



Published in final edited form as:

Environ Fluid Mech (Dordr). 2010 ; 10(4): 471–489. doi:10.1007/s10652-009-9163-2.

A FRAMEWORK FOR EVALUATING REGIONAL-SCALE NUMERICAL PHOTOCHEMICAL MODELING SYSTEMS

Robin Dennis^{a,*}, Tyler Fox^b, Montse Fuentes^c, Alice Gilliland^a, Steven Hanna^d, Christian Hogrefe^e, John Irwin^f, S.Trivikrama. Rao^{a,**}, Richard Scheffe^b, Kenneth Schere^a, Douw Steyn^g, and Akula Venkatram^h

^aAtmospheric Modeling and Analysis Division, National Exposure Research Laboratory, US Environmental Protection Agency, RTP, NC 27711 USA

^bAir Quality Assessment Division, Office of Air Quality Planning and Standards, US Environmental Protection Agency, RTP, NC 27711 USA

^cDepartment of Statistics, North Carolina State University, Raleigh, NC 27695 USA

^dHanna Consultants, Kennebunkport, ME 04046 USA

^eBureau of Air Quality Analysis and Research, NYS Dept. of Environmental Conservation, Albany, NY 12233 USA

^fJohn S. Irwin and Associates, Raleigh, NC 27615 USA

^gDepartment of Earth and Ocean Sciences, The University of British Columbia, Vancouver, BC, V6T1Z4 Canada

^hDepartment of Mechanical Engineering, University of California, Riverside, CA 92521 USA

Abstract

This paper discusses the need for critically evaluating regional-scale (~200-2000 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, 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.

** Corresponding author: U.S. EPA – E243-02, R.T.P., NC 27711; rao.st@epa.gov; phone: 919-541-4542; fax: 919-541-1379 .

* Authors are listed in alphabetical order.

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 simulation models (AQMs) play a key role in the development and implementation of air pollution control rules and regulations in the U.S.A. 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 assessment of the degree to which an AQM can simulate the spatio-temporal features imbedded in air quality data. This paper discusses the approaches for rigorously evaluating three-dimensional photochemical AQMs.

Over the last several 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 not designed 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.A. 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 (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 are used in widely differing applications ranging from research on atmospheric processes to air quality forecasting. For regulatory applications, a model must be able to provide an adequate description of the relationships among atmospheric processes and variables in addition to adequate quantitative estimates of species concentrations. By contrast, a forecast model is judged 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-term emissions. In this application, we focus on diagnostic 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 sometimes need to compare the relative performance of different models in explaining the observations. Evaluation procedures must 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 within the model. 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. “**Probabilistic evaluation**” acknowledges the uncertainty in model inputs and formulation of processes by focusing 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.

Operational Evaluation

Operational evaluations make use of routine observations of ambient pollutant concentrations, emissions, meteorology, and other relevant variables. Typical modeled variables used in operational model exercises for air quality include the meteorological state and derived variables: temperature, moisture (humidity), wind speed and direction, planetary boundary layer height, surface radiation, clouds and precipitation. Air quality variables include ozone (O_3), carbon monoxide (CO), nitrogen oxides (NO , NO_x), and fine particulate matter mass and its species ($PM_{2.5}$, SO_4 , NO_3 , NH_3 , OC, EC).

There are three performance measures that are widely being used in AQM evaluation (and most other types of model evaluation) – mean bias (MB), root mean square error (RMSE), and correlation (R) [12]. Sometimes, the statistical confidence levels in these statistics are 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.

Limitations of Standard Metrics—The standard metrics, namely, MB, RMSE, and R, do not take into consideration that predictions from three-dimensional regional AQM models are volume-averaged ensemble mean 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, note that the smoothing technique relies on a statistical model to combine observations, and, thus, the comparison is tantamount to a comparison of the results of two different models, and not a direct comparison of model output and corresponding observations. Further, it is important to note that 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 the numerical photochemical AQMs. 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/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 [14]. 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 data-based grid cell estimates 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 [15,16]. 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 shows how the model’s fidelity is greatest 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 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).

Graphical Techniques—Some graphical techniques in operational model evaluation have been alluded to earlier in conjunction with standard statistical metrics. Scatter plots of percentile values and time-series plots are useful for regional AQM analyses [5,17]. It is useful to aggregate the results across coherent space and/or time regions based on Principal Component Analysis to represent distributional quantities, and not single point observations [18,19]. For example, O₃ concentration distributions over all monitoring sites in a region can be plotted as a daily time series over a month or longer period for model results and observations. The hourly O₃ 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 speciated bar graphs are useful for comparing simulated and observed chemical constituents of size-segregated particulate matter [20]. Performance goal plots (“soccer” plots) that summarize model performance by plotting performance goals and criteria for fractional bias versus fractional error, and concentration performance plots (“bugle” plots) that display fractional bias or error as a function of concentration have been suggested [21]. Taylor diagram [22] which combines model error and correlation in a single point, has been found to be useful for comparing the performance of several models [23].

For regional models in particular, a basic comparison of the 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 [24] and e-folding distance [25].

Emission models are part of regional AQM systems, and, hence, they need to be evaluated. Generally, estimates from emissions models cannot be directly compared with observed values because emission observations 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 in emissions estimates for these point sources, and, thus, are assimilated into the AQM through the emissions inputs. 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 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 difficulty of real-world emission measurements for AQMs, there are diagnostic methods that may provide insights into biases and errors in the emissions. These techniques will be discussed as part of the next section.

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 processes. 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., ten-minute 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. Use of 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 [26,27]. 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: Separating roles of model inputs from modeled processes—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. Sensitivity tests are one of the most common ways to ascertain whether inputs have a notable influence on model performance issues. A sensitivity test examines the response of a model's outputs to perturbations in the model's inputs. A fundamental description of sensitivity analyses of environmental models is given by Saltelli et al. [28], and Cullen and Frey [26] provide specific discussions related to AQMs. However, because of the nonlinear characteristics of regional AQMs, the sensitivity test may only be valid for a certain 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 [29,30]. 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 [31,32].

Advanced instrumented modeling tools (e.g. sulfur tacking 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, 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 $\text{PM}_{2.5}$ predictions to emission precursors, source regions and sectors, boundary conditions, and more is calculated during the model

simulations [35-37]. The DDM tool is able to capture both the first and second order sensitivities to these inputs, which is important for these non-linear chemical systems. We will next discuss some examples of diagnostic evaluation studies that identified key meteorological and emission issues that can play a strong role in AQM performance.

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 [38] 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 [30] 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 (Figure 2b). The results in Figure 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 also help to inform the emission inventory of potential missing sources. 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 [39] and used detailed tracking of emission contributions across space for inverse modeling [40]. In all cases, top-down methodologies can inform improvements needed to bottom-up inventories that are critical for AQM performance.

Dynamic Evaluation

Dynamic evaluation looks at a retrospective case(s) to evaluate whether the model has properly predicted air quality response to known emission reductions and/or meteorological changes. The change in concentration is being 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). An example of dynamic evaluation would be modeling assessments of the weekday/weekend concentration differences where mobile source emissions are known to significantly change [41]. These studies can provide insight into the ozone response to NO_x emissions in core urban areas with very dense mobile emissions. However, there can be fairly substantial uncertainty in the estimate of these mobile emissions as well as in modeling the impacts of roadways in a regional model. More recently, an evaluation of an AQM’s response to a regulatory

emission reduction program has been assessed [25,31,42]. 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. 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 changes are also being influenced by different meteorological conditions from one 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].

Probabilistic Evaluation

All regional numerical AQMs use first-order closure and model outputs represent population mean. 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 predictions. There are no widely-used prescribed methods for determining such confidence through a probabilistic evaluation. A method was suggested by Lewellen et al. [45] that 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 meteorological input uncertainty, using a probability distribution function (pdf), to air quality predictions.

Yet another technique uses an ensemble of modeling methods to define the pdf [47,48]. The ensemble method is a subset of a full Monte Carlo uncertainty exercise, where a few model runs are made using varying inputs and other assumptions in hopes that the limited number of runs 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. [49], 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 [50-52] 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-hr ozone concentrations, much smaller than the effect on absolute concentrations for the “base case” simulation. Bayesian Model Averaging (BMA) [53] has been used to calibrate the ensemble predictions by weighting each individual ensemble member generated in the Pinder et al. study [54] 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 [55]. 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. As an 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 Figure 3.

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 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 [50].

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 all possible sources and scales of source variation in time and space and have measurement uncertainties. Measurements are taken at specific locations.
- It should be recognized that even with the perfect model science and perfect model input and numerical algorithms, there will be differences between modeled values observations because the model predicts the population mean while any given observation is a single event out of a population.

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 the evaluation criteria appear to be different in different 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 model 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 (which we have called “standard metrics”) are often used in combinations, and have varying levels of utility and interpretations. A fundamental difficulty lies in the fact that model output (based on volume-averages) and observations (based on pointwise 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, we need probabilistic evaluation.

In order to conduct diagnostically-oriented model evaluations, high-quality data on ambient air quality, 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 [55]. 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 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/>).

Acknowledgments

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. Will the circle be unbroken: A history of the National Ambient Air quality Standards. *Journal of Air & Waste Management Association*. 2007; 57:652–697.
2. Frost GJ, et al. Effects of changing power plant NO_x emissions on O₃ in the eastern United States: Proof of concept. *Journal of Geophysical Research*. 2006; 111:D12306.
3. Gégó E, Gilliland A, Godowitch J, Rao ST, Porter PS, Hogrefe C. Modeling analyses of the effects of changes in nitrogen oxides emissions from the electric power sector on ozone levels in the eastern United States. *Journal of Air and Waste Management Association*. 2008; 58:580–588.
4. Otte TL, et al. Linking the Eta Model with the Community Multiscale Air Quality (CMAQ) Modeling System to build a national air quality forecasting system. *Weather and Forecasting*. 2005; 43:1648–1665.
5. Mathur R. 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.* 2008; 113:D17302. doi: 10.1029/2007JD009767.
6. Eder B, et al. A demonstration of the use of national air quality forecast guidance for developing local air quality index forecasts. *Bulletin of the American Meteorological Society*. 2009 (in press).
7. Hanna SR, Davis JM. Evaluation of a photochemical grid model using estimates of concentration probability density functions. *Atmospheric Environment*. 2002; 36:1793–1798.

8. Fox DG. Judging air quality model performance: A summary of the AMS Workshop on Dispersion Model Performance. *Bulletin of American Meteorological Society*. 1981; 62:599–609.
9. Dabberdt WF, Carroll MA, Baumgardner D, Carmichael G, Cohen R, Dye T, Ellis J, Grell G, Grimmond S, Hanna S, Irwin J, Lamb B, Madronich S, McQueen J, Meagher J, Odman T, Pleim J, Schmid HP, Westphal DL. Meteorological research needs for improved air quality forecasting. *Bulletin of American Meteorological Society*. 2004; 85:563–586.
10. Steyn DG, Galmarini S. Evaluating the Predictive and Explanatory Value of Atmospheric Numerical Models: Between Relativism and Objectivism. *The Open Atmospheric Science Journal*. 2008; 2:38–45. doi:10.2174/1874282300802010038.
11. Irwin JS, Civerolo K, Hogrefe C, Appel W, Foley K, Swall J. A procedure for inter-comparing the skill of regional-scale air quality model simulations of daily maximum 8-hr ozone concentrations. *Atmospheric Environment*. 2008; 42:5403–5412.
12. Weil JC, Sykes RI, Venkatram A. Evaluating air quality models: review and outlook. *Journal of Applied Meteorology*. 1992; 31:1121–1145.
13. Swall JL, Foley KM. The impact of spatial correlation and incommensurability on model evaluation. *Atmospheric Environment*. 2009; 43:1204–1217.
14. Chang JC, Hanna SR. Air quality model performance. *Meteorology and Atmospheric Physics*. 2004; 87:167–196.
15. Rao ST, Zurbenko IG, Neagu R, Porter PS, Ku JY, Henry RF. Space and time scales in ambient ozone data. *Bulletin of American Meteorological Society*. 1997; 78:2153–2166.
16. Hogrefe C, Rao ST, Zurbenko IG, Porter PS. Interpreting information in time series of ozone observations and model predictions relevant to regulatory policies in the eastern United States. *Bulletin of American Meteorological Society*. 2000; 81:2083–2106.
17. Appel KW, Gilliland AB, Sarwar G, Gilliam RC. Evaluation of the Community Multi-scale Air Quality (CMAQ) model Version 4.5: Sensitivities impacting model performance; Part 1 Ozone. *Atmospheric Environment*. 2007; 41(40):9603–9615.
18. Gégo EL, Porter PS, Irwin JS, Hogrefe C, Rao ST. Assessing the comparability of ammonium, nitrate and sulfate concentrations measured by three air quality monitoring networks. *Pure and Applied Geophysics*. 2005; 162:1919–1939.
19. Kang D, Mathur R, Rao ST, Yu S. Bias-adjustment techniques for improving ozone air quality forecasts. *Journal of Geophysical Research*. 2008; 113(D23308):1–17.
20. Appel KW, Bhawe PV, Gilliland AB, Sarwar G, Roselle SJ. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance; Part II - particulate matter. *Atmospheric Environment*. 2008; 42:6057–6066.
21. Morris RE, McNally DE, Tesche TW, Tonnesen G, Boylan JW, Brewer P. Preliminary evaluation of the Community Multiscale Air Quality Model for 2002 over the southeastern United States. *Journal of Air & Waste Management Association*. 2005; 55:1694–1708.
22. Taylor KE. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*. 2001; 106(D7):7183–7192.
23. Vautard R, et al. Evaluation and intercomparison of Ozone and PM10 simulations by several chemistry transport models over four European cities within the CityDelta project. *Atmospheric Environment*. 2007; 41:173–188.
24. Stohl A, Hittenberger M, Wotawa G. Validation of the lagrangian particle dispersion model FLEXPART against large-scale tracer experiment data. *Atmospheric Environment*. 1998; 32:4245–4264.
25. Godowitch JM, Hogrefe C, Rao ST. 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. *Journal of Geophysical Research*. 2008b; 113 doi:10.1029/2007JD009537.
26. Cullen, AC.; Frey, HC. Probabilistic Techniques in Exposure Assessment. A Handbook for Dealing with Variability and Uncertainty in Models and Inputs. Plenum Press; New York: 1999. p. 335
27. Hudman RC, et al. Surface and lightning sources of nitrogen oxides over the United States: magnitudes, chemical evolution, and outflow. *Journal of Geophysical Research*. 2007; 112:D12S05. doi:10.1029/2006JD007912.

28. Saltelli, A.; Tarantola, S.; Campolongo, F.; Ratto, M. *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*. John Wiley & Sons; 2004.
29. Biswas J, Rao ST. Uncertainties in episodic ozone modeling stemming from uncertainties in the meteorological fields. *Journal of Applied Meteorology*. 2001; 40:117–136.
30. Otte TL. The Impact of Nudging in the Meteorological Model for Retrospective Air Quality Simulations. Part II: Evaluating Collocated Meteorological and Air Quality Observations. *Journal of Applied Meteorology and Climatology*. 2008; 47:1868–1887.
31. Gilliland AB, Hogrefe C, Pinder RW, Godowitch JM, Foley KL, Rao ST. Dynamic evaluation of regional air quality models: Assessing changes in O₃ stemming from changes in emissions and meteorology. *Atmospheric Environment*. 2008; 42:5110–5123.
32. Godowitch JM, Gilliland AB, Draxler R, Rao ST. Modeling assessment of point source NO_x emission reductions on ozone air quality in the eastern United States. *Atmospheric Environment*. 2008a; 42:87–100.
33. Pinder RW, Adams PJ, Pandis SN, Gilliland AB. Temporally resolved ammonia emission inventories: Current estimates, evaluation tools, and measurement needs. *Journal of Geophysical Research-Atmospheres*. 2006; 111 doi:10.1029/2005JD006603.
34. Vizuete W, Kioumourtzoglou M, Jeffries H, Henderson B. Effects of radical source strengths on ozone formation in models for Houston, Texas. *Atmospheric Environment*. 2008a in review at.
35. Vizuete W, et al. Modeling ozone formation from industrial emission events in Houston, Texas. *Atmospheric Environment*. 2008b doi: ATMENV-D-07-01368R2.
36. Cohan DS, Hakami A, Hu Y, Russell AG. Nonlinear response of ozone to emissions: Source apportionment and sensitivity analysis. *Environmental Science and Technology*. 2005; 39:6739–6748. [PubMed: 16190234]
37. Napelenok SL, Cohan DS, Hu Y, Russell AG. Decoupled direct 3D sensitivity analysis for particulate matter (DDM-3D/PM). *Atmospheric Environment*. 2006; 40:6112–6121.
38. Seaman NL. Meteorological modeling for air-quality assessments. *Atmospheric Environment*. 2000; 34:2231–2259.
39. Bhave PV, Pouliot GA, Zheng M. Diagnostic model evaluation for carbonaceous PM_{2.5} using organic markers measured in the southeastern U.S. *Environmental Science & Technology*. 2007; 41:1577–1583. [PubMed: 17396644]
40. Napelenok SL, Pinder RW, Gilliland AB, Martin RV. A method for evaluating spatially-resolved NO_x emissions using Kalman filter inversion, direct sensitivities, and space-based NO₂ observations. *Atmospheric Chemistry and Physics*. 2008; 8:6469–6499.
41. Chow JC. Introduction to Special Topic: Weekend and weekday differences in ozone levels. *Journal of Air & Waste Management Association*. 2003; 53:771.
42. Godowitch, JM.; Pouliot, G.; Rao, ST. On the use of a dynamic evaluation approach to assess multi-year change in modeled and observed urban nitrogen oxide concentrations. In: Steyn, DG.; Rao, ST., editors. *Proceedings of the 30th NATO/SPS International Technical Meeting on Air Pollution, Modeling and Its Application*; San Francisco, CA. 2009.
43. Porter PS, Rao ST, Zurbenko IG, Dunker AM, Wolff GT. Ozone air quality over North America: Part II-An analysis of trend detection and attribution techniques. *Journal of Air & Waste Management Association*. 2001; 51:283–306.
44. Camalier L, Cox W, Dolwick P. The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmospheric Environment*. 2007; 41:7127–7137.
45. Lewellen, WS.; Sykes, RI.; Parker, SF. An evaluation technique which uses the prediction of both concentration mean and variance. *Proceedings of the DOE/AMS Air Pollution Model Evaluation Workshop*; 1985. p. 24 Savannah River Lab Report Number DP-1701-1, Section 2
46. Irwin JS, Rao ST, Petersen WB, Turner DB. Relating error bounds for maximum concentration estimates to diffusion meteorology uncertainty. *Atmospheric Environment*. 1987; 21:1927–1937.
47. Galmarini S, et al. Ensemble dispersion forecasting, part I: concept approach, and indicators. *Atmospheric Environment*. 2004a; 38:4607–4617.
48. Galmarini S, et al. Ensemble dispersion forecasting, part II: application and evaluation. *Atmospheric Environment*. 2004b; 38:4619–4632.

49. Warner TT, Sheu R-S, Bowers JF, Sykes RI, Dodd GC, Henn DS. Ensemble simulations with coupled atmospheric dynamic and dispersion models: Illustrating uncertainties in dosage simulations. *Journal of Applied Meteorology*. 2002; 41:488–504.
50. Hogrefe C, Rao ST. Demonstrating attainment of the air quality standards: integration of observations and model predictions into the probabilistic framework. *Journal of Air & Waste Management Association*. 2001; 51:1060–1072.
51. Jones JM, Hogrefe C, Henry RF, Ku J-Y, Sistla G. An assessment of the sensitivity and reliability of the relative reduction factor (RRF) approach in the development of 8-hr ozone attainment plans. *Journal of Air & Waste Management Association*. 2005; 55:13–19.
52. Hogrefe C, Civerolo KL, Hao W, Ku JY, Zalewsky EE, Sistla G. Rethinking the assessment of photochemical modeling systems in air quality planning applications. *J. Air & Waste Manage. Assoc.* 2008 doi10.3155-1047-3289.58.8.1086.
53. Raftery AE, Gneiting T, Balabdaoui F, Palokowski M. Using Bayesian Model Averaging to Calibrate Ensembles. *Monthly Weather Review*. 2005; 133:1155–1174.
54. Pinder RW, Gilliam RC, Appel KW, Napelenok SL, Gilliland AB. Efficient probabilistic estimates of surface ozone concentration using an ensemble of model configurations and direct sensitivity calculations. *Environmental Science & Technology*. 2009; 43:2388–2393. [PubMed: 19452891]
55. Foley, K.; Pinder, R.; Napelonak, S. New Directions in Air Quality Model Evaluation: Probabilistic Model Evaluation; Poster presented at the CMAS Conference; Chapel Hill, NC. 2008. Available from <http://www.cmascenter.org/conference/2008/agenda.cfm>



Figure 1.
A framework for evaluating regional-scale air quality models.

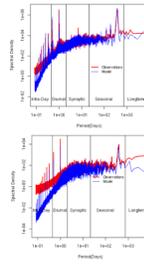


Figure 2.

(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 2005; (b) same as (a) except for wind speed time series.

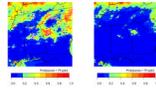


Figure 3. Spatial plots of the probability of the 4th highest daily maximum 8-hr ozone concentration exceeding 75 ppb for (a) the base case CMAQ model simulation and (b) after a 50% reduction in NO_x emissions