Environ Fluid Mech (2010) 10:471–489
DOI 10.1007/s10652-009-9163-2

ORIGINAL ARTICLE

A framework for evaluating regional-scale numerical photochemical modeling systems

Robin Dennis · Tyler Fox · Montse Fuentes · Alice Gilliland · Steven Hanna · Christian Hogrefe · John Irwin · S. Trivikrama Rao · Richard Scheffe · Kenneth Schere · Douw Steyn · Akula Venkatram

Received: 21 August 2009 / Accepted: 16 December 2009 / Published online: 5 March 2010

Abstract This paper discusses the need for critically evaluating regional-scale (∼200–2,000 km) three-dimensional numerical photochemical air quality modeling systems to establish a model's credibility in simulating the spatio-temporal features embedded in the observations. Because of limitations of currently used approaches for evaluating regional air quality models, a framework for model evaluation is introduced here for determining the suitability of a modeling system for a given application, distinguishing the performance between different models through confidence-testing of model results, guiding model development,

R. Dennis · A. Gilliland · S. T. Rao (✉) · K. Schere
Atmospheric Modeling and Analysis Division, National Exposure Research Laboratory,
US Environmental Protection Agency, Research Triangle Park, NC 27711, USA
e-mail: rao.st@epa.gov

T. Fox · R. Scheffe
Air Quality Assessment Division, Office of Air Quality Planning and Standards, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA

M. Fuentes
Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

S. Hanna
Hanna Consultants, Kennebunkport, ME 04046, USA

C. Hogrefe
NYS Department of Environmental Conservation, Bureau of Air Quality Analysis and Research,
Albany, NY 12233, USA

J. Irwin
John S. Irwin and Associates, Raleigh, NC 27615, USA

D. Steyn
Department of Earth and Ocean Sciences, The University of British Columbia, Vancouver, BC V6T1Z4, Canada

A. Venkatram
Department of Mechanical Engineering, University of California, Riverside, CA 92521, USA

and analyzing the impacts of regulatory policy options. The framework identifies operational, diagnostic, dynamic, and probabilistic types of model evaluation. Operational evaluation techniques include statistical and graphical analyses aimed at determining whether model estimates are in agreement with the observations in an overall sense. Diagnostic evaluation focuses on process-oriented analyses to determine whether the individual processes and components of the model system are working correctly, both independently and in combination. Dynamic evaluation assesses the ability of the air quality model to simulate changes in air quality stemming from changes in source emissions and/or meteorology, the principal forces that drive the air quality model. Probabilistic evaluation attempts to assess the confidence that can be placed in model predictions using techniques such as ensemble modeling and Bayesian model averaging. The advantages of these types of model evaluation approaches are discussed in this paper.

Keywords Air quality model · Photochemical model · Model evaluation · Performance evaluation

1 Introduction

Regional-scale air quality models are designed to simulate air quality in a domain with a horizontal scale of several hundred to several thousand kilometers and a vertical scale of several kilometers. The horizontal grid cell size is usually on the order of a few kilometers and the smallest vertical grid spacing is on the order of tens of meters. Such three-dimensional numerical photochemical air quality models (AQMs) play a key role in the development and implementation of air pollution control rules and regulations in the United States and elsewhere [1–3], and they are also being used for short-term forecasting of air quality [4–6]. The prerequisite to such applications is an assessment of the degree to which an AQM can simulate the spatio-temporal features embedded in air quality data. This paper discusses multiple approaches for rigorously evaluating three-dimensional photochemical AQMs.

Over the last three decades, several workshops and research papers have addressed the evaluation of AQMs [7–9]. However, these workshops and papers have addressed short-range to mesoscale range plume or puff-type AQMs rather than regional-scale three-dimensional numerical photochemical modeling systems. The statistical metrics developed to evaluate short-range dispersion models are limited in their ability to evaluate the ability of regional-scale models to simulate the complex relationships among the variables that constitute the photochemical system. Most evaluation methods for short-range models focus on generating statistics of the deviations between the modeled concentrations of a few species and the corresponding observations. While such statistics are useful, they provide little insight into the adequacy of models for the many processes that constitute the complex three-dimensional air quality system. Recognition of these shortcomings led the U.S. Environmental Protection Agency (EPA) and the American Meteorological Society (AMS) to convene an invited group of nearly 100 experts at a workshop during August 7–8, 2007. The objectives of the workshop were to (1) examine current approaches for the evaluation of regional scale models, (2) discuss new approaches to advance air quality and related model evaluation methods and procedures, and (3) develop a set of recommendations for model evaluation methods, procedures, and metrics for different components of regional AQMs for further testing and use by the air quality modeling community. This paper is motivated by the discussions held among the workshop participants.

2 Model evaluation framework

Three-dimensional time-dependent numerical models of the atmosphere describe processes at a wide range of spatial and temporal scales, and they are used in widely differing applications ranging from research on atmospheric processes to air quality forecasting. For regulatory applications, a model must provide an adequate estimate of concentration response to forcing variables, such as emissions and meteorology, in addition to adequate quantitative estimates of species concentrations. By contrast, a forecast model is judged solely by its ability to simulate the temporal evolution of chosen forecast variables. Hence, *model evaluation criteria* are dependent on the context in which models are to be applied [10]. Nevertheless, the following three primary objectives can be identified:

(1) *Determining the suitability of a model system for a specific application and configuration.* The main goal of a model evaluation exercise (including regional AQMs) is to demonstrate that the model is "performing adequately" when compared with observations, for the purposes for which the model is applied. The purpose of model application as well as the relevant model outputs should be stated at the outset. For air quality management, we are mainly interested in the model's ability to correctly estimate the air quality response to changes in potential source emissions. In this application, we focus on assessments of the model's simulation of the governing processes and the interaction among them. Emphasis in air quality forecasting is chiefly on the outcome state of the model, a prediction of next-day air quality.

(2) *Distinguishing the performance among different models or different versions of the same model.* We need to compare the relative performance of different models in comparing their results to observations so we can better understand models' strengths and limitations. Evaluation procedures must to be able to distinguish the relative performance with specified levels of statistical significance [11]. The model inter-comparisons can identify model deficiencies and areas requiring further model development.

(3) *Guiding model improvement.* Evaluation exercises should shed light on the uncertainties in the simulation of atmospheric processes attributable to model parameterizations and model input. The results of these exercises should lead to improved AQMs.

Figure 1 introduces a model evaluation framework, incorporating the above three major objectives. "Operational evaluation" refers to generating statistics of the deviations between model estimates and observations, and comparing their magnitudes to some selected criteria. "Diagnostic evaluation" examines the ability of the model to simulate each of the interacting processes that govern the air quality system. "Dynamic evaluation" focuses on the model's ability to predict changes in air quality concentrations in response to changes in either source emissions or meteorological conditions. Recognizing that there is uncertainty in model inputs and formulation of processes, "Probabilistic evaluation" focuses on the modeled distributions of selected variables rather than individual model estimates at specific times and locations.

3 Evaluation methods

This section provides details on the approaches embodied in the proposed model evaluation framework. We provide some illustrative examples of their application to regional AQMs.

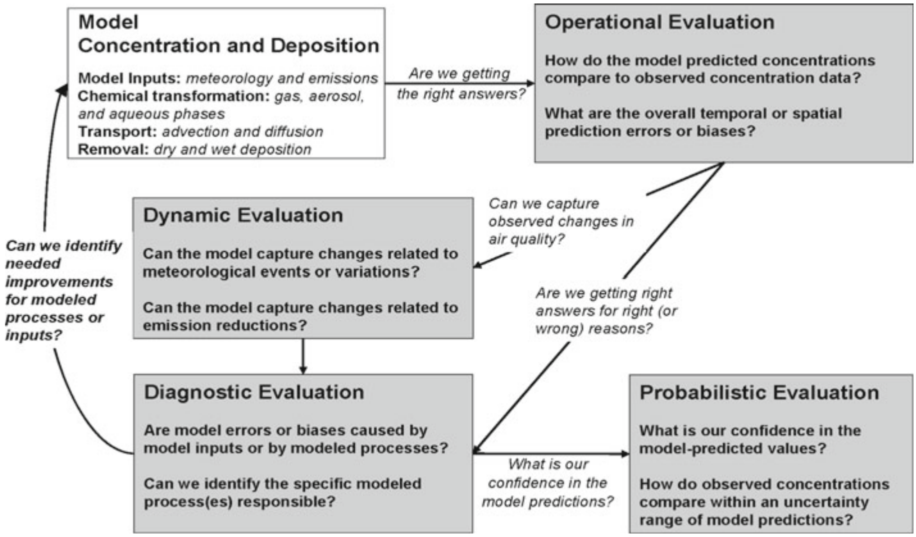

Model
Concentration and Deposition
Model Inputs: meteorology and emissions
Chemical transformation: gas, aerosol, and aqueous phases
Transport: advection and diffusion
Removal: dry and wet deposition
Are we getting the right answers?
Operational Evaluation
How do the model predicted concentrations compare to observed concentration data?
What are the overall temporal or spatial prediction errors or biases?
Dynamic Evaluation
Can the model capture changes related to meteorological events or variations?
Can the model capture changes related to emission reductions?
Can we capture observed changes in air quality?
Are we getting right answers for right (or wrong) reasons?
Can we identify needed improvements for modeled processes or inputs?
Diagnostic Evaluation
Are model errors or biases caused by model inputs or by modeled processes?
Can we identify the specific modeled process(es) responsible?
What is our confidence in the model predictions?
Probabilistic Evaluation
What is our confidence in the model-predicted values?
How do observed concentrations compare within an uncertainty range of model predictions?

Fig. 1 A framework for evaluating regional-scale photochemical modeling systems

3.1 Operational evaluation

Operational evaluations make use of routine observations of ambient pollutant concentrations, emissions, meteorology, and other relevant variables. The modeled meteorological variables considered in operational model evaluation include temperature, moisture (humidity), wind speed and direction, planetary boundary layer height, surface radiation, clouds and precipitation. Air quality variables include concentrations of ozone (O_3), carbon monoxide (CO), nitrogen oxides (NO, NO_x), and fine particulate matter mass and its species (fine particulate matter [$PM_{2.5}$], sulfate [SO_4], nitrate [NO_3], ammonium [NH_4], organic and elemental carbon [OC, EC]).

The three performance measures most widely used in AQM evaluation (and most other types of model evaluation) are mean bias (MB), root mean square error (RMSE), and correlation (R) [12]. However, statistical confidence levels in these statistics are rarely calculated. This information can be used to answer questions such as "Is the model mean bias significantly different from zero at the 95% confidence level?", or "Is the correlation coefficient for one model significantly different from the correlation coefficient for another model?" It is important to note that observations and corresponding modeled values may contain different spatio-temporal correlation structures, complicating the interpretations of confidence intervals and other statistics for judging model performance.

The standard metrics (MB, RMSE, and R) do not take into consideration that predictions from 3-dimensional regional AQM models are volume-averaged ensemble mean (representing average weather conditions, physical processes, and chemical reaction rates) concentrations, whereas observations are point measurements reflecting individual events. This inconsistency is referred to as the *incommensurability* or *change of support* problem [13]. One way of dealing with this problem is to use spatial smoothing such as block-kriging on the observed data to produce values that can be compared with the grid-averaged model estimates. However, such smoothing techniques rely on a statistical model to interpolate observations, and, thus, the evaluation is based on a comparison of the results of two different models, and not a direct comparison of model output and corresponding observations. Furthermore, observations contain measurement errors while model outputs contain errors due to inadequacies in both the model input data and the model's representations of the relevant atmospheric processes.

Often, dense observations at the ground level and aloft are not available to adequately define the initial and boundary conditions for numerical photochemical AQMs [14]. It is well-recognized that without completely knowing the 3-D initial chemical state of the atmosphere, its future state cannot be simulated accurately. Also, whereas the observations contain stochastic variations, models do not. Thus, one should expect differences between model outputs and their corresponding observations. Most operational model evaluations conducted and published to date have simply paired the observations and modeled values in computing statistical metrics such as MB, RMSE, and R without properly taking into account the points mentioned above. Hence, any agreement found between the paired observations and modeled results should be considered fortuitous.

The spatio-temporal patterns of model predictions and observations can be compared by determining the fractional overlap of spatial patterns or time series of predictions and observations [15]. The evaluation could determine whether the scales of variability in the predicted and observed patterns are comparable using correlation and spectral analysis. Differences between maps of model predictions and maps computed from observations yield a spatial difference field. Investigation of spatial patterns can be done using statistical measures of spatial dependency, such as the *variogram* function, and temporal dependency structure can be studied with methods such as spectral analysis. For example, time series of ozone (O_3) have been decomposed into spectral bands representing intra-day, diurnal, synoptic, seasonal, and longer-term fluctuations [16,17]. Figure 2a illustrates the comparison between these component spectra estimated from 15 years of observed and CMAQ model-predicted hourly O_3 data. The figure reveals the model's ability in capturing the variability associated with diurnal and synoptic features in the time series of O_3. There are apparent problems in the model's simulation of the variability inherent in high-frequency (hour-to-hour) variations, as well as a tendency for the model to underestimate the variability of the seasonal and longer-term O_3 signal, possibly due to the inaccuracies in the regional model's boundary conditions, emissions, and representation of the free tropospheric processes.

Empirical orthogonal functions can also be used for analysis of spatial/temporal data. This approach provides a decomposition of the spatial response surfaces in terms of the principal components that explain the spatial structure at different scales. For this second-order assessment (based on the correlation structure), graphical displays can be used such as the spatial variogram and estimated temporal spectrum for both model output and data-based grid cells, and also for the difference field (differences maps between model and data-based grid cells).

Some graphical techniques in operational model evaluation have been alluded to earlier in conjunction with standard statistical metrics. While scatter plots of percentile values of pollutant concentrations and time-series plots have been useful for regional AQM analyses [5,18], it may be more appropriate to aggregate results across coherent space and/or time regions based on techniques such as Principal Component Analysis to represent distributional quantities, and not single point observations [19,20]. For example, daily time series of summary statistics for O_3 concentrations over all monitoring sites in a region (where pollutants are spatially-coherent) can be plotted as box plots over a month or longer period for model results and observations. The hourly O_3 concentration values for a month (or a season) at a site (or averaged over sites within a given sub-region) can be used to track the diurnal variation of modeled and observed averages, variances, bias, etc. Time series of model bias and error distributions are also useful. Pie charts or bar graphs of particulate matter species

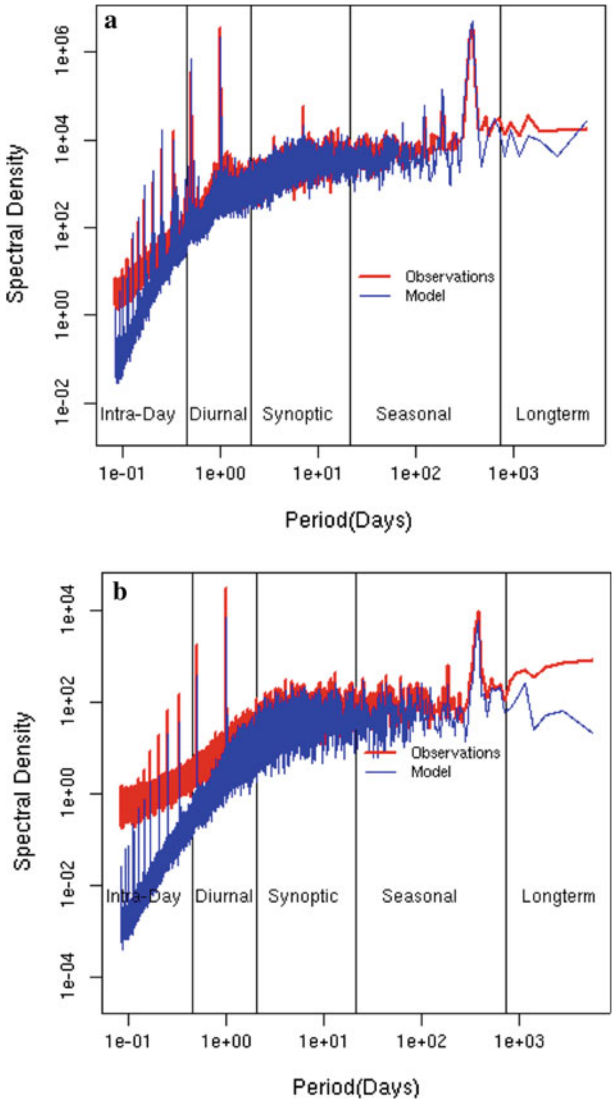


Fig. 2 **a** Power spectra of O_3 time series from CMAQ model results (*blue line*) and observations from ground monitoring networks (*red line*). Time series of model and observed data used in the analysis covers a 15-year period ending in 2002; **b** same as (**a**) except for wind speed

are useful for comparing simulated and observed chemical constituents of size-segregated particulate matter [21]. Scatterplots can be used to compare distributions of observed and modeled parameters, such as that for $PM_{2.5}$ shown in Fig. 3 [22]. From an operational evaluation perspective it is recommended that standard statistics (R, MB, RMSE) be calculated from the distributional comparisons of observed and modeled variables; *this is a more appropriate alternative to strict pair-wise comparisons.*

Performance goal plots ("soccer" plots) that summarize model performance by plotting performance goals and criteria for fractional bias versus fractional error, and concentra-

Fig. 3 Comparison of CMAQ model forecast and observed daily-average $PM_{2.5}$ distributions for January–February 2005. Observed $PM_{2.5}$ concentrations are from the AIRNOW network. At each observation location the time-series of modeled and observed daily-average $PM_{2.5}$ is examined and percentiles of the distributions are computed. The figure illustrates the relationship between the modeled and observed percentiles (denoted by different colors) at each location. Also shown are the 1:1, 1:2, and 2:1 lines [22]

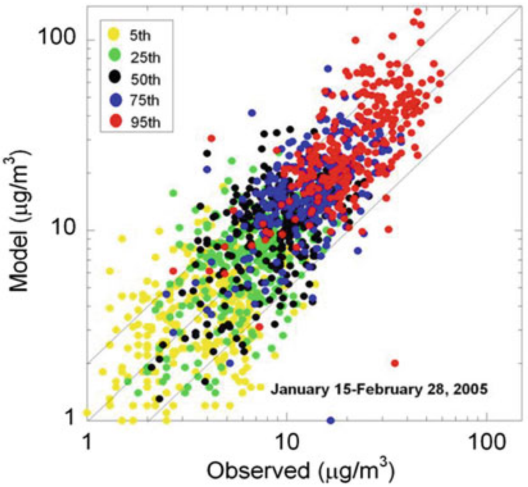
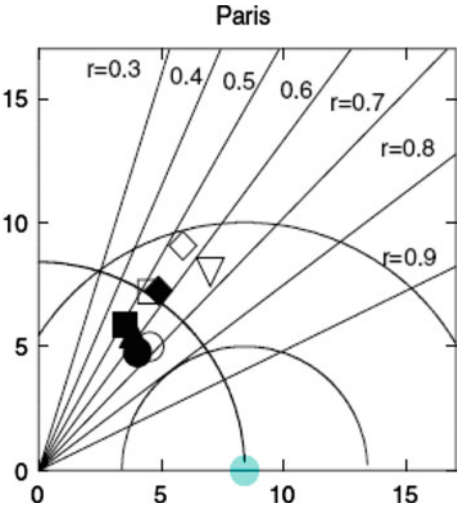


Fig. 4 Taylor diagram for model results for O_3 in Paris region for 1999 [25]. *Symbols* represent results for distinct models. Values along axes are in ppb



tion performance plots ("bugle" plots) that display fractional bias or error as a function of concentration have been suggested [23]. A Taylor diagram [24], which combines model error and correlation statistics in a single plot, has been found to be useful for comparing the performance of several models [25]. Figure 4 provides an illustration of the Taylor diagram where, for each model included, the standard deviation of simulated values (radius) and the time correlation between simulated and observed values (angle from horizontal) are indicated. The standard deviation of observations is shown as the point on the horizontal axis, and circles centered on this point represent points of equal simulation error standard deviation. As shown in the figure, error standard deviations are smallest for models with the highest correlations.

For regional models in particular, a basic comparison of the spatial extent and magnitude of the modeled concentration field through a concentration isopleth or colored grid plot overlaid with the observations or compared with a similarly analyzed field from the data-based grid cell values from kriging or other spatial analysis techniques, can often provide a strong initial indication of how well the model is predicting the spatial texture and magnitude of the species of interest. This type of screening analysis is often the essential first step in putting into perspective the representativeness of the statistical measures and deciding on subsequent steps in the operational evaluation. The spatial extent comparison can be made more objective by using pattern comparison techniques, such as the figure of merit [26] and e-folding distance [27].

Emission models are an integral part of regional AQM systems and need to be evaluated. However, estimates from emissions models cannot be directly compared with observed values because emission observations generally do not exist on the regional-scale. The sole exception to this general case is the Continuous Emissions Monitoring Systems (CEMS), which measure primary pollutant emissions on the tall stacks of large electrical generating units. These data are used directly as emission inputs into AQMs. For other emissions sectors, the primary assessment tool is quality assurance and control of the process, such as aggregating emissions estimates by state or by source sector and comparing these estimates to previous or independent emissions estimates. Examining statistical distributions of emissions across a model domain can help identify outliers or questionable data for further examination. Studying the spatial distribution of emissions surrogates (e.g., population, road networks) or the temporal allocation of emissions (e.g., seasonal and daily patterns) may also help spot obvious errors. While operational evaluation methods are applicable to only a few limited sets of emissions data because of the lack of real-world emission measurements for AQMs, diagnostic methods may provide insights into biases and errors in the emissions. These techniques will be discussed as part of the next section.

3.2 Diagnostic evaluation

Operational evaluations do not provide information on the adequacy of models for representing the many interacting processes that lead to the concentrations that are finally modeled. Diagnostic evaluation methods are designed to probe into the physical and chemical process models or representations. Regional AQM diagnostic evaluations are complicated by the fact that the system is non-linear: a change in a given model input does not always lead to a proportional response in the model output.

An examination of the chemical processes in the AQM requires precursor concentrations such as speciated volatile organic compounds and NO_y along with radiation data and photolysis rate estimates at relatively high temporal resolution (e.g., 10-min averages). Diagnostic evaluation of aerosol chemistry also requires extensive data for the individual aerosol species, their size distributions, and their chemical precursors. The direct and indirect influences of the meteorology on the chemical concentrations require data on meteorological parameters that are not typically available, such as the planetary boundary layer heights and cloud heights and cover, both of which have a large impact on air quality concentration levels. These types of diagnostic evaluation can be obtained through process-oriented field studies, but for very limited locations and periods of time due to the resources required. Some field studies and special data sets include both surface data and aloft measurements via aircraft or tower. Using information from such studies can help to evaluate the modeled chemistry and transport processes in the free troposphere and focus on larger regional impacts and emission budgets aloft [27]. Given the large investments in, and limited availability of these field studies, many diagnostic evaluation studies are tailored to focus on the information and data available from short-duration special studies.

Diagnostic evaluation is aimed at understanding the reasons for poor and good model performance. It can help to build additional confidence in the model even when operational model performance statistics are deemed acceptable. A sensitivity test, which examines a model's response to perturbations in its inputs, is a common way to ascertain whether inputs have a notable influence on model performance issues. A fundamental description of sensitivity analyses of environmental models is given by Saltelli et al. [28]. Cullen and Frey [29] provide specific discussions related to AQMs. However, because of the nonlinear response of a regional AQM, sensitivity tests may be valid only for a limited range of input variables. Air quality simulations can be performed using multiple meteorological inputs to assess how much meteorological model errors and differences impact the air pollutant [30,31]. Emissions have also been varied either through incremental changes to emission inputs or comparison across different inventory estimates to test the impact on air quality endpoints [32]. Figure 5 illustrates an evaluation of total nitrate estimates from the CMAQ model. Figure 5a shows an operational comparison of simulated total nitrate ($HNO_3 + NO_3$ aerosols) with measurements from the CASTNet network on a monthly basis during 2001. In Fig. 5b the sensitivity of CMAQ results are diagnostically probed as a function of the treatment of microphysics and soil moisture in the meteorology model (MM5). In another experiment, the sensitivity of ozone estimates from the CMAQ model to the representation of the chemical mechanism is illustrated in Fig. 6. In this example, the CMAQ results for ozone using the Carbon Bond 4 (CB4) chemistry are compared to those using the 1999 version of the Statewide Air Pollution Research Center (SAPRC99) chemical mechanism. The differences seen in the spatial plots are a representation of the chemical uncertainties in the model results. Other chemical diagnostic techniques for model evaluation include the use of the ozone production efficiency [27] for gas-phase photochemistry and the gas ratio for gas-to-aerosol partitioning.

Advanced instrumented modeling tools (e.g. direct decoupled method, adjoint models, sulfur tracking method) have also been introduced into model evaluation research, where contributions from various processes or inputs on pollutant concentrations are tracked during the simulation. The tracking information from these instrumented modeling tools can sometimes replace the need for numerous brute-force sensitivity simulations. For example,

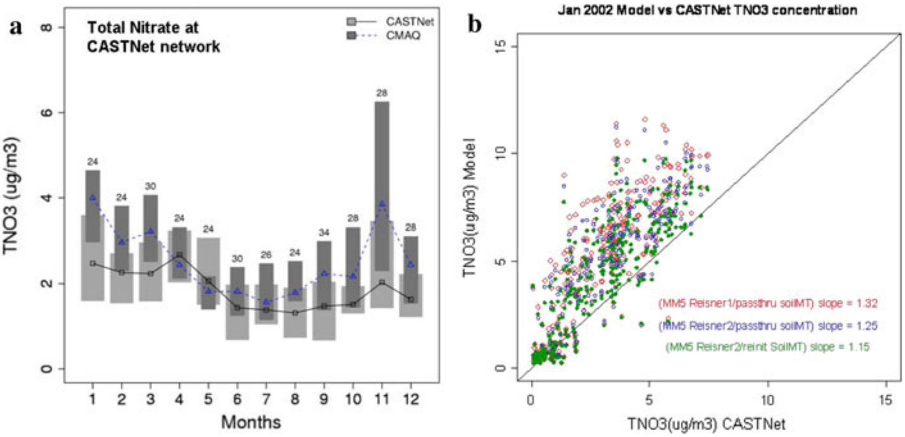


Fig. 5 **a** Comparison of monthly simulated distribution of total nitrate ($\mu g/m^3$) for 2001 from CMAQ model with CASTNet network measurements. **b** Comparison of January 2002 total nitrate concentrations between CMAQ model and CASTNet measurements. CMAQ results are shown for three different simulations, using different microphysics and soil temperature options in MM5 meteorology model

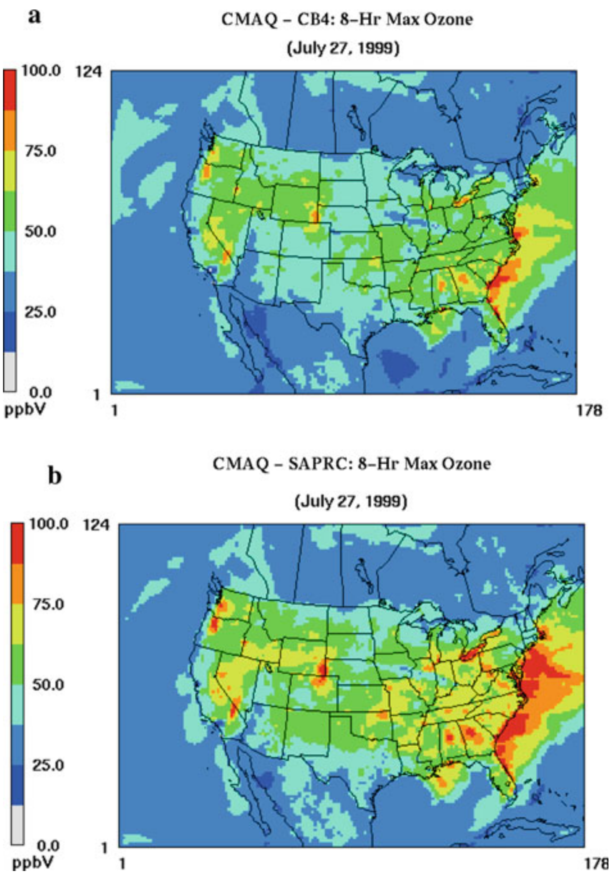


Fig. 6 CMAQ model results for 8-h maximum daily ozone concentrations on July 27, 1999 using **a** the CB4 chemical mechanism and **b** the SAPRC99 chemical mechanism

process analysis tools have been embedded into AQMs to characterize the impact of transport processes, chemical production and loss pathways, and sensitivity to NO_x or radical emission sources on ozone concentrations [33,34]. Another example of an instrumented modeling tool is the Direct Decoupled Method (DDM) that has been incorporated into the CMAQ modeling system, where the integral sensitivity of O_3 and $PM_{2.5}$ predictions to emission precursors, source regions and sectors, and boundary conditions is calculated during the model simulations [35,36]. The DDM tool is able to capture both the first and second order sensitivities to these inputs, which, depending upon the size of the perturbations studied, are important for non-linear chemical systems.

Meteorological models have long been used to forecast weather, but AQM predictions are sensitive to a number of different meteorological variables that are not as critical to weather prediction. Evaluation of such models for the purpose of providing weather forecasting guidance may not be sufficient to assure their reliable use in air quality applications. Seaman [37] provided a comprehensive summary of the key meteorological issues most relevant for air quality modeling. For retrospective air quality modeling, meteorological simulations often include various approaches for data assimilation or nudging, so that agreement between meteorological observations and predictions is optimized. Otte [31] provides an example of a diagnostic study that demonstrates that assimilation of observations into the meteorological predictions can contribute to improved ozone predictions, in addition to improved meteorological predictions. However, power spectra of modeled and observed temperatures and wind speeds reveal large underestimation of the variability in the high-frequency intra-day band even with 4-dimensional data assimilation (Fig. 2b). The results in Fig. 2b imply that one should expect large differences to be found in the hour-to-hour comparisons of modeled and observed values of meteorological and chemical variables since the variability in the short scales is not well-represented in the model.

For observationally-based methods such as receptor models, speciated observations are needed on shorter time scales in order to decipher the source signatures to distinguish between different source types. In many cases, the data are only available for limited time periods and specific locations. However, receptor models can be the first major step to understanding the types of sources contributing to air pollution at a given location and can help identify potential missing sources in an emission inventory. Inverse modeling also can be limited by data if the network does not provide high-resolution spatial and temporal data or if the observed species does not provide a conservative indicator for the emitted species (e.g., ammonium is not a conservative indicator for ammonia emissions). Additionally, since inverse modeling relies on the AQM to estimate the relationship between the emissions and the resulting concentration, model error should be included in the calculations whenever possible and such methods are only helpful if the known emission uncertainties are much larger than the error intrinsic to the AQM processes that also impact the concentrations. Recent advances have introduced approaches that integrate receptor modeling methods into AQMs [38] and used detailed tracking of emission contributions across space for inverse modeling [39]. In all cases, top-down methodologies can inform improvements needed for bottom-up inventories that are critical for AQM performance.

3.3 Dynamic evaluation

Dynamic evaluation looks at a retrospective case(s) to evaluate whether the model has properly predicted air quality response to known emission and/or meteorological changes. The change in concentration is evaluated instead of the “base” concentration itself, unlike operational and diagnostic aspects of model evaluation. This method is used in addition to traditional indicator ratios that focus on a model’s potential response to a change in emissions through chemical relationships (e.g., O_3/NO_y). One example of dynamic evaluation includes modeling assessments of the weekday/weekend concentration differences where mobile source emissions are known to significantly change [40]. These studies can provide insight into the ozone response to NO_x emissions in core urban areas with very dense mobile emissions. A model should also be able to track the impacts of emissions changes over longer time periods. Figure 7 displays an 18-year smoothed record of NO_x and CO concentrations at several urban monitoring stations and the analogous record from CMAQ model simulations. The data show that the modeled NO_x concentrations are about 50% lower than the observations, at least partially due to subgrid-scale emission gradients. However, there is good agreement between observed and simulated trends, with both sets of data showing approximately 30 ppb reduction of ambient NO_x concentrations over this time period. The CO analysis indicates that the modeled concentrations are about 50% lower than observations for the earlier time periods, with the underestimation decreasing to about 20% for the later time periods. The observations also show a steeper decrease over time than the CMAQ model, implying that the emissions inventory for CO was more severely underestimated in the early time period.

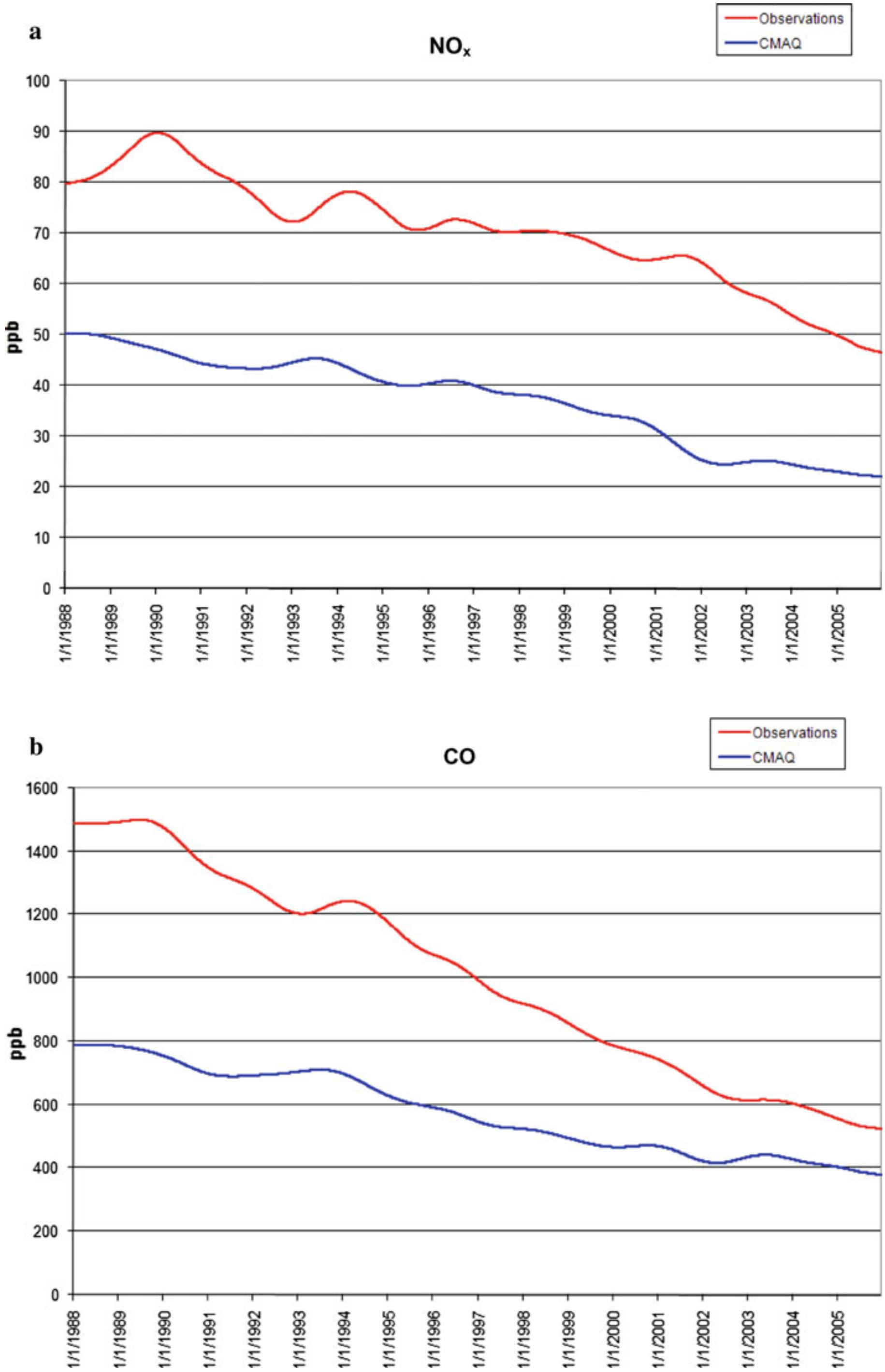


Fig. 7 Observed and simulated long-term smoothed time series (1988–2005) of **a** NO_x (averaged over 3 stations) and **b** CO concentrations (averaged over 34 stations) in the eastern United States

More recently, an evaluation of an AQM's response to a regulatory emission reduction program has been assessed [27,32,41]. The "NO_x SIP Call" was an unusual example of an emission control program that required a large reduction in emissions in a short span of time from the electricity generating sector [42]. Since those emissions are monitored with Continuous Emission Monitoring Systems, it was a unique opportunity for dynamic evaluation where the emission change could be directly measured and then tested in an AQM. Evaluation of the model's prediction of air quality response to such emission changes is challenged by the question of whether the year to year air quality changes are also being influenced by different meteorological conditions from 1 year to another. In a multi-year simulation, one could examine how the seasonality and trends in the air quality data are simulated by the model. Further work in this area of dynamic evaluation should include sensitivity studies with varying meteorology with the same emission reductions, as well as statistical methods that are traditionally used to adjust the observed pollutant concentrations for meteorological influences [43,44].

3.4 Probabilistic evaluation

All regional numerical AQMs use first-order closure, and, hence, the model outputs represent population means reflecting average weather and chemical conditions. It is of course possible to restructure the model system to solve the equations using second-order or higher closure. Thus, the model solves for the ensemble mean and the variance. A distribution shape is assumed (the clipped normal) and thus the full distribution is obtained. If regional AQMs were to use second-order closure, the computational times required would be much larger. Thus, the current crop of first-order closure regional AQMs are inherently deterministic (for a given scenario with a given set of inputs, the same concentrations are predicted). They also do not explicitly account for underlying uncertainties in the data, science process algorithms, or numerical routines that constitute the modeling system. Probabilistic model evaluation should allow quantification of the confidence in regional AQM-predicted values and determination of how observed concentrations compare within an uncertainty range of model estimates. There are no widely-used prescribed methods for determining such confidence levels through a probabilistic evaluation. A method suggested by Lewellen et al. [45] depends on knowledge of the probability distribution function (pdf) of the AQM predictions. This probabilistic model evaluation methodology was applied by Hanna and Davis [7] to regional AQM (UAM-V) predictions of ozone in the eastern U.S. It was shown that, across the full distribution range for all observing sites, the observations generally fell within the 95% confidence bounds of the regional AQM predictions. For that exercise, the pdf of the model predictions was determined from a previous Monte Carlo uncertainty study for that model on that domain and episode. Also, Irwin et al. [46] used the Monte Carlo approach to propagate uncertainty in meteorological inputs, using a probability distribution function (pdf), to air quality predictions.

Yet another technique uses an ensemble of modeling methods to approximate a pdf [47–54]. The ensemble method is a subset of a full Monte Carlo uncertainty exercise, where a few model simulations are made using varying inputs and other assumptions in hopes that a limited number of simulations will "cover" the full uncertainty range. The use of the ensemble method with prognostic meteorological models linked with a dispersion model was tested by Warner et al. [55], who showed that the method was able to adequately account for the uncertainties in the concentration pdf due to mesoscale and regional meteorological variations.

A series of studies [56–58] have shown that the effect of model-to-model uncertainty on the simulated response to emission reductions is typically on the order of a few percent of daily maximum 8-h ozone concentrations, much smaller than the effect on absolute concentrations for the "base case" simulation. Bayesian Model Averaging (BMA) [59] has been used to calibrate the ensemble predictions by weighting each individual ensemble member generated in the Pinder et al. [60] study based on how closely it matches observed ozone values. This approach provides an estimated probability distribution of pollutant concentrations at any given location and time, which can be used to estimate a range of likely, or "highly probable", concentration values or the probability of exceeding a given threshold value for a particular pollutant [61]. Figure 8 illustrates a month-long time series of daily 8-h maximum O_3 concentrations from a 200-member CMAQ model ensemble along with the observed concentration time series for this single observation site. This technique is useful for diagnosing structural process-based errors in the AQM system. When the envelope of ensemble results brackets the observations there is more confidence that the modeled system processes can replicate reality. On the other hand when the observations fall outside of or barely within the ensemble envelope, there is an indication that the model is biased across many process combinations with respect to replicating reality.

This type of model assessment is particularly useful in examining the relative efficacies of various emission control options in meeting a given air quality objective and in selecting the emission control strategy having the greatest probability of success in meeting the intended objective for future air quality. For example, the probability of exceeding a given threshold ozone concentration over the southeastern United States for the base case and an emission reduction case utilizing the ensemble and BMA approach is presented in Fig. 9.

Another potential approach to the probabilistic evaluation of AQMs is the use of order statistics and extreme value theory to compare the tail of observed and simulated concentration distributions. For some applications, we are particularly interested in the modeling system's ability to simulate a specific aspect of the observed distribution, such as the 4th-highest daily maximum ozone concentration over a summer season. In addition to directly comparing the observed and simulated 4th-highest concentrations, one can utilize extreme value theory to

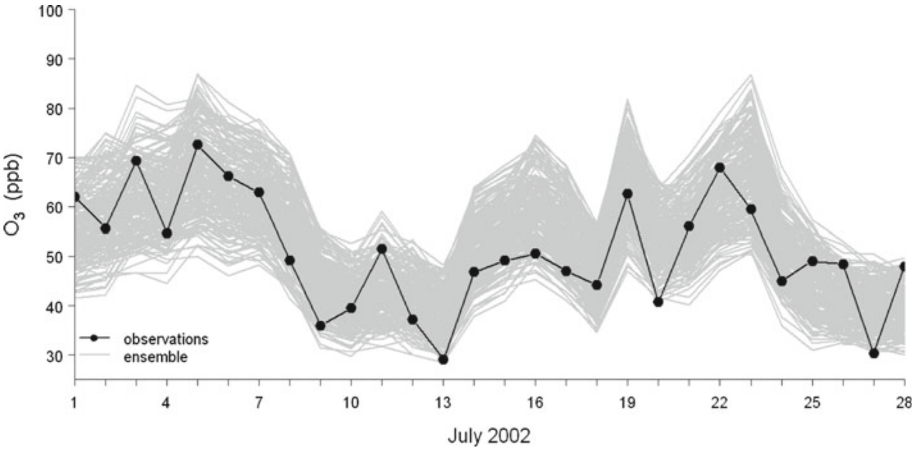


Fig. 8 Time series of daily maximum 8-h O_3 concentrations (ppb) for July 2002 at a monitoring site located in the Birmingham, Alabama metropolitan area. *Gray lines* are results from individual members of a 200-member CMAQ model ensemble; *black line/symbols* are observed data from the monitor

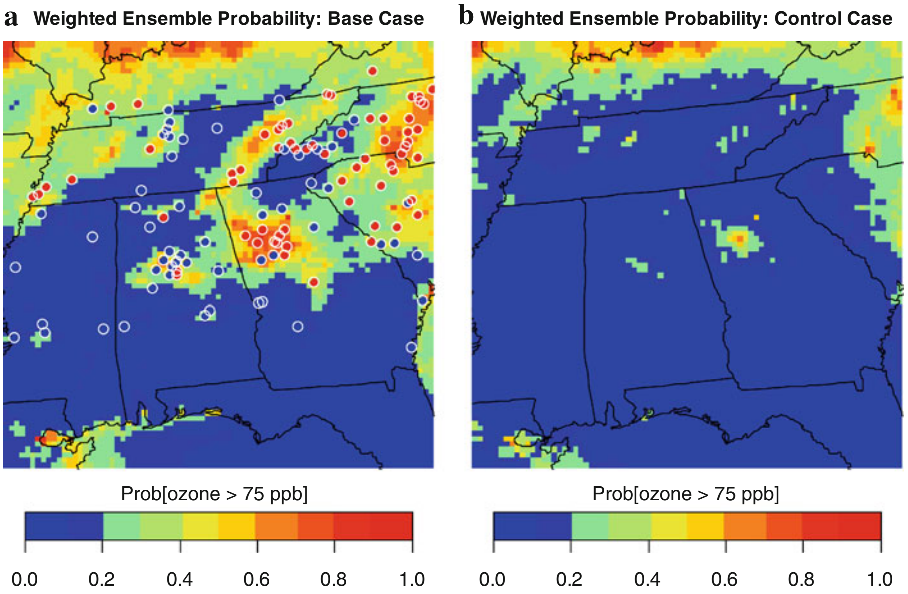

a Weighted Ensemble Probability: Base Case
b Weighted Ensemble Probability: Control Case
Prob[ozone > 75 ppb]
0.0
0.2
0.4
0.6
0.8
1.0
Prob[ozone > 75 ppb]
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 9 Spatial plots of the probability of the 4th highest daily maximum 8-h ozone concentration exceeding 75 ppb for **a** the base case CMAQ model simulation and **b** after a 50% reduction in NO_x emissions. Observations are shown in *white circles* in plot (**a**)

estimate the probability that the observed or simulated 4th-highest concentration exceeds a certain concentration threshold (say 84 ppb) or to estimate the 95% confidence bounds of the observed and simulated 4th-highest concentrations given the other sample values of the observed and simulated distributions. For example, if at a station the observed and simulated 4th-highest ozone concentration were 92 and 87 ppb, respectively, but the width of the 95% confidence interval was 5 ppb in both cases, one might conclude that these two values are not significantly different given the discrete observed and modeled sample distributions. An illustration of this approach and an application to air quality planning is provided by Hogrefe and Rao [56].

4 Summary

In this paper, we have examined approaches to the evaluation of regional-scale air quality modeling systems, as they are currently used in a variety of applications. It is evident from this examination that model evaluation exercises are based on a set of presumptions, which are often not explicitly stated. These premises are:

- Observations of air pollution contain the influences of multiple sources that vary in space and time. Further, observational values are affected by measurement uncertainties that can include instrumental errors and biases as well as spatial representativeness uncertainties.
- It should be recognized that even with the perfect model science and perfect model input and numerical algorithms, there will be differences between modeled and observed values because the model predicts the population mean while an observation is a single event out of a population, and stochastic variations embedded in the observations are not modeled in current regional-scale numerical air quality models.

Our examination of modeling practices leads us to conclude that models cannot be validated in the formal sense, but rather can be shown to have predictive and diagnostic value. The process whereby this value is demonstrated is called model evaluation. Because evaluation criteria can differ between applications, the criteria for “success” should be context-relative [10].

Our review of current practices reveals that model evaluation is driven by three broad objectives: to determine a model’s suitability for an intended application, to distinguish between models, and to guide model development. These objectives can be achieved via four types of model evaluation: *Operational Evaluation*, in which model predictions are compared with data in an overall sense using a variety of statistical measures; *Diagnostic Evaluation*, in which the relative interplay of chemical and physical processes captured by the model are analyzed to assess if the overall operation of the model is correct; *Dynamic Evaluation*, in which the ability of the modeling system to capture observed changes in emissions or meteorology is analyzed; and, *Probabilistic Evaluation*, in which various statistical techniques are used to capture joint uncertainty in model predictions and observations.

There exist many measures and techniques for quantifying model performance in an operational sense. These measures (or “standard metrics”) are often used in combination and with varying levels of utility and interpretations. A fundamental problem in using these measures is that model output (based on volume-averages) and observations (based on point-wise measurements) are in principle incommensurable, and that model predictions represent population averages while observations reflect individual events out of a population. Since this fundamental problem is generally ignored in the first three types of model evaluation, probabilistic evaluation methods are recommended.

To conduct diagnostically-oriented model evaluations, high-quality 3-D data on ambient air concentrations, emissions and meteorology are needed. These data needs are often quite extensive, and in many cases not fully met. Hence, most model evaluations to date begin and end with the operational evaluation. An outstanding example of the inadequacy of evaluation data sets is the need to resolve three-dimensional pollution fields, when only two dimensional data are available. Our understanding of pollutant transport aloft and re-entrainment in the PBL is limited due to the lack of these 3-D datasets [14]. Similarly, process evaluation of chemical sub-models often requires measurements of chemical species that are only available in specialized research studies, and not generally in routine environmental monitoring programs.

To properly address the issues related to the model evaluation, an international effort [62] is currently underway to apply the model evaluation framework presented in this paper involving several regional air quality models being used in North America and Europe (see http://aqmeii.jrc.ec.europa.eu/).

Acknowledgements Although this paper has been subjected to the U.S. Environmental Protection Agency review and approved for publication, it does not necessarily reflect the views and policies of the Agency.

References

1. Bachmann J (2007) Will the circle be unbroken: a history of the national ambient air quality standards. J Air Waste Manag Assoc 57:652–697

2. Frost GJ et al (2006) Effects of changing power plant NO_x emissions on O_3 in the eastern United States: proof of concept. J Geophys Res 111:D12306
3. Gégo E, Gilliland A, Godowitch J, Rao ST, Porter PS, Hogrefe C (2008) Modeling analyses of the effects of changes in nitrogen oxides emissions from the electric power sector on ozone levels in the eastern United States. J Air Waste Manag Assoc 58:580–588
4. Otte TL et al (2005) Linking the Eta model with the community multiscale air quality (CMAQ) modeling system to build a national air quality forecasting system. Weather Forecast 43:1648–1665
5. Mathur R (2008) Estimating the impact of the 2004 Alaskan forest fires on episodic particulate matter pollution over the eastern United States through assimilation of satellite-derived aerosol optical depths in a regional air quality model. J Geophys Res 113:D17302. doi:10.1029/2007JD009767
6. Eder B et al (2009) A demonstration of the use of national air quality forecast guidance for developing local air quality index forecasts. Bull Am Meteorol Soc doi:10.1175/2009BAMS2734.1
7. Hanna SR, Davis JM (2002) Evaluation of a photochemical grid model using estimates of concentration probability density functions. Atmos Environ 36:1793–1798
8. Fox DG (1981) Judging air quality model performance: a summary of the AMS workshop on dispersion model performance. Bull Am Meteorol Soc 62:599–609
9. Dabberdt WF et al (2004) Meteorological research needs for improved air quality forecasting. Bull Am Meteorol Soc 85:563–586
10. Steyn DG, Galmarini S (2008) Evaluating the predictive and explanatory value of atmospheric numerical models: between relativism and objectivism. Open Atmos Sci J 2: 38–45. doi:10.2174/1874282300802010038
11. Irwin JS et al (2008) A procedure for inter-comparing the skill of regional-scale air quality model simulations of daily maximum 8-hr ozone concentrations. Atmos Environ 42:5403–5412
12. Weil JC, Sykes RI, Venkatram A (1992) Evaluating air quality models: review and outlook. J Appl Meteorol 31:1121–1145
13. Swall JL, Foley KM (2009) The impact of spatial correlation and incommensurability on model evaluation. Atmos Environ 43:1204–1217
14. Rao ST (2009) Environmental monitoring and modeling needs in the 21st century. EM Magazine, October
15. Chang JC, Hanna SR (2004) Air quality model performance. Meteorol Atmos Phys 87:167–196
16. Rao ST, Zurbenko IG, Neagu R, Porter PS, Ku JY, Henry RF (1997) Space and time scales in ambient ozone data. Bull Am Meteorol Soc 78:2153–2166
17. Hogrefe C, Rao ST, Zurbenko IG, Porter PS (2000) Interpreting information in time series of ozone observations and model predictions relevant to regulatory policies in the eastern United States. Bull Am Meteorol Soc 81:2083–2106
18. Appel KW, Gilliland AB, Sarwar G, Gilliam RC (2007) Evaluation of the community multi-scale air quality (CMAQ) model version 4.5: sensitivities impacting model performance; part 1 Ozone. Atmos Environ 41(40):9603–9615
19. Gégo EL, Porter PS, Irwin JS, Hogrefe C, Rao ST (2005) Assessing the comparability of ammonium, nitrate and sulfate concentrations measured by three air quality monitoring networks. Pure Appl Geophys 162:1919–1939
20. Kang D, Mathur R, Rao ST, Yu S (2008) Bias-adjustment techniques for improving ozone air quality forecasts. J Geophys Res 113(D23308):1–17
21. Appel KW, Bhave PV, Gilliland AB, Sarwar G, Roselle SJ (2008) 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
22. Mathur R, Yu S, Kang D, Schere KL (2008) Assessment of the wintertime performance of developmental particulate matter forecasts with the Eta-community multiscale air quality modeling system. J Geophys Res 113:D02303. doi:10.1029/2007JD008580
23. Morris RE, McNally DE, Tesche TW, Tonnesen G, Boylan JW, Brewer P (2005) Preliminary evaluation of the community multiscale air quality model for 2002 over the southeastern United States. J Air Waste Manag Assoc 55:1694–1708
24. Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 106(D7):7183–7192
25. Vautard R, Builtzes PHJ, Thunis P, Cuvelier C, Bedogni M, Bessagnet B, Honore C, Moussiopoulos N, Pirovano G, Schaap M, Stern R, Tarrason L, Wind P (2007) Evaluation and intercomparison of ozone and PM10 simulations by several chemistry transport models over four European cities within the CityDelta project. Atmos Environ 41:173–188
26. Stohl A, Hittenberger M, Wotawa G (1998) Validation of the Lagrangian particle dispersion model FLEXPART against large-scale tracer experiment data. Atmos Environ 32:4245–4264

27. Godowitch JM, Hogrefe C, Rao ST (2008) Diagnostic analyses of a regional air quality model: changes in modeled processes affecting ozone and chemical-transport indicators from NO_x point source emission reductions. J Geophys Res 113. doi:10.1029/2007JD009537
28. Saltelli A, Tarantola S, Campolongo F, Ratto M (2004) Sensitivity analysis in practice, a guide to assessing scientific models. Wiley, New York
29. Cullen AC, Frey HC (1999) Probabilistic techniques in exposure assessment. A handbook for dealing with variability and uncertainty in models and inputs. Plenum Press, New York, 335 pp
30. Biswas J, Rao ST (2001) Uncertainties in episodic ozone modeling stemming from uncertainties in the meteorological fields. J Appl Meteorol 40:117–136
31. Otte TL (2008) The impact of nudging in the meteorological model for retrospective air quality simulations. Part II: Evaluating collocated meteorological and air quality observations. J Appl Meteorol Climatol 47:1868–1887
32. Gilliland AB, Hogrefe C, Pinder RW, Godowitch JM, Foley KL, Rao ST (2008) Dynamic evaluation of regional air quality models: assessing changes in O_3 stemming from changes in emissions and meteorology. Atmos Environ 42:5110–5123
33. Godowitch JM, Gilliland AB, Draxler R, Rao ST (2008) Modeling assessment of point source NO_x emission reductions on ozone air quality in the eastern United States. Atmos Environ 42:87–100
34. Pinder RW, Adams PJ, Pandis SN, Gilliland AB (2006) Temporally resolved ammonia emission inventories: current estimates, evaluation tools, and measurement needs. J Geophys Res Atmos 111. doi:10.1029/2005JD006603
35. Cohan DS, Hakami A, Hu Y, Russell AG (2005) Nonlinear response of ozone to emissions: source apportionment and sensitivity analysis. Environ Sci Technol 39:6739–6748
36. Napelenok SL, Cohan DS, Hu Y, Russell AG (2006) Decoupled direct 3D sensitivity analysis for particulate matter (DDM-3D/PM). Atmos Environ 40:6112–6121
37. Seaman NL (2000) Meteorological modeling for air-quality assessments. Atmos Environ 34:2231–2259
38. Bhave PV, Pouliot GA, Zheng M (2007) Diagnostic model evaluation for carbonaceous $PM_{2.5}$ using organic markers measured in the southeastern U.S. Environ Sci Technol 41:1577–1583
39. Napelenok SL, Pinder RW, Gilliland AB, Martin RV (2008) A method for evaluating spatially-resolved NO_x emissions using Kalman filter inversion, direct sensitivities, and space-based NO_2 observations. Atmos Chem Phys 8:6469–6499
40. Chow JC (2003) Introduction to special topic: weekend and weekday differences in ozone levels. J Air Waste Manag Assoc 53:771
41. Godowitch JM, Pouliot G, Rao ST (2009) On the use of a dynamic evaluation approach to assess multiyear change in modeled and observed urban nitrogen oxide concentrations. In: Steyn DG, Rao ST (eds) Proceedings of the 30th NATO/SPS international technical meeting on air pollution, modeling and its application, San Francisco, CA
42. Environmental Protection Agency (2008) NOx budget trading program, EPA-430-R-08-008, 62 pp
43. Porter PS, Rao ST, Zurbenko IG, Dunker AM, Wolff GT (2001) Ozone air quality over North America: part II—An analysis of trend detection and attribution techniques. J Air Waste Manag Assoc 51:283–306
44. Camalier L, Cox W, Dolwick P (2007) The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. Atmos Environ 41:7127–7137
45. Lewellen WS, Sykes RI, Parker SF (1985) An evaluation technique which uses the prediction of both concentration mean and variance. In: Proceedings of the DOE/AMS air pollution model evaluation workshop, Savannah river lab report number DP-1701-1, section 2, 24 pp
46. Irwin JS, Rao ST, Petersen WB, Turner DB (1987) Relating error bounds for maximum concentration estimates to diffusion meteorology uncertainty. Atmos Environ 21:1927–1937
47. Galmarini S et al (2004) Ensemble dispersion forecasting, part I: concept approach, and indicators. Atmos Environ 38:4607–4617
48. Galmarini S et al (2004) Ensemble dispersion forecasting, part II: application and evaluation. Atmos Environ 38:4619–4632
49. Dabberdt WF, Miller E (2000) Uncertainty, ensembles and air quality dispersion modeling: applications and challenges. Atmos Environ 34:4667–4673
50. Delle Monache L, Deng X, Zhou Y, Stull R (2006) Ozone ensemble forecasts: 1 A new ensemble design. J Geophys Res 111:D05307. doi:10.1029/2005JD006310
51. Mallet V, Sportisse B (2006) Ensemble-based air quality forecasts: a multimodel approach applied to ozone. J Geophys Res 111:D18302. doi:10.1029/2005JD006675
52. McKeen S et al (2005) Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. J Geophys Res 110:D21307. doi:10.1029/2005JD005858
53. van Loon M et al (2007) Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble. Atmos Environ 41:2083–2097

54. Riccio A, Giunta G, Galmarini S (2007) Seeking for the rational basis of the median model: the optimal combination of multi-model ensemble results. Atmos Chem Phys 7:6085–6098
55. Warner TT, Sheu R-S, Bowers JF, Sykes RI, Dodd GC, Henn DS (2002) Ensemble simulations with coupled atmospheric dynamic and dispersion models: illustrating uncertainties in dosage simulations. J Appl Meteorol 41:488–504
56. Hogrefe C, Rao ST (2001) Demonstrating attainment of the air quality standards: integration of observations and model predictions into the probabilistic framework. J Air Waste Manag Assoc 51:1060–1072
57. Jones JM, Hogrefe C, Henry RF, Ku J-Y, Sistla G (2005) An assessment of the sensitivity and reliability of the relative reduction factor (RRF) approach in the development of 8-hr ozone attainment plans. J Air Waste Manag Assoc 55:13–19
58. Hogrefe C, Civerolo KL, Hao W, Ku JY, Zalewsky EE, Sistla G (2008) Rethinking the assessment of photochemical modeling systems in air quality planning applications. J Air Waste Manag Assoc. doi:10.3155-1047-3289.58.8.1086
59. Raftery AE, Gneiting T, Balabdaoui F, Palokowski M (2005) Using Bayesian model averaging to calibrate ensembles. Mon Weather Rev 133:1155–1174
60. Pinder RW, Gilliam RC, Appel KW, Napelenok SL, Gilliland AB (2009) Efficient probabilistic estimates of surface ozone concentration using an ensemble of model configurations and direct sensitivity calculations. Environ Sci Technol 43:2388–2393
61. Foley K, Pinder R, Napelenok S (2008) New directions in air quality model evaluation: probabilistic model evaluation. Poster presented at the CMAS Conference, Chapel Hill, NC. Available at http://www.cmascenter.org/conference/2008/agenda.cfm
62. Rao ST, Schere KS, Galmarini S, Steyn DG (2009) AQMEII: air quality model evaluation international initiative. In: Proceedings of the 30th NATO/SPS international technical meeting on air pollution modeling and its application, San Francisco