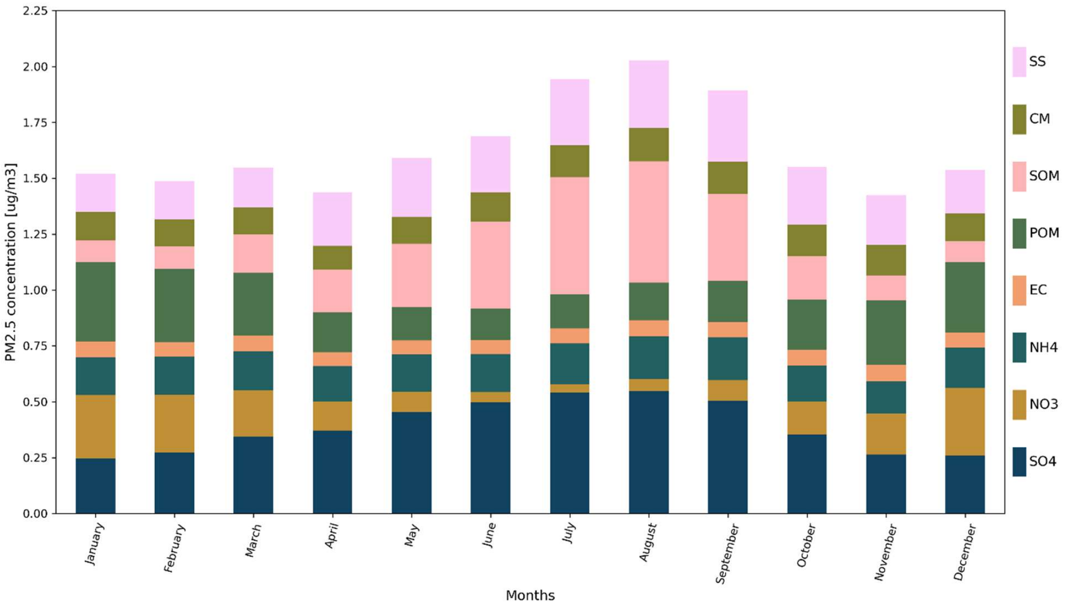


Figure 16. Time series of predicted monthly mean variation of eight daily $PM_{2.5}$ chemical component concentrations (µg·m^{-3}) area-weighted averaged over North American continent grid cells and 2013 to 2016 simulations. These components are SO4, NO3, NH4, EC, POM, SOM, CM, and SS (shown ordered from bottom to top in stacked bar graphs).

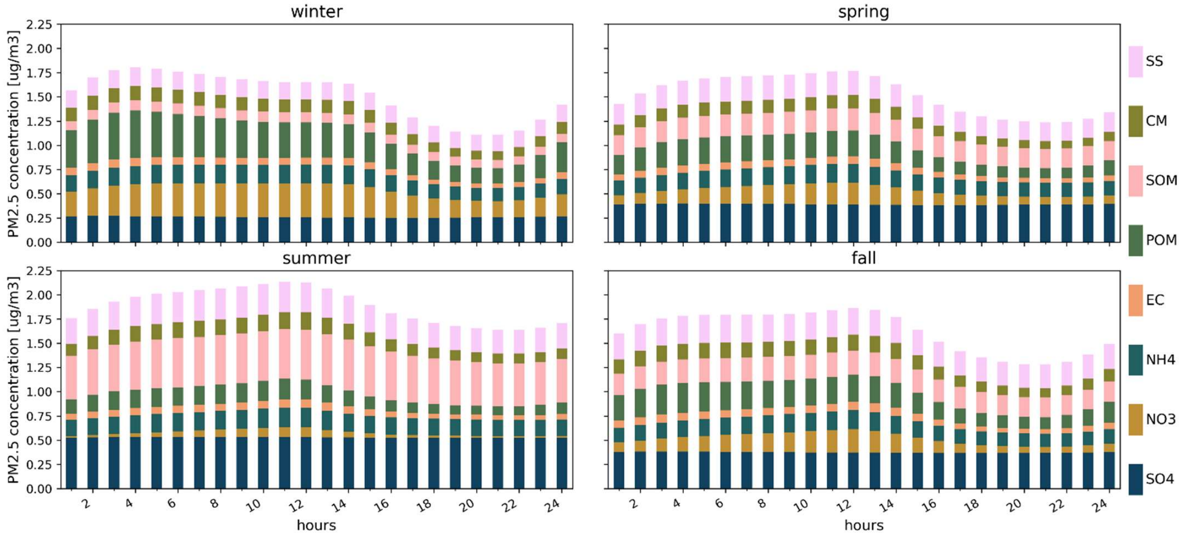


Figure 17. Time series of predicted mean diurnal (UTC) variation of eight hourly $PM_{2.5}$ chemical component concentrations (µg·m^{-3}) area-weighted averaged over North American continent grid cells and 2013 to 2016 simulations for each of four seasons: (a) winter; (b) spring; (c) summer; and (d) autumn. These components are SO4, NO3, NH4, CM, SS, EC, POM, and SOM (shown ordered from bottom to top in stacked bar graphs).

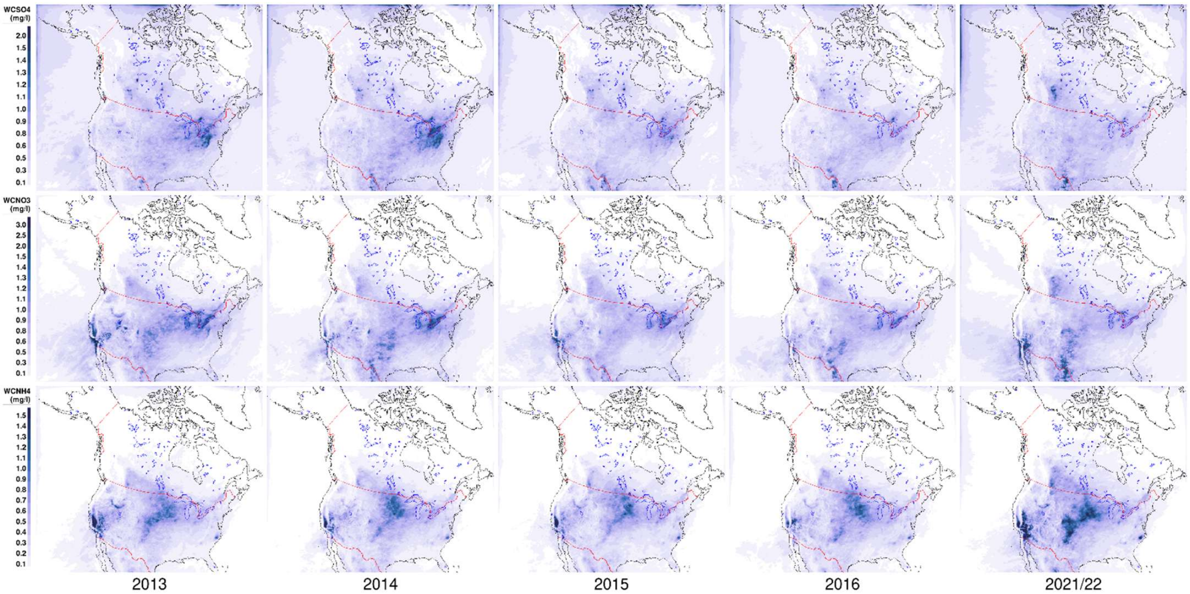
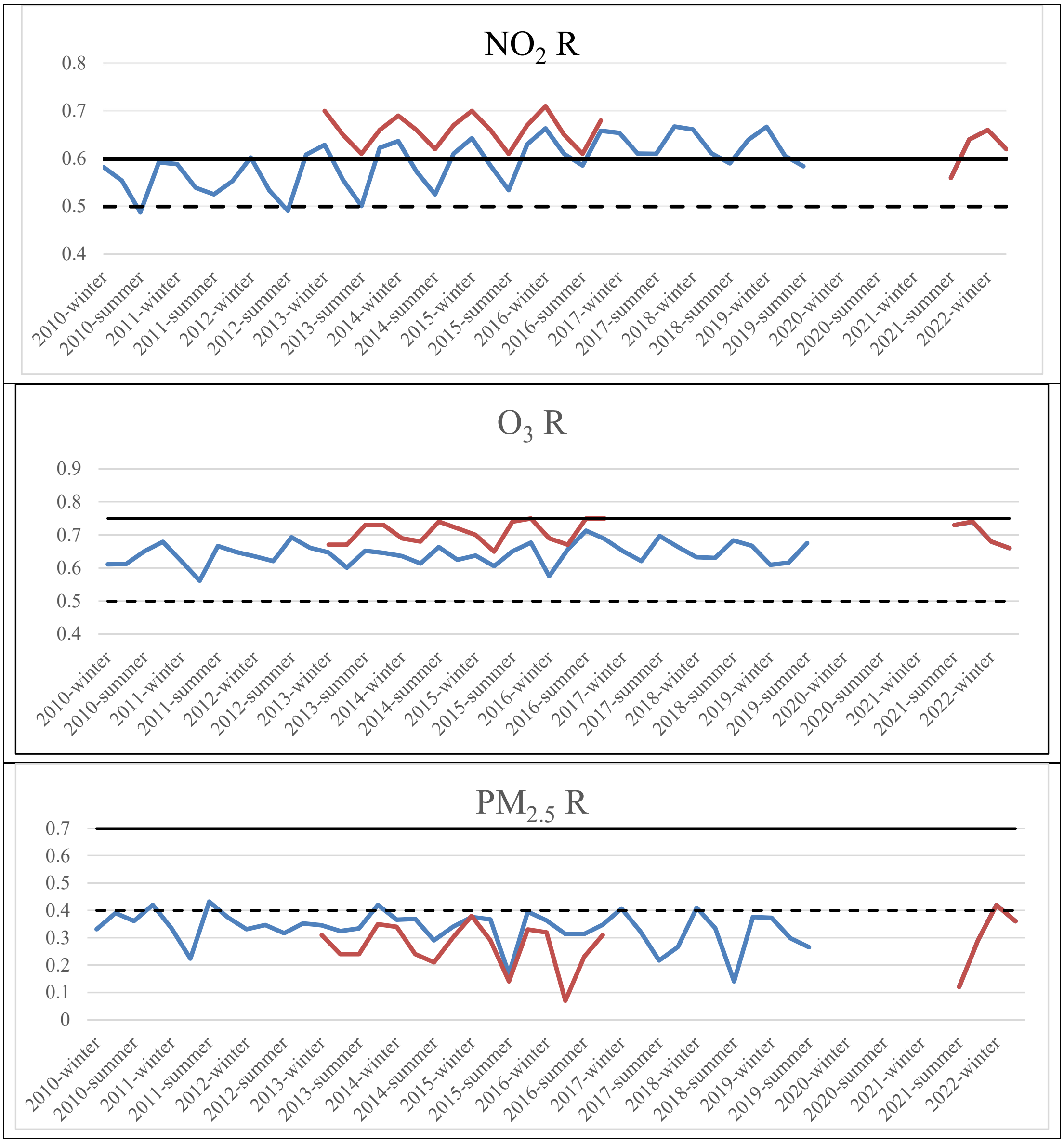


Figure 18. Spatial distribution of predicted mean annual concentrations in precipitation (mg·L^{-1}) of (top row) $SO_4^=$, (middle row) NO_3^-, and (bottom row) NH_4^+ for five years (from left to right, 2013–2016 and 2021/22).

https://doi.org/10.5194/egusphere-2025-4324
Preprint. Discussion started: 23 December 2025


NO$_2$ R

O$_3$ R

PM$_{2.5}$ R

Figure 19. Time series of seasonal correlation coefficient (R) scores for surface (top) NO$_2$, (middle) O$_3$, and (bottom) PM$_{2.5}$ hourly abundances for all available North American NRT measurement stations for the periods Jan. 2010–June 2019 and June 2021-May 2022. The solid blue line denotes scores for the operational RAQDPS at the time while the solid orange line denotes scores for the RAQDPS023 forecasts and hindcasts. The dashed and solid black lines mark the “acceptable” and “good” benchmark thresholds, respectively, taken from Emery et al. (2017) for O$_3$ and PM$_{2.5}$ and from Zhai et al. (2024) for NO$_2$.