ON THE EVALUATION OF REGIONAL-SCALE PHOTOCHEMICAL AIR QUALITY 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], Akula Venkatram[h]

[a]*Atmospheric Modeling Division, National Exposure Research Laboratory, US Environmental Protection Agency, RTP, NC 27711 USA*

[b]*Air Quality Assessment Division, Office of Air Quality Planning and Standards, US Environmental Protection Agency, RTP, NC 27711 USA*

[c]*Department of Statistics, North Carolina State University, Raleigh, NC 27695 USA*

[d]*Hanna Consultants, Kennebunkport, ME 04046 USA*

[e]*Bureau of Air Quality Analysis and Research, NYS Dept. of Environmental Conservation, Albany, NY 12233 USA*

[f]*John S. Irwin and Associates, Raleigh, NC 27615 USA*

[g]*Department of Earth and Ocean Sciences, The University of British Columbia, Vancouver, BC Canada*

[h]*Department of Mechanical Engineering, University of California, Riverside, CA 92521 USA*

[*] Authors are members of EPA/AMS Regional Model Evaluation Workshop Steering Committee, and are listed in alphabetical order.

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

Submitted to Atmospheric Environment, 4 September 2008

Abstract

This work provides a comprehensive view of the process of evaluating regional-scale (~200-2000 km) three-dimensional numerical photochemical air quality modeling systems, including meteorological, emissions, and air quality components. We have examined approaches to the evaluation of regional air quality modeling systems, as they are currently used in a variety of applications. From this examination, we conclude that such models cannot be validated in the formal sense, but rather can be shown to have predictive and diagnostic value. A framework for model evaluation is introduced here to provide a context for the evaluation process. The objectives of the model evaluation process include determining the suitability of a modeling system for a specific application, distinguishing the performance between different models through confidence-testing of model results, and guiding further model development. The evaluation framework presents some methods for operational, diagnostic, dynamic, and probabilistic model evaluation. Operational evaluation techniques include statistical and graphical analyses aimed at determining whether the estimated values of the modeled variables are comparable to measurements in an overall sense. Diagnostic evaluation focuses on process-oriented analyses that 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 predict changes in air quality given changes in source emissions or meteorology, the principal forces that drive the air quality model. Probabilistic evaluation attempts to assess the level of confidence in the model predictions through techniques such as ensemble model simulations. Current and emerging needs for observational data from the model evaluation perspective are also discussed, as well as the

challenges in using point measurement data in comparisons with volume-averaged model output from three-dimensional numerical models.

Keywords: air quality model, photochemical model, model evaluation, diagnostic evaluation

1. Introduction

Regional-scale (spatial scale on the order of ~200-2000 km) three-dimensional numerical photochemical air quality simulation models (AQMs) are being used for air quality management decisions and for short-term forecasting of air quality. These models play a key role in the development and implementation of air pollution control rules and regulations in the U.S. and elsewhere (Bachmann, 2007). They are used to inform the selection of particular source emissions controls since AQMs can predict the efficacy of different control strategies in reducing pollutant concentrations to the level of the relevant air quality standards. These models are also used for tracking and evaluating air quality management programs through accountability studies, such as those recently completed on the large NO_x emissions reductions from the power generation sector in the eastern U.S. (Frost et al., 2006; Gégo et al., 2008, 2007;Gilliland et al., 2008). To build confidence in the model estimates, a model must be critically evaluated to assess whether it is properly simulating the spatial and temporal features imbedded in air quality observations on the scales resolved by the model. The evaluation also assesses whether the physical and chemical processes are simulated correctly in the model, leading to proper model response to changes in meteorology and emissions, the principal classes of input data required by AQMs. To this end, a new perspective is needed to establish the best methods for assessing the performance of regional-scale AQMs.

Over the last several decades, there were several workshops and position statements discussing the evaluation of AQMs and the importance of better characterizing model uncertainties (e.g., Hanna and Gifford, 1971; Fox, 1981, 1984; Demerjian, 1985; Dabberdt et al.,

2004; NRC, 2007). Suggestions for model evaluation methods have been provided to account for the fact that models do not predict stochastic variations seen in observations (Venkatram, 1979, 1988; Weil et al., 1992; ASTM, 2005; Dabberdt et al., 2004). For the most part however, previous workshops and position statements have addressed short-range to mesoscale range AQMs rather than regional-scale three-dimensional numerical photochemical modeling systems. While our scientific understanding, model developments, and computational capabilities have grown tremendously over the last few decades, the model performance methodologies set forth in the 1980's and 1990's still represent the most comprehensive effort to date to provide guidance for AQM evaluation. However, while many of the earlier methods can be extended from local and mesoscale AQMs to regional AQMs, there are limitations in the extension of some of the evaluation procedures and metrics. Obviously, the temporal and spatial scales of the modeled phenomena are significantly different between these model types, as are the density and characteristics of the observations available for model evaluation.

During August 7-8, 2007 the U.S. Environmental Protection Agency (EPA) and the American Meteorological Society (AMS) convened an invited group of nearly 100 experts to a Workshop to (a) discuss and determine the most appropriate current methods in use in regional AQM evaluation exercises, (b) discuss new approaches to advance air quality and related model evaluation methods and procedures, and (c) develop a set of recommendations for model evaluation methods, procedures, and metrics for different components of the regional AQMs for further testing and use by the air quality modeling community. Workshop sessions focused on topics including: evaluating the performance of meteorological processes within regional-scale AQM systems, evaluating the performance of source and sink processes within AQM systems,

evaluating the performance of chemistry and aerosol processes within AQM systems, and methods and processes for evaluating the performance of AQM system components. Stemming from the discussions at and following this Workshop, this paper discusses current and proposed procedures most relevant to the evaluation of regional-scale numerical photochemical AQMs.

2. Model Evaluation Framework

We begin by agreeing with Oreskes et al. (1994) that, as for all environmental model systems, AQMs cannot be validated, in the sense of being proved "true", since "truth" (i.e., all possible conditions) is in principle inaccessible to us. We do assert, however, that such modeling systems do have both predictive and diagnostic (process-oriented) value, and that this value must be demonstrated through model evaluation exercises. Russell and Dennis (2000), in a critical review of regional-scale photochemical air quality modeling, define model evaluation as:

> " *Evaluation*: assessment of the adequacy and correctness of the science represented in the model through comparison against empirical data, such as laboratory tests, in situ tests and the analysis of natural analogs. Evaluation is a process of model confirmation relative to current understanding. Multiple, confirmatory evaluations can never demonstrate the veracity of a model: confirmation is a matter of degree. However, an evaluation can raise doubts about the science in a model."

The approach suggests that all models are wrong in some sense, and right in some other sense. Our responsibility is to discover in which way(s) our models are "right", and then only use the model in those way(s). It then becomes clear that utility of a model cannot be established on an all-or-nothing basis. If this were so, establishing complete utility would be tantamount to

establishing the "truth" of the model or, conversely, showing that a model has no utility is effectively invalidation. These extremes are logically and philosophically unattainable. It is however, possible to establish that the model has some utility, but it is clear that the utility must be a continuously variable measure.

Three-dimensional time-dependent numerical models of the atmosphere exist for processes and phenomena at a wide range of spatial and temporal scales, and are used in widely differing applications (from research to policy-making). For example, Computational Fluid Dynamics (CFD) models are applied to domains of a few hundred meters and grid sizes of 1 to 5 m, while regional photochemical AQMs are applied to domains of 200 to 2000 km with grid cell sizes ranging from 2 to 40 km. This paper addresses the issues relevant to the regional scale. Since most processes in the atmosphere are scale-dependent, it would be very surprising if criteria for evaluating atmospheric models were not also scale-dependent. It seems inevitable therefore that the utility scale will be relative to the temporal and spatial scales at which the model is applied. Both Hogrefe et al. (2001) and Beven (2002) provide strong arguments that this should be the case.

Furthermore, keeping in mind the two major uses of a regional AQM, it seems reasonable that evaluation requirements for diagnostic uses of an AQM system will be different from those of a forecast model. This should be so since the diagnostic application of models requires that the interrelationships between processes and their attendant parameters in the model match those discerned from observations. By contrast, a forecast model will be judged to have utility if the temporal evolution of chosen variables (the forecast variables) corresponds to those that actually

occur. While in both cases we want the model to accurately predict the answers for the right reasons, the evaluation metrics employed and their interpretation depend upon the application endpoints. We argue therefore that model evaluation criteria should be dependent on the context in which they are to be applied.

What then are the overall primary objectives of AQM evaluation? There seem to be three such objectives:

(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 (Britter et al., 1995). The last phrase in the previous sentence is important because there is always a reason why we are running the model, whether it is run in operational mode, forecast mode, or research mode. The reason should be precisely stated as well as the model outputs that are being evaluated (e.g., the daily maximum 8-hr average ozone concentration at a routine monitoring site anywhere on the given domain for a particular time period). In the case of research activities associated with model improvements, the model outputs may be very specific but need to be precisely stated in any event (e.g. the daily average sulfate concentration during the Texas 2000 experiment over the entire given model domain in the layer from 100 to 1000 m). In the case of NOAA-EPA's ozone forecasting guidance, the output of interest may be the number of routine observing sites in the eastern U.S. where the 8-hr ozone standard is exceeded on a given day.

Two types of model application are for air quality management and for short-term air quality forecasting. In the former, we are mainly interested in the model's ability to correctly estimate the air quality response to potential source-term emissions reductions. In this application, diagnostic assessments of the model's individual and interactive processes are desired since we are focused on the model response to a change in one of the driving parameters. Evaluation of the outcome state of the model is a necessary, but not sufficient, step in this evaluation. The emphasis in air quality forecasting, by contrast, is chiefly on the outcome state of the model, a prediction of next-day (or short-term) air quality. Criteria for acceptance of model results in a particular application may be established a priori, or may be fluid depending upon the needs and requirements of the application.

(2) Distinguishing the performance among different models or different versions of the same model.

We are sometimes in a position of determining whether the performance of one model is significantly different from that of another model for an intended application when multiple models have been employed. Even more often, we are faced with the question of whether to move an application to the latest version of a model when we have already been using the previous version of the model for other applications. Evaluation procedures must to be able to distinguish the performance in outcome states among models as compared to observations, with specified levels of significance. Some of the existing AQM evaluation methods (e.g., Chang and Hanna, 2004; ASTM, 2005; Irwin et al., 2008) include the ability to determine whether the performance measures of two models are significantly different from each other, but there is uncertainty about the degrees of freedom to be chosen. Evaluation procedures must also be

available to distinguish among models with respect to their process-level scientific credibility, even when outcome state performance is comparable.

(3) Guiding model improvement. Ultimately, model development and refinement are dependent upon model evaluation to guide and inform the process. Evaluation exercises shed light on the systematic biases and errors in model outcome states as well as indicating sensitivities and uncertainties in the atmospheric processes simulated within the model. The results of these exercises should lead to new directions in model development and improvement, as well as sometimes pointing to the need for additional measurements for better diagnostics in evaluation exercises.

Given these philosophies and objectives of model evaluation, a conceptual framework is presented here to guide model evaluation exercises. Figure 1 presents a model evaluation framework, which is based on the purpose and specific questions being asked as part of the evaluation. As a first step in model evaluation, model predictions are compared to observed data and some statistical measures are computed to gauge model performance in an overall sense, which is referred to here as "**operational evaluation**." However, the ability of a model to predict the outcome state pollutant of interest does not address whether the predicted concentrations stem from correct or incorrect physical/chemical modelled processes, which should be addressed via "**diagnostic evaluation"**. For secondary pollutant species that are not directly emitted, diagnostic evaluation methods are critical for insuring credibility of the model and for identifying potential model improvements. Figure 1 also includes an evaluation approach referred to as "**dynamic evaluation**" that focuses on the model's ability to predict

changes in air quality concentrations in response to changes in either source emissions or meteorological conditions. This exercise requires historical case studies where known emission changes or meteorological changes occurred that could be confidently estimated. Dynamic evaluation also requires that these changes have a discernable impact on air quality. Operational, diagnostic, and dynamic evaluation approaches complement one another by not only characterizing how well the model captured the air quality levels at that time, but how well the model captures the role and contributions of individual inputs and processes and the ability of the AQM to respond properly to changes in these factors. These three approaches in concert provide a comprehensive evaluation of model performance for specific model applications and support the priority directions for further model improvement.

A fourth aspect of model evaluation in Figure 1, referred to as **"probabilistic evaluation"**, attempts to capture the uncertainty or level of confidence in model results for air quality management or forecasting applications. To better determine the significance of the model performance, it is necessary to know the uncertainty in the model predictions and in the observations. Many methods exist to estimate the uncertainty (e.g., ensemble runs, direct calculation of variances in predicted concentrations, Monte Carlo runs, analytical error propagation methods for simple-model algorithms). A classic example would be ensemble modeling being used for meteorological forecasting. Currently in operational use in several national weather services globally, this technique makes use of multiple model runs of different models or the same model with different parameter or process choices. Results from the multiple model runs allow the forecaster to describe local and regional forecasts in terms of probability of occurrence. With computer efficiencies improving exponentially, methods such as ensemble

modeling that characterize a range of uncertainties in an AQM context, become increasingly realistic for decision-making or forecasting. Probabilistic model evaluation has not been extensively used in regional three-dimensional photochemical AQM applications, although such methods have been used in plume dispersion modeling for many years (Lewellen et al., 1985). Hanna and Davis (2002) showed how the methods could be used to evaluate ozone predictions by a regional AQM. Additional research and advancements are needed to develop innovative approaches that consider the confidence in AQM predictions for various applications (see Gégo et al., 2003).

3. Evaluation Methods

In this section, we further describe the elements and approaches of the model evaluation framework, providing illustrative examples of their application to regional AQM systems. Typically, such systems incorporate models for meteorological characterization or forecasting, source emissions estimation, as well as the air quality (i.e., chemical-transport) model itself. Where applicable we show the evaluation of each of these components of the AQM system.

Operational Evaluation. Operational evaluations focus on the direct comparison of model outputs with analogous observations. An operational evaluation makes use of routine observations of ambient pollutant concentrations, emissions, meteorology, and other relevant variables. Fortunately, there has been a steady increase in the number of routine pollutant concentration and deposition monitoring stations in the U.S. and in the number of chemical constituents that are measured. Emissions estimations have also been steadily improving

although there are still few direct continuous measurements of emissions, with the principal exception being Continuous Emissions Monitoring Systems (CEMS) on large U.S. electrical generating units (http://www.epa.gov/ttnnaaqs/ozone/areas/etscem.htm). Routine meteorological observations have primarily been enhanced with improved surface meso-networks. However, for all variables, there is still a dearth of routinely measured vertical profiles in the lower troposphere. Remote sounders are improving, but few are in place as part of the routine network. 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 model variables include ozone (O_3), carbon monoxide (CO), nitrogen oxides (NO, NO_x), and fine particulate matter mass and species ($PM_{2.5}$, SO_4, NO_3, NH_3, OC, EC). It is recommended that the evaluation of meteorological variables be coordinated with air quality to determine how errors in the meteorology model affect AQM performance (Seaman, 2000; Hanna and Yang, 2001).

Once the evaluation goals and model outputs of concern are precisely identified, the next task in the model evaluation is to "look at" the data (Tukey, 1977). This should be done prior to applying statistical software. For example, the time series of observed and predicted concentrations at a location could be plotted and studied. Simulated concentration contours can be compared with observed patterns. This allows the scientist to determine if there are any obvious biases or problems that the eye can see. In this manner, outliers can quickly be seen.

The next step is to clearly identify the quantitative performance measures that are to be considered in the evaluation and the criteria for deciding whether the model is performing adequately for that situation (Cullen and Frey, 1999). It is usually desirable to calculate quantitative statistical performance measures that depend on the model type and the output characteristics, as well as the available observations. These performance measures can be compared with those calculated for previous related evaluations, or the performance measures for two or more different models or different versions of the same model can be compared (Irwin et al., 2008). There are three performance measures that are commonly used in AQM evaluation (and most other types of model evaluation) – the mean bias, the root mean square error, and the correlation. There are various ways of defining these, using different normalizing factors, but the basic measures are similar (Weil et al., 1992, Hanna et al., 1993). In any case, it is important to calculate the statistical confidence levels. Standard statistical tests 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?" To use these tests, there are assumptions such as the hourly concentrations being independent, and these are not always satisfied. Appendix A provides several standard statistical metrics commonly used in AQM evaluation.

Limitations of Standard Metrics The standard metrics included in Appendix A do not take into consideration that the three-dimensional regional AQM model predictions, M_i, are volume-averaged concentrations representing ensemble mean conditions, whereas the observations, O_i, are point measurements reflecting individual events. This phenomenon is known as an *incommensurability* or *change of support* problem. Because of the spatial averaging (smoothing)

inherent in grid cell predictions, it is fundamentally impossible to use the model output to determine values that are precisely comparable to point observations. One way to account for the effects of small-scale spatial variability is to use a spatial smoother such as the block-kriging technique (Cressie, 1993) on the observed data to produce data-based grid cell values. The comparisons of model and data-based grid-cell values would then be conducted using the standard metrics discussed in Appendix A. Graphical displays such as time series, frequency distributions, and concentration isopleths could also be used with modeled values and the data-based grid cell values. However, it should be noted that the smoothing technique is a model itself, 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.

Many of the standard metrics assume that modeled and observed values conform to the same distribution (e.g., normal). While the observations tend to be log-normal, the predictions from three-dimensional AQMs appear more normally distributed (averages are normally distributed according to the Central Limit Theorem.) This issue of using the standard metrics in operational model evaluations for comparing data sets characterized by different distributions is one that is often overlooked.

It is important to conduct comparisons of the spatio-temporal patterns of the model predictions and the observations. This can be done by simply determining the fractional overlap of spatial patterns or time series of predictions and observations (e.g., Chang and Hanna 2004). Some methods allow the evaluator to weigh underpredictions (false negatives) more than overpredictions (false positives) (Warner et al., 2004). Or, in some cases, the evaluation could

extend to determining whether the scales of variability in the predicted and observed patterns are comparable, using, for instance, correlation and spectral analysis. Differences between maps of model predictions and maps computed from data-based grid cell estimates yield a spatial difference field. It is important to study the spatial pattern of these differences (Gégo et al., 2007). Investigation of spatial patterns can be done using statistical measures of spatial dependency, such as the *variogram* function (Cressie, 1993), and temporal dependency structure can be studied with methods such as spectral analysis. For example, Rao et al. (1997) and Hogrefe et al. (2000) have decomposed the time series of O_3 into spectral bands representing intra-day, diurnal, synoptic, seasonal, and longer-term fluctuations. Figure 2 illustrates the comparison between these component spectra estimated from 15 years of observed and CMAQ model-predicted hourly ozone 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 inaccuracies in the regional model's boundary conditions and representation of the free tropospheric processes. Empirical Orthogonal functions (EOFs) can also be used for analysis of spatial/temporal data (Jolliffe, 2002). This approach provides a decomposition of the spatial response surfaces in terms of "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. If possible, it is useful to aggregate the results across coherent space and/or time regions so as to represent distributional quantities, and not single point observations. For instance, O_3 concentration distributions over all monitoring sites in a region can be plotted as a daily time series over a month 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 subregion) can be used to plot 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 (Appel et al., 2008) are useful for comparing simulated and observed chemical constituents of size-segregated particulate matter (PM). Morris et al. (2005) illustrate the use of 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. Vautard et al. (2007) make use of the Taylor diagram (Taylor, 2001) which combines model error and correlation in a single point, and is particularly useful for comparing the performance of several models. Examples of the soccer and Taylor plot graphical techniques are illustrated in Figure 3. In Figure 3(b), the Taylor plot, each symbol in the plot represents a distinct model. The plot shows, for each model, the standard deviation of simulated values (radius) and the time correlation between simulated and observed values (angle from horizontal). 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 extent and magnitude of the modeled concentration field through a concentration isopleth or colored grid plot overlain 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 extent 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 (Stohl et al., 1998).

Before leaving the topic of operational evaluation, the issue of evaluation of source emissions inventories and emissions models should be acknowledged. Emission models are part of regional AQM systems. However, unlike the deterministic meteorology and AQMs whose results are based on time-dependent differential equations, emissions models are typically based on engineering and empirical approaches using surrogate or indirect measures of real emissions fluxes. Generally, emissions model results are not directly verifiable since emission observations do not exist on the regional-scale. (The sole exception to this general case is the CEMS, mentioned earlier, 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 (see Cullen and Frey, 1999) 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. While the above metrics and comparisons could be applied to any chemical trace constituent for which observations are available, the comparisons are traditionally focused on the air pollutant endpoint of interest (e.g., ozone, total and individual aerosol species that comprise $PM_{2.5}$). While these operational evaluations are important to establish model performance for the pollutants of concern, the comparisons do not identify whether the modeled concentrations are correct for the right physical/chemical reasons, what inputs or processes have a strong influence on model performance, and whether the model is capturing these factors well. Evaluation approaches that look into these types of questions in light of model evaluation are traditionally referred to as diagnostic evaluation methods and cover a wide variety of evaluation studies that consider the physical, chemical, meteorological, and/or emissions processes in an AQM system. In simple terms, diagnostic evaluation must consider how the pollutant's mass budget is impacted by the chemical and physical processes and the sources and sinks. Diagnostic

evaluation is a critical piece that brings feedback from model evaluation studies back to continued model improvement.

Regional AQM diagnostic evaluations are complicated by the fact that the system is non-linear. That is, a change in a given model input does not always lead to the same change in a model output. In some cases, even the sign of the change in the output will switch as the magnitude of the input changes (e.g., the effect of changes in NO_x emissions on ozone concentrations).

In order to proceed into diagnostic evaluation from an operational evaluation of the endpoint pollutants, observational data beyond traditional networks that track air quality attainment thresholds are often needed. To focus on the chemical processes, precursor concentrations at relatively high temporal resolution (e.g., ten-minute averages) are needed. For example, hourly data such as speciated volatile organic compounds and NO_y have been collected at field studies focused on ozone chemistry, along with radiation data and photolysis rate estimates. Diagnostic evaluation of aerosol chemistry also requires extensive data for the individual aerosol species, their size distributions, and their chemical precursors. Having measurements of both the chemical endpoint(s) of interest and the precursors is also helpful in constraining the emission or source/sink budgets, which will be discussed further in a later section. In addition to the large data demands for the chemical species themselves, the direct and indirect roles of the meteorology on the chemical concentrations can require data on meteorological parameters that are not typically available, such as the planetary boundary layer heights and clouds, both of which have a large impact on air quality concentration levels.

These types of process-oriented field studies can provide a rich data set, 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 (Hudman et al., 2007; Brown et al., 2006). 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 special studies.

Diagnostic Evaluation: Separating roles of model inputs from modeled processes

Rather than presenting a summary of diagnostic evaluation studies, a perspective on the challenges and approaches of diagnostic evaluation will be discussed with the motivation of identifying where model improvements are needed. If operational evaluation results show poor model performance for an air pollutant of concern, a number of factors could be driving the model performance. (The term poor performance is used here generally. It could imply poor performance as compared to other AQMs, previous versions of the same model, different emission or meteorological inputs, etc.) Even if the evaluation shows good performance, we need diagnostic evaluation to help build confidence in model predictions. One of the first challenges as an operational study transitions into diagnosing the cause(s) of model errors is to attempt to determine if the model performance issue is driven by inputs, such as meteorology or

emissions, or by chemical/physical processes simulated within the AQM. This is fundamental to determining what improvements are needed and what uncertainties are involved.

Sensitivity tests are one of the most common and traditional 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. (2004), and Cullen and Frey (1999) provide specific discussions related to AQMs. However, as mentioned above, because of the nonlinear characteristic of regional AQMs, the sensitivity test may only be valid for a certain range of input variables. As an example of a regional AQM sensitivity study, air quality simulations can be performed using multiple meteorological inputs to assess how much meteorological model errors and differences impact the air pollutant (e.g., Otte, 2008). Emissions have also been varied either through incremental changes to emission inputs or comparison across different inventory estimates (e.g. Gilliland et al., 2008, Pinder et al., 2006, etc.) to test the impact on air quality endpoints. Advanced instrumented modeling tools 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 (e.g., Henderson, 2008; Vizuete et al., 2008a, 2008b). Figure 4 is an example result from Henderson (2008) where contributions from production, loss, emissions, and transport were tracked for two model

simulations with different grid cell sizes. 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, boundary conditions, and more is calculated during the model simulations (e.g., Cohan et al., 2005; Napelenok et al., 2006). The DDM tools are able to capture both the first and second order sensitivities to these inputs, which is important for these non-linear chemical systems.

A key purpose of these sensitivity tests and instrumented modeling tools is to identify whether the inputs (e.g., emissions, transport, boundary conditions) have a large enough influence that known errors in the inputs could be a driving influence on the air quality predictions. If not, a more internal focus on the AQM itself should be the priority; however, emissions and/or meteorology are often found to have a dominant influence on air quality predictions and warrant improvement. 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 (2000) provided a comprehensive summary of the key meteorological issues most relevant for air quality modeling. Hanna and Yang (2001) evaluated the boundary layer outputs of several

mesoscale meteorological models (e.g., MM5, RAMS, OMEGA), stressing meteorological variables used by AQMs. For example, it was found that the rmse of wind speed predictions was about 1 or 2 m/s at best, and that the models generally underestimated the strength of nocturnal inversions. 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 (2008) 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.

Another example is the growth and evolution of the planetary boundary layer (PBL) and entrainment of pollutants trapped aloft in the nighttime residual layer as the PBL grows during the day. These entrained pollutants can be transported long distances downwind by the nocturnal jets and can have a distinct impact on near-surface concentrations of air pollutants the next day (Zhang and Rao, 1999). While observations of the evolution of the PBL are very limited, recent measurement studies have demonstrated new approaches that can provide these critical data (e.g., Emeis et al., 2004) on PBL heights and growth. With this information, the importance of PBL heights for hourly AQM predictions can be investigated more thoroughly. Without these data, inferences about hourly air quality predictions and how they relate to PBL evolution (e.g., rate of rise of the PBL in the morning, decay of the PBL in the afternoon) could be incorrect. The time evolution of the mixing height growth in the morning and the PBL decay in the afternoon is a significant determinant of near-surface pollutant concentrations. The interactions of pollutant emissions at key times, such as rush hour traffic emissions, into reduced mixing layers can have major impacts not only on the direct primary emitted pollutant concentrations

(e.g., CO, NO_x, VOCs, NH_3, PM) but also on the secondary pollutants produced from chemical reactions among the primary pollutants (e.g., O_3, secondary organic aerosol, nitrate aerosol).

Regarding the role of emissions in air quality predictions, "top-down" diagnostic evaluation studies consider predicted concentrations and infer what the emissions should have been. Inverse modeling, indicator ratios, and source apportionment are all examples of top-down emission evaluation methods that use information outside of the emission inventories to evaluate and inform current emission estimates (Parrish, 2005). Inverse modeling (e.g., Gilliland et al., 2003, 2006; Napelenok et al., 2008) uses methods to estimate what emissions would result in the minimum least squares error in the resulting concentration indicator. Indicator ratios may be used to estimate how much of an emitted pollutant comes from one source type versus others (e.g., NO_x/CO ratios in Parrish et al., 2002). Receptor models estimate, based on pre-determined chemical speciation profiles for different sources and observed concentrations at the receptor, the relative contribution of various source types (e.g., Wittig and Allen, 2008). All of these types of analyses can provide important insights about the magnitude, location, and/or sector(s) contributing to emissions and can be complementary to one another. In all cases, the methods rely on observational data to constrain the conclusions, and models or assumptions are introduced as part of the emission estimation.

Since space limitations do not allow for a detailed discussion of these top-down approaches to evaluate emissions, the emphasis here is on the value and requirements of top-down approaches to evaluate emissions. 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 (e.g., Bhave et al., 2007) and used detailed tracking of emission contributions across space for inverse modeling (e.g., Napelenok et al., 2008). In all cases, top-down methodologies can inform continued improvement to the bottom-up inventories that are critical for AQM performance.

These types of diagnostic studies often demonstrate the important influence that meteorological predictions and emission inventories can have on air quality predictions; however, improvements can also be found in the chemical or physical processes within the AQM. Evaluation of the chemistry in AQMs has benefited greatly from observed measurements of O_3 precursors and indicator ratios. For example, a number of different indicators have been used to evaluate the oxidant chemistry as simulated in AQMs. Probing indicators related to the total oxidized nitrogen include the ratio of total nitrogen species to NO_y (Parrish et al., 1993;

Trainer et al., 1993), the photochemical age ratio ((NOy-NOx)/NOy), and the (O_3 versus (NO_y-NO_x) (i.e., NOz)) ratio (Arnold et al., 2003). These indicators all can be used with available observations to assess whether the oxidant chemistry as simulated in the model is similar to the observed chemical oxidant state.

A key application of AQMs is to estimate O_3 and/or $PM_{2.5}$ changes in response to emission changes. Many traditional and more recent diagnostic probes focus specifically on the potential response of model predictions to emission changes. For example, the indicator ratios of H_2O_2/HNO_3 (Sillman, 1995; Sillman et al., 1998; Kleinman, 1994; Kleinman et al., 1997), and O_3/NO_x (Tonnesen and Dennis, 2000a,b; Arnold et al., 2003) are both response-surface probes that have been used to characterize how O_3 will change with NO_x and VOC levels in a given area. More recently, the potential for nitrate replacement and less reduction in total $PM_{2.5}$ than anticipated with SO_2 emission reductions has been studied using the Gas Ratio, which is a ratio of free ammonia to total nitrate (Ansari and Pandis, 1998; Pinder et al., 2008; Dennis et al., 2008). By comparing modeled results to observations from special field studies, these types of diagnostic probes help to extend diagnostic evaluation from assessment of predicted concentrations to evaluation of the model's ability to respond correctly to emission changes.

Dynamic Evaluation A new area of model evaluation referred to as "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 model predictions where mobile source emissions are known to significantly change (Chow, 2003). 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 (Gilliland et al., 2008). Figure 5 illustrates principal findings from that study. 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. Additional 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 observed pollutant concentrations for meteorological influences (e.g., Porter et al., 2001; Camalier et al., 2007).

Probabilistic Evaluation All regional numerical AQM systems use first-order closure. That is, the variables that are being solved are the ensemble means. It is of course possible to write the model system and solve the equations using second-order or higher closure. For example, the SCIPUFF model (Sykes et al, 2007) uses second-order 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 run times 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. (1985) that depends on knowledge of the probability distribution function (pdf) of the AQM predictions. As stated above, their AQM, named Second Order Closure Integrated Puff (SCIPUFF; Sykes et al., 1984, 1988, 2007), assumes a pdf shape (the clipped normal) and automatically predicts both the mean and the variance of the concentration distribution. The Lewellen et al. (1985) probabilistic model evaluation methodology was applied by Hanna and Davis (2002) 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. (1987) used the Monte Carlo approach to propagate meteorological input uncertainty, using pdf's, to air quality predictions.

As described by Lewellen et al. (1985), in order to carry out a probabilistic AQM evaluation, it is necessary to somehow "know" the pdf of the model outputs. The pdf can be estimated by many methods. For example, the simplest way is to assume the same pdf everywhere at all times – such as an exponential distribution shape with the standard deviation equal to the mean, or a clipped normal distribution shape with an assumed standard deviation and mean (can be prescribed independently). Or, the shape can be prescribed and then the model calculates the variance along with the mean (as done in SCIPUFF). Another method, as described by Hanna and Davis (2002) involves a long Monte Carlo exercise with over 100 model runs. Yet another technique uses an ensemble of modeling methods to define the pdf. 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 SCIPUFF was tested by Warner et al. (2002), who showed that the method was able to adequately account for the uncertainties in the concentration pdf due to mesoscale and regional meteorological variations. It should also be mentioned that earlier works started down this path in the early 1980s (e.g., Lamb and Hati, 1987; Schere and Coats, 1992). Lamb and Hati proposed using several (an ensemble, although they did not use that word) possible mesoscale regional wind fields to drive a regional ozone model, and therefore produce the desired estimate of uncertainty.

The ensemble method is a quite simple "brute force" approach where a number, N, of model runs are made for the same scenario. Widely used for numerical weather prediction, applications for air quality modeling are only recently being reported (Galmarini et al., 2004a,b; McKeen et al., 2005; Mallet and Sportisse, 2006a,b; DelleMonache et al., 2006a,b; Zhang et al., 2007). The concept behind ensemble modeling is that uncertainties in model inputs and model formulations cause uncertainties in the predicted pollutant concentration fields. The model runs may be made with different models, different algorithms in the same model, or different boundary and initial conditions, to name a few. The ensemble method is based on similar principles as the Monte Carlo model uncertainty method, which is widely used for other environmental and risk analysis models. The main difference is the ensemble method uses far fewer model runs, carefully chosen so that the full uncertainty range is retained. By performing multiple simulations with a set of different model inputs and model parameterizations, the impact of their uncertainties on the resulting model predictions can be quantified by providing a range of model predictions. Computational resource limitations typically constrain the number of discrete simulations populating an ensemble. Ensemble members must be carefully chosen such that each member of the ensemble, among other criteria, must be shown to "perform well". Thus, the method is empirical with some arbitrariness.

In weather forecasting, ensembles of meteorological model simulations are routinely used operationally for mesoscale weather predictions, including for predicting the track and strength of hurricanes. Ensembles are often constructed using different large-scale fields for initialization as well as selecting different cumulus and boundary layer parameterizations within the

meteorological model (e.g. Jones et al., 2007), or using different meteorological models (Biswas and Rao, 2001). For air quality modeling applications, sensitive input fields and model parameters in addition to the meteorological fields include anthropogenic and biogenic emissions, photolysis rates, and chemical mechanism (e.g. Hanna et al., 2001; Della Monache et al., 2006a,b; Carvalho et al., 2007). The most sensitive parameters can be determined through sensitivity runs (use of different physics options, chemical mechanisms, grid resolution, etc.) or in a more formal way through adjoint modeling studies (Menut, 2003; Menut et al., 2000). Realistic distributions for these sensitive parameters can then be obtained through literature review and expert solicitations (e.g. Hanna et al., 2001; Fine et al., 2003). Pinder et al. (2008) describe a technique for generating ensemble members based on discrete model simulations of a single AQM system, in combination with a direct sensitivity technique, that can efficiently produce hundreds to thousands of ensemble members. Figure 6 illustrates a month-long time series of daily 8-hour 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.

The ranges of model predictions stemming from ensembles can be utilized in a number of ways. For example, in forecasting applications for discrete events such as the occurrence of rainfall or the exceedance of a threshold concentration for a certain pollutant, one can estimate

the probability of that event occurring by determining in how many ensemble members the event occurs or does not occur. From a model evaluation perspective, if an observed value falls within the range spanned by the ensemble predictions, this can be interpreted as there being no discernable difference between the observed quantity and model predictions given the underlying uncertainty of model predictions. In this context, the comparison of observations and ensemble predictions is also important to evaluate and refine the design of the ensemble experiment itself. In order for ensembles to provide useful information, it is necessary to determine whether the predicted spread is a true measure of underlying model uncertainty. This can be assessed for example through Talagrand diagrams (Delle Monache et al., 2006b; Vautard et al., 2006) constructed from raw or bias-adjusted ensemble members and observations. Deviations from the flat shape of an ideal Talagrand diagram can indicate whether the spread of the ensemble is too small because the observed event often falls outside the range of values sampled by the ensemble or whether the ensemble forecasts are systematically biased toward overprediction or underprediction.

It is important to note the distinction between a true measure of model uncertainty and the results obtained from a finite set of model simulations to create an ensemble. The former describes the full spectrum of the population of results constituting model uncertainty based upon data and model formulation/parameterization uncertainties. The latter is a limited view of a portion of the uncertainty spectrum; in a sense it is a measure of the "spread of our ignorance". Techniques to formally propagate uncertainty through AQM systems are not commonly used owing to the non-linearities inherent in AQM formulations and the difficulty of uncertainty propagation through non-linear systems that also contain parametric algorithms. Some studies

have relied on Monte Carlo techniques (Hanna et al., 1998; Hanna and Davis, 2002) to estimate model uncertainties. The Monte Carlo method is based on probabilities and nonparametric statistical methods amenable to standard statistical analysis. However, the variables in AQM systems (including meteorological, emissions, and air chemistry) are not all independent from each other, making Monte Carlo techniques problematic to apply and interpret results from such exercises. It is possible to account for correlations among input variables in the Monte Carlo methods, but the magnitudes of these correlations are not well known, thus limiting the usefulness of Monte Carlo techniques for quantifying regional AQM uncertainty. As numerical air quality simulation models are used as surrogates for the real chemical atmosphere, they defy attempts to estimate "true" uncertainty since the grid-based formulation resolves only a limited set of spatial/temporal scales. The true stochasticity of the atmosphere is not captured. For example, if the same conditions were used to repeatedly run a model simulation, the same results would be obtained with each run of the model. However, in the real atmosphere, the "same" conditions existent on different days would not result in the same weather or air quality conditions given the stochastic variations across the spectrum of scales intertwined in nature.

Information from ensembles can also be used to estimate distributions of model-observation differences that can guide the interpretation of modeled concentration endpoints (e.g. predicted daily maximum 1-hr ozone concentrations) in particular applications. For example, Biswas and Rao (2001) and Rao et al. (2001) used a number of different meteorological and photochemical modeling configurations to simulate daily maximum 1-hr ozone over the eastern U.S. during multiple high-ozone episodes during the summer of 1995. Their results showed that the choice of modeling options typically introduces a variability of 20% of simulated individual daily

maximum 1-hr ozone concentrations, suggesting that this value could be viewed as an empirical lower bound on the model's uncertainty in simulating this quantity. Another potential use of this information is in designing statistical tests that can then be used to address the question of whether a difference between observed and simulated concentration is significant at a certain confidence level given the estimated error distribution.

For many air quality planning applications, the quantity of greatest interest is the modeling system's relative response to emission reductions (U.S. EPA, 2007). Therefore, to estimate the effect of model uncertainties on this quantity, ensembles can be constructed in which this response is calculated using a variety of meteorological and photochemical models and modeling options. To this end, two simulations have to be carried out by each ensemble member; one reflecting the base case emission scenario and the other reflecting the control case emission scenario. In a series of studies, Hogrefe and Rao (2001), Sistla et al. (2004), Jones et al. (2005), and Hogrefe et al. (2008) 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.

Another potential approach to the probabilistic evaluation of AQMs is the use of rank 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 4^{th}-highest daily maximum ozone concentration over a summer season. In addition to directly comparing the

observed and simulated 4th-highest concentrations, one can utilize statistical 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 in Hogrefe and Rao (2001).

There are challenges in using observational data for probabilistic model evaluations. The air quality observed on a given day is a sample or individual event from a larger population. For example, the air quality on July 1 this year might be a member of an ensemble consisting of the set of air quality observations on July 1 from Y_{i-5} to Y_{i+5}, where Y_i is the current year. Since it may not be practical to model the air quality on July 1 over this temporal range, another possibility is to seek multiple days over a shorter time window when the meteorological (and emissions) conditions are "similar" to July 1. The distribution of modeled results for these days and the distributions of the observations can then be compared.

Fuentes and Raftery (2005) developed a new Bayesian approach to evaluate the spatial pattern of concentrations simulated by AQMs, and showed how it can also be used to remove the spatial bias in model output. The Bayesian approach is ideal for this application because it provides a natural framework to compare data from very different sources taking into account

different uncertainties, and it also provides posterior distributions of quantities of interest that can be used for scientific inference. They do not treat monitoring data as the "ground truth". Instead, they assume that there is some smooth underlying (but unobserved) field that measures the "true" concentration/flux of the pollutant at each location. Monitoring data are these "true values" plus some measurement and representativeness errors. The AQM output can also be written in terms of this true underlying (unobservable) process, with some parameters that explain the bias and microscale error components in the model. The *truth* is assumed to be a smooth underlying spatial process with some parameters that explain the large-scale and short-scale dependency structure of the air pollutants. They evaluate the model by comparing the distribution of the monitoring data at a given location, to the predictive posterior distribution of the model at that given point in space. The bias in the AQM result is removed by obtaining the posterior distribution of the bias parameters given monitoring data and model output. This technique does not account for the temporal dependence in the data; more recent research efforts in adapting Bayesian approaches are focusing on space-time data fusion of model results and observations (Gégo et al., 2007).

In an example illustrating the spatial Bayesian approach, observed values of SO_2 weekly averaged (week of July 11, 1995) concentrations at six selected CASTNet sites that are representative of different meteorological, land use, and altitude conditions are used. Figure 7 shows the predictive posterior distribution (ppd) of the CMAQ model interpolated output at the six CASTNet locations for SO_2 (ppb). The circle in each graph indicates the CASTNet value at the given site. As expected, large uncertainty at the Indiana site is obtained. This site is very close to several coal-fired power plants, and so the SO_2 levels can be very high or very low

depending on wind speed, wind direction, and on the atmospheric stability, so there is more variability. Taking into account this variability and other sources of uncertainty as characterized by the corresponding ppds, the CASTNet value at this site is not significantly different than the CMAQ output. The sites in Maine and Florida have the lowest SO_2 levels and variability. The agricultural site in Illinois and the site in North Carolina have similar behavior in terms of SO_2 levels. The site in North Carolina is not far from the Tennessee power plants, and the site in Illinois is also relatively close to some Midwestern power plants. The site in Michigan, which is very close to Lake Michigan and relatively far from power plants, also has low SO_2 levels. By using a Bayesian approach for model evaluation, we can characterize different sources of uncertainty in CMAQ model estimates and CASTNet observations that should lead to a more reliable and accurate evaluation of the CMAQ model.

4. Data Needs for Model Evaluation

Model evaluations often rely on observational data sets that are not designed to support modeling assessments. Consequently, there are numerous incommensurabilities between model evaluation needs and observations. However, there does exist an enormous body of routine air quality observations (Figure 8), which combined with periodic intensive field campaigns and satellite missions, provides compositional, temporal and vertical data often mined for operational and diagnostic model evaluations. This section examines the data needs for model evaluation with regards to the available observational datasets.

Meteorological Data. While there is a wealth of surface-based observations, vertical profiles of temperature, winds and planetary boundary layer (PBL) heights are most relevant to diagnosing physical processes in large Eulerian modeling systems. Constructing the spatial distribution and temporal evolution of the PBL is a fundamental need for model evaluation. PBL height is a derived quantity based largely on vertical temperature profiles and refractive index structure parameters, Cn^2. The National Weather Service's (NWS) radiosonde network's twice daily soundings at nearly 100 locations across the U.S. lack adequate temporal resolution to characterize the diurnal development and decay of the PBL. Radar profilers are capable of providing the necessary temporal resolution. NOAA has deployed 35 unmanned Doppler Radar sites (NPN - http://www.profiler.noaa.gov/npn/) profiling the troposphere (10-15 km). The NPN, concentrated in the central United States, is designed for violent weather forecasting. While there is lack of a consensus methodology to synthesize raw radar profiler data into temporal observation patterns for model evaluation, the main issue is deployment to provide adequate spatial coverage. The Photochemical Assessment Measurement Stations (PAMS) program and a variety of State funded programs support over 30 boundary layer radar profilers (http://www.madis-fsl.org/cap/profiler.jsp) that provide highly resolved wind profiles and Cn^2 coefficients of the boundary layer (up to 5 km). The boundary layer radar profilers, especially when complemented by temperature profiles generated by a Radio-Acoustic Sounding System (RASS), offer a source of relatively untapped data for model evaluation which would benefit from an organized data synthesis effort for selected model evaluation applications.

Emissions Data. Emission inventory evaluation should address improving emissions of relatively inert pollutants relating to physical processes (e.g., elemental carbon (EC) and CO) as

well as precursor emissions driving chemical transformation of ozone and aerosols (e.g., nitrogen oxides, NO_X, ammonia, NH_3, sulfur dioxide, SO_2, and volatile organic compounds, VOC). Accuracy for these gases is required, given the multiple interactions across pollutant species. From an evaluation perspective, natural emissions should be considered as important as anthropogenic sources as all emission fields affect the modeling system's predicted concentrations. These needs are consistent with the recommendations in the recent NARSTO emission inventory assessment (NARSTO, 2006) but with less emphasis on toxic and hazardous air pollutants. Direct measurement of emissions is ideal, such as CEMs on major point sources (e.g., power plants). Otherwise, engineering and emission model estimates need to be independently evaluated. Evaluation through measurement programs can be broken into four categories: (1) direct near-source measurements; examples include ammonia flux measurements (Aneja et al., 2008) and remote sensing of roadway emissions through open path FTIR methods (Bradley et al., 2000), (2) dedicated mass balance studies in controlled environments such as roadway tunnels (e.g., Gertler et al., 1997), (3) inference analysis based on ambient measurements, typically in the form of statistical relationships through source apportionment models or simple ratio analysis (e.g., Parrish et al., 2002) and (4) inverse modeling using ambient measurements at the surface or from satellites with assumed well characterized model physics (Gilliland et al., 2003; Napelenok et al., 2008). Overlaps exist across these methods as well as between source methods (ammonia flux measurements) used to develop emission factors and evaluate emissions. These approaches need to continue to be improved and expanded.

Ambient Air Quality Data.

Conservative Tracers. Very useful conservative tracers for evaluation and interpretation of physical and PBL processes are CO and EC. Total oxidized nitrogen, NO_Y, can also be useful. There are, however, overlaps between evaluating emissions and PBL processes. Each is dependent on the other and judgment plus additional evidence is required. Measurements of conservative tracers need to be hourly for best interpretive support. While a useful conservative tracer, EC is defined by measurement and analysis protocols which are not uniform across the networks and are subject to periodic modification. In addition, EC typically is a 24-hour average value. Hourly EC measurements would be very useful. The new NCore network (U.S. EPA, 2006) will provide an initial 75 stations measuring hourly trace level concentrations of CO and NO_Y.

Key Chemical Indicator Species. Fast reacting oxidizers [e.g., hydroxyl ($OH\bullet$), hydroperoxy ($HO_2\bullet$), organicperoxy ($RO_2\bullet$) and nitrate ($NO_3\bullet$) radicals] and key species associated with their production [e.g., O_3, formaldehyde (HCHO), nitrous acid (HONO), and true nitrogen dioxide (NO_2)] and their termination products [e.g. HNO_3, peroxyacetyl nitrate (PAN), plus higher PANs, and hydrogen and organic peroxide (HOOH, ROOH)] as well as photochemically-aged products [e.g., total oxidized nitrogen (NO_Y), and NO_Z (NO_Y-NO_X)] are key to diagnosing the oxidative capacity of the atmosphere, the resultant photochemical production and the sensitivity of O_3 and aerosol production to NO_X and VOC controls. Despite the importance of these indicator species for diagnostic model evaluation, the general difficulty of accurately measuring these species, due to limited atmospheric lifetimes and analytical interferences, prevents incorporation into traditional networks. For example, routine network NO_2 measurements are confounded by interferences from various oxidized nitrogen compounds. These species are

measured in intensive field campaigns, but seldom is a sufficiently complete set measured to fully support diagnostic evaluation, a long standing model evaluation need that should be addressed.

Species Involved in Physics and Chemistry. Secondarily formed species such as ozone, organic carbon, sulfate, nitrate, ammonium and mercury are of direct relevance to operational and probabilistic evaluations. Current networks provide relatively rich spatial coverage for these species, although more than one network is involved. (Since different measurement protocols may be used in different networks, care must be taken in using data across networks for model evaluation (for example, Gégo et al., 2005)) Mercury and ammonium are limited to precipitation networks and both the long time averaging periods and artifacts associated with filter-based measurements need to be taken into account. For dynamic and interpretive evaluations, complete gas plus particle budgets are needed (e.g., SO_2 plus sulfate, HNO_3 plus nitrate, NH_3 plus ammonium, and NO_Y) as well as primary aerosols and precursor organic species and speciated organic aerosols. NCore will augment current networks with CO, SO_2, NH_3 and NO_Y at 75 sites across the U.S., addressing some, but not all, of the gas plus particle budget needs. The major NO_Y species of PAN and HNO_3 will still not be measured in routine networks. Speciated organic aerosol measurements remain limited to specialized research field campaigns. Ambient speciated mercury (Hg^0, Hg^{+2} gases and particulate Hg) observations are key for mass balance interpretation and dynamic evaluations, but are only available in region-specific efforts, such as the Southeastern Aerosol Research and Characterization Study (SEARCH). Precipitation chemistry networks such as the National Atmospheric Deposition Program (NADP) and Mercury

Deposition Network (MDN) provide relatively adequate coverage of nitrogen and sulfur and improving coverage of mercury aqueous phase ions.

Vertical Profiles. Moving beyond surface observations into the mixed layer and the free troposphere is becoming increasingly important as the ozone air quality standard is lowered. Vertical information regarding O_3, NO_Y and its constituents, SO_2 and sulfate, and fine particles, is very important to this end. Vertical O_3 profiles through ozonesondes often are deployed during intensive field campaigns or for limited periods (e.g., NASA's IONS); however routine operations generally are conducted at remote locations through NOAA's global monitoring efforts (http://www.esrl.noaa.gov/gmd/ozwv/ozsondes/index.html). Aircraft-derived trace gas profiles from intensive field campaigns provide an important component for model evaluation efforts. Satellites are an emerging source of total atmospheric column information. However, there are still large uncertainties in the retrieval algorithms, which are themselves models. Continued advances are expected from studies based on the 2004 launching of NASA's Aura platform (http://aura.gsfc.nasa.gov/) with multiple trace gas capability, including glyoxal, a relatively stable VOC. The wealth of aerosol optical depth information from numerous satellite platforms has been applied to evaluating AQMs (Mathur, 2008). The proposed addition of future geosynchronous satellite missions measuring trace gases and aerosol properties focused on North America, anticipated through NASA's GEOCAPE project in the 2015 timeframe, would add a nearly continuous stream of data greatly extending the current twice-daily scans from polar orbiting satellite platforms.

Intensive Field Campaigns. Intensive field campaigns are critical tools for probing specific processes and often are the only sources of vertical profile and highly resolved chemical and temporal information. However, model evaluation often is a secondary objective of intensive studies which typically are focused on characterizing a specific process (e.g. chemical processing in clouds, missing chemical production, transoceanic fate and transport). Well-conceived field campaigns focused on model evaluation can illuminate and make quantitative the impact an improved process module can impart to a fully integrated modeling system.

Spatial and Temporal Gaps. Operational evaluations require the best spatial or horizontal-scale information feasible, including information on subgrid variability of concentrations. These measurement needs are associated with regional and urban spatial scales, with some attention to hemispheric transport scales for boundary conditions. The demand on models to address near-field characterizations to support human exposure assessments, near-roadway phenomena and urban residuals remaining after implementation of region-wide emission strategies will require attention. Diagnostic (and potentially dynamic) evaluations require hourly data for all species simultaneously. These needs are not met with current routine monitoring networks. In general, U.S. air quality networks include an abundance of surface-based measurements, many of which are collected over 24 hour (chemical speciation networks) or even weekly (CASTNET) averaging periods. Ground-based enhancements in deployment of continuous sulfate, carbon and nitrate aerosols, ammonia and nitric acid gases are obvious temporal gaps that remain.

5. Summary and Outlook

We have examined approaches to the evaluation of regional 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 reflect influences of all possible sources and scales of source variation in time and space; have measurement uncertainties; and are measurements at specific points.
- Eulerian grid modeling results have stochastic and deterministic uncertainties resulting from the emission and meteorological inputs; have deterministic uncertainties in the modeling algorithms; and provide volume-average estimates.
- It should be recognized that even with perfect model inputs and perfect model science and numerical algorithms, there will be differences between model output and observations because of fundamental differences between model output and observations.

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. The model evaluation process includes model-observation and model-model comparisons, and employs a range of standard metrics to quantify the comparison. Because the evaluation criteria appear to be different in different applications, we argue that the criteria for "success" should be context-relative as described by Steyn and Galmarini (2008).

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 are achieved in 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 point measurements) are in principle incommensurable, and that model predictions represent population averages whereas observations reflect individual events out of a population. This fundamental problem is generally ignored in the first three types of model evaluation, necessitating the need for 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 begin and end with 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 requires these 3-D datasets. 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.

Our review of recent model evaluation literature leads to two general conclusions:

- Most of the published evaluations of photochemical grid models compute and interpret model evaluation metrics without recognizing that perfect agreement is impossible.
- There are few (if any) published evaluations of inter-comparisons of photochemical grid models on a common domain that address all the premises listed above. It is unclear what the consequence would be if we were to more comprehensively address the premises listed above. Presently, we do not have sufficient information to answer this question and consequently we are not able to provide a specific set of evaluation procedures and metrics that can be generally recommended.

Our review of approaches and techniques used in the evaluation of regional AQMs indicates that there is much interest in evaluations of these models, primarily because of model application in the policy realm. It also indicates that there are many different approaches, and as yet, incomplete consensus on many central issues. We believe our categorization of model evaluation types, motives and metrics will help lead to much needed common approaches to this important problem. Despite the lack of consensus on particular approaches to regional scale model evaluation, there has emerged an agreement on the value of the process of evaluation,

including aspects of each of the four types of evaluation: operational, diagnostic, dynamic, and probabilistic. A specific model application and evaluation exercise should endeavor to employ several methods from these evaluation categories consistent with the available observational databases and the goals of the model application. Components of regional air quality modeling systems (emissions model, meteorological model, air quality model) should be evaluated as interdependent components all affecting the final air quality model estimates. A robust evaluation outcome relies on the use of several evaluation categories and methods providing different perspectives on model performance. No one method, metric, statistic, or graph is likely to provide a complete model evaluation. The model evaluation process is likely to involve a nonlinear path, with the results of a given analysis suggesting follow-on analyses that may be unique to a particular model application. Of note is the growing interest in probabilistic evaluation and the use of modeling ensembles, and the possibility of incorporating more sophisticated space-time analysis approaches than have been used in the past. Further research is encouraged in the development, application, and refinement of such probabilistic evaluation methods.

Acknowledgments

The United States Environmental Protection Agency through its Office of Research and Development partially funded and collaborated in the research described here. It has been subjected to Agency review and approved for publication.

Appendix A. Statistical Metrics Commonly Used in Regional Air Quality Model Operational Evaluations

Standard metrics Standard metrics for air quality performance evaluation focus on measures that compare the observed value and modeled quantity at a number of locations across space and/or time. Each of these metrics assumes the existence of a number *N* of pairs of modeled and observed concentrations M_i and O_i where the index *i* might vary across time at a given location, or across space for a given time, or both. Definitions and discussions of the standard metrics are contained in several widely-used papers on AQM evaluation, including Fox (1984), Venkatram (1979, 1988), Hanna (1989), Weil et al. (1992), Chang and Hanna (2004), and ASTM (2005). The metrics used for small and mesoscale models are also valid for regional models, in most cases. An important concern that is unique to 3-D (grid volume) regional models is that the model prediction represents an average over the grid volume whereas the observation is at a single point.

Metrics based on differences

This first section concerns metrics that have the dimensions of the concentrations themselves, with no normalizing by terms such as the average concentration.

The mean bias function, defined as, $B_{MB} = \frac{1}{N}\sum(M_i - O_i)$ is a measure of the average over-or-under estimation of the model for all of the data considered. Note that the mean bias can be calculated as i) the average of the differences or ii) the difference of the average modeled and observed concentration. For the mean bias metric, individual positive and negative errors tend to compensate (i.e., cancel) each other. Thus, if the model over-estimates at some locations and

under-estimates at others within a region, the overall bias averaged over the region may be relatively small, even if the magnitude of the local errors are relatively large.

The mean absolute gross error and the Root Mean Square Error (RMSE) are measures of the scatter. As mentioned above, it is possible to have an excellent (near-zero) mean bias but still have much scatter. The mean absolute gross error is defined as $E_{MAGE} = \frac{1}{N}\sum|M_i - O_i|$. It characterizes the average spread of the departure between model predictions and observations. In this metric, errors cannot compensate each other; it therefore provides an additional useful measure of the model agreement with observations. The RMSE is similar to the absolute gross error except, instead of using the absolute values of the differences between observed and modeled values, the squares of those differences are used: $E_{RMSE} = [\frac{1}{N}\sum(M_i - O_i)^2]^{\frac{1}{2}}$. This metric is more heavily influenced by a few large errors. Thus E_{MAGE} is more robust than E_{RMSE}, since large differences between observed and modeled values at some isolated pairs have less impact on E_{MAGE}.

Normalized Metrics for Mean Bias

The differences between observed and modeled values can be normalized in several ways as discussed by Weil et al. (1992), Chang and Hanna (2004) and ASTM (2005). A simple way of doing this is to normalize the concentration difference by a combination of the observed and/or predicted concentrations. For example, the mean normalized bias contains the observed concentration as the divisor, $B_{MNB} = \frac{1}{N}\sum\frac{M_i - O_i}{O_i}$. This is the mean of the relative model bias.

This metric has the undesirable characteristic that, if the values of the individual observations are very small, then very large numbers can be obtained. The possibility of very large numbers can be partially eliminated by normalizing by the mean observed concentration, yielding the normalized mean bias $B_{NMB} = \frac{\sum (M_i - O_i)}{\sum O_i}$. As for the dimensional bias described earlier, positive and negative errors tend to compensate each other in any mean bias calculation.

The measures of scatter avoid compensating errors. For example, one could use the mean normalized absolute error $E_{MNAE} = \frac{1}{N} \sum \frac{|M_i - O_i|}{O_i}$, or the normalized mean absolute error, $E_{NMAE} = \frac{\sum |M_i - O_i|}{\sum O_i}$. The latter metric is less influenced by large outliers.

With the metrics in the above paragraph, overpredictions are artificially given more weight than underpredictions, as B_{MNB} and B_{NMB} are bounded by -1 for underpredictions and E_{MAGE} and E_{NMAE} are bounded by 0 for underpredictions; thus these are *asymmetric* metrics. Note that if the normalization is by the predicted concentration, the same type of asymmetry would result.

To overcome the asymmetry of the previous normalized metrics, one could normalize by the average of the observed and predicted concentrations. This leads to the fractional bias, $B_{FB} = \frac{1}{N} \sum \frac{(M_i - O_i)}{(M_i + O_i)/2}$, taking values between −2 and +2. A more widely used metric is the normalized mean fractional bias where the overall mean of (M_i+ O_i) is used as the normalizing factor.

Many AQM evaluation experts prefer the symmetry offered by the geometric mean, MG = exp(<ln(M_i/O_i)>), where <> represents an average (see Weil et al., 1992; Chang and Hanna, 2004). With MG, a factor of 100 underprediction has symmetry with a factor of 100 overprediction. As shown by Chang and Hanna (2004), all of these metrics are strongly influenced by what is used for the minimum or threshold concentration, which is determined by the background and the instrument response uncertainty.

Normalized Metrics for Scatter

A simple and robust normalized measure of scatter is FAC2, or the fraction of predictions within a factor of two of observations. This measure is insensitive to very large outliers. Sometimes FAC5 or FAC10 are used, too.

In keeping with our earlier definitions, to avoid compensating positive and negative errors, one could use the fractional absolute error $B_{FAE} = \frac{1}{N}\sum \frac{|M_i - O_i|}{(M_i + O_i)/2}$, which takes values between 0 and 2. These two metrics are very compressed beyond ±1. Or, the Normalized Mean Square Error (NMSE) is defined as the root mean square of all of the (M_i-O_i), divided by the average of (M_i- O_i)/2.

Here too, many model evaluation experts prefer the geometric variance, VG = exp (<(ln(M_i/O_i)^2>)$^{1/2}$, as explained by Weil et al. (1992). MG concerns the mean bias and VG concerns the scatter, and both are symmetrical. A perfect model would have MG = VG = 1.0.

Factor-based metrics

Another method, using factor-based metrics, to overcome the asymmetry problem has recently been introduced by Yu et al. (2006). They suggest using a ratio between modeled and observed quantities that would be defined differently if the observed quantity exceeds the modeled than if the modeled exceeds the observed. They define $F_i = M_i / O_i$ if the modeled value is greater or equal than the observed, and $F_i = -O_i / M_i$ otherwise. The sign of the function F_i would give the sense of the departure, and its absolute value the magnitude of the departure. They calculate the B_{MB}, E_{MAGE} metrics using a function of the factor $F_i's$ rather than differences between observed and modeled values, obtaining then, the mean normalized factor bias, and the mean normalized absolute factor error:

$$B_{MNFB} = \frac{1}{N}\sum G_i \quad ,$$

$$E_{MNAFE} = \frac{1}{N}\sum |G_i|$$

where $G_i = (\frac{M_i}{O_i} - 1)$ if $M_i \geq O_i$ and $G_i = (1 - \frac{O_i}{M_i})$ otherwise. They define another factor-based metric that depends not on the relationship between M_i and O_i at each location, but on the overall magnitude of the model bias. They use $\overline{M}$ to denote the sample mean of the modeled values, and $\overline{O}$ to denote the sample mean of the observed values. Then, they calculate the normalized mean bias factor:

$$B_{NMBF} = \frac{\sum (M_i - O_i)}{\sum O_i}$$

if $\overline{M} \geq \overline{O}$, and

$$B_{NMBF} = \frac{\sum (M_i - O_i)}{\sum M_i}$$

otherwise. This is a generalization of the metric B_{NMB} but introducing a normalizing function that is different when $\overline{M} \geq \overline{O},$ than when $\overline{M} < \overline{O}$.

The normalized mean absolute error factor would be a generalization of E_{NMAE} to account for the lack of symmetry; it is defined as $E_{NMAEF} = \frac{\sum |M_i - O_i|}{\sum O_i},$ if $\overline{M} \geq \overline{O},$ and

$E_{NMAEF} = \frac{\sum |M_i - O_i|}{\sum M_i},$ otherwise.

All these factor-based metrics overcome the asymmetry problem between overestimation and underestimation.

Categorical metrics A model's ability to correctly predict the exceedance of a threshold value of a variable can be measured by a series of categorical metrics (Joliffe and Stephenson, 2003; Eder et al., 2006). The threshold is typically set at some meaningful value that is proportional to a health or welfare effect in the case of meteorological and AQMs. For example, Figure A-1 illustrates a scatter plot of modeled and observed maximum 8-hr O_3 concentrations, with the threshold of 85 ppb (level of the U.S. National Ambient Air Quality Standard) indicated by solid line boundaries separating the plot into four quadrants (a, b, c, d). Traditional measures of categorical performance include Accuracy $\left(A = \left(\frac{b+c}{a+b+c+d}\right)100\%\right)$, Bias $\left(B = \left(\frac{a+b}{b+d}\right)\right)$,

False Alarm Ratio $\left(FAR = \left(\frac{a}{a+b}\right)100\%\right)$, Critical Success Index $\left(CSI = \left(\frac{b}{a+b+d}\right)100\%\right)$, and Hit Rate $\left(H = \left(\frac{b}{b+d}\right)100\%\right)$. *A* measures the percentage of modeled estimates that correctly predicts the exceedance or non-exceedance of the measured variable. This metric can be greatly influenced by the non-exceedance category where exceedance is a clear minority event. *B* indicates fractional levels of model underprediction or overprediction of the threshold value (<1 indicates underprediction; >1 indicates overprediction). The *FAR* measures the percentage of time a threshold exceedance was modeled but did not actually occur in the observations. The *CSI* indicates how well the exceedances were estimated by considering false alarms and missed forecasts of exceedances; it is not skewed by the number of correctly forecast non-exceedances, as can happen with *A*. Finally, H, also known as the probability of detection, indicates the percentage of actual exceedances that are correctly estimated by the model.

These categorical metrics traditionally have been used with paired data, model/observations paired in space/time or at least in space. Forecasts from regional AQMs are often evaluated on a regional or metropolitan scale rather than at a single monitor, so the use of the above metrics can be limiting for that purpose. More recently a set of new categorical metrics has been demonstrated based on area-weighting concepts (Kang et al., 2007). For example, the area-weighted *A* and *FAR* metrics are defined by matching observed and modeled threshold exceedances within an area (i.e., model grid cells) surrounding the observation location. This new concept for application of categorical metrics better represents how air quality forecasts are issued in practice.

Metrics for determining the ability of the model to match space and time variations of the observed data The Figure of Merit (FOM) has been used for over 30 years and the Measure of Effectiveness (MOE) is a more recent metric for assessing the ability of the model to match observed space and/or time patterns.

FOM is a simple measure where fractional overlap is calculated of the areas enclosed by a modeled and observed contour corresponding, for example, to the NAAQS. For some problems, it is easy for a small 20 degree error in wind direction to cause the observed and modeled areas to not overlap at all, even though they are close to each other and have the same area. If agreement in time is also required, a slight error in timing can cause the modeled and predicted contours to miss each other. This was seen in the European Tracer Experiment (ETEX), where a cloud of tracer material was released in France and tracked for 2000 km and three days (Girardi et al., 1998). After three days the cloud was about 1000 km in diameter. Twenty regional linked meteorological and dispersion models were tested, and in many cases, the modeled and observed clouds "looked" about the same, but FOM in space and/or time was small because the observed cloud was displaced a few hundred km in space or a few hours in time.

MOE (Warner et al., 2004) is similar to FOM but allows for different weighting for false negatives (large underpredictions) versus false positives (large overpredictions). It is worse for a model to have false negatives because in that case the public would be told that they have no cause for alarm and it would happen that concentrations were high.

Determining statistical confidence limits on any metric Once a set of observed concentrations is available and sets of one or more model predictions are also available for the observed locations and times, it is possible to carry out statistical analyses with the paired data. For example, if O_{ij} represents the observation at location i and time j, and M_{ijk} represents the modeled value for model k at location i and time j, then at each i and j there is a set (O_{ij}, M_{ij1}, M_{ij2}, ...M_{ijK}), where K is the number of models. ASTM (2005), Efron (1983), and Hanna (1989) describe how bootstrap resampling and jackknife techniques can be used to determine statistical confidence limits on any metric and on the difference in the metrics calculated for pairs of models. The most common use of the method is, for example, to determine whether or not one model's mean bias is significantly different from zero at the 95% confidence level. Another widely-used test is to determine whether the difference between two models' metric is significantly different from zero at the 95% confidence level. For example, if model 1 has a FAC2 of 0.72 and model 2 has a FAC2 of 0.75, is this difference statistically significant?

REFERENCES

Aneja, V.P., J. Blunden, K. James, W.H. Schlesinger, R. Knighton, W. Gilliam, G. Jennings, D. Niyogi, and S. Cole, 2008: Ammonia assessment from agriculture: U.S. status and needs, *Journal of Environmental Quality*, **37**, 515-520.

Ansari, A.S., S.N. Pandis, 1998: Response of inorganic PM to precursor concentrations. *Environmental Science & Technology,* **32**, 2706-2714.

Appel, K.W., P.V. Bhave, A.B. Gilliland, G. Sarwar, and S.J. Roselle, 2008: Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance; Part II - particulate matter, *Atmospheric Environment*, **42**, 6057-6066.

Arnold, J.R., R.L. Dennis, and G.S. Tonnesson, 2003: Diagnostic evaluation of numerical air quality models with specialized ambient observations: Testing the Community Multiscale Air Quality modeling system (CMAQ) at selected SOS 95 ground sites, *Atmospheric Environment*, **37**, 1185-1198.

ASTM, 2005: Standard Guide D6569-05 for Statistical Evaluation of Atmospheric Dispersion Model Performance. Subcommittee D22.11 (John S. Irwin, chair).

Bachmann, J., 2007: Will the circle be unbroken: A history of the National Ambient Air Quality Standards. *Journal of Air & Waste Management Association*, **57**, 652-697.

Beven, K., 2002: Towards a coherent philosophy for modeling the environment. *Proc. R. Soc. London Ser. A-Math. Phys. Eng. Sci.,* **458**, 2465-2484

Bhave, P.V., G.A. Pouliot, and M. Zheng, 2007: Diagnostic model evaluation for carbonaceous $PM_{2.5}$ using organic markers measured in the southeastern U.S., *Environmental Science & Technology*, **41**, 1577-1583.

Biswas, J., and S.T. Rao, 2001: Uncertainties in episodic ozone modeling stemming from uncertainties in the meteorological fields. , *Journal of Applied Meteorology*, **40**, 117–136.

Bradley, Y.K.S., K.B. Brooks, L. Hubbard, P. Popp, and D.H. Stedman, 2000: Motor Vehicle Fleet Emissions by OP-FTIR, *Environmental Science & Technology*, **34**, 897-899

Britter, R.E., C. Collier, R. Griffiths, P. Mason, D. Thomson, R. Timmis and B. Underwood, 1995: Atmospheric dispersion modeling – Guidelines on the justification of choice and use of models, and the communication and reporting of results. Royal Meteorol. Soc. Policy Statement, RMS, 104 Oxford Rd., Reading RG1 7LJ, UK, 8 pp.

Brown, S. S., T. B. Ryerson, A. G. Wollny, C. A. Brock, R. Peltier, A. P. Sullivan, R. J. Weber, W. P. Dube, M. Trainer, J. F. Meagher, F. C. Fehsenfeld, and A. R. Ravishankara, 2006: Variability in nocturnal nitrogen oxide processing and its role in regional air quality. *Science*, **311**, 67-70.

Camalier, L., Cox, W., and Dolwick, P., 2007: The effects of meteorology on ozone in urban areas and their use in assessing ozone trends, *Atmospheric Environment*, **41**, 7127–7137.

Carvalho, A.C., L. Menut, R. Vautard, and J. Nicolau, 2007: Air quality ensemble forecast coupling ARPEGE and CHIMERE over Western Europe., Proceedings of the 29th NATO/CCMS International Technical Meeting on Air Pollution Modeling and its Application, September 24-28, 2007, Aveiro, Portugal, pp. 357 – 363.

Chang, J.C. and S.R. Hanna, 2004: Air quality model performance. *Meteorology and Atmospheric Physics*, **87**, 167-196.

Chow, J.C., 2003: Introduction to Special Topic: Weekend and weekday differences in ozone levels. *Journal of Air & Waste Management Association*, **53**,771.

Cohan, D.S., A. Hakami, Y. Hu and A.G. Russell, 2005: Nonlinear response of ozone to emissions: Source apportionment and sensitivity analysis, *Environmental Science and Technology*, **39**, 6739–6748.

Cressie, N.A., 1993: Statistics for Spatial Data. Revised Edition. Wiley, New York.

Cullen, A.C. and H.C. Frey, 1999: *Probabilistic Techniques in Exposure Assessment. A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. Plenum Press, New York, 335 pp.

Dabberdt, W.F., M.A. Carroll, D. Baumgardner, G. Carmichael, R. Cohen, T. Dye, J. Ellis, G. Grell, S. Grimmond, S. Hanna, J. Irwin, B. Lamb, S. Madronich, J. McQueen, J. Meagher, T. Odman, J. Pleim, H.P. Schmid and D.L. Westphal, 2004: Meteorological research needs for improved air quality forecasting, *Bulletin of American Meteorological Society,* **85**, 563–586.

Delle Monache, L., X. Deng, Y. Zhou, R. Stull, 2006a: Ozone ensemble forecasts: 1. A new ensemble design, *Journal of Geophysical Research*, **111**, D05307, doi:10.1029/2005JD006310.

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, R. B. Stull, 2006b: Probabilistic aspects of meteorological and ozone regional ensemble forecasts, *Journal of Geophysical Research*, **111**, D24307, doi:10.1029/2005JD006917.

Demerjian, K.L., 1985: Quantifying uncertainty in long range transport models: A summary of the AMS Workshop on Sources and Evaluation of Uncertainty in Long Range Transport Models. *Bulletin of American Meteorological Society*, **66**, 1533-1540.

Dennis, R. L., P. Bhave, and R.W. Pinder, 2008: Observable indicators of the sensitivity of $PM_{2.5}$ nitrate to emission reductions, part II: sensitivity to errors in total ammonia and total nitrate of the CMAQ-predicted nonlinear effect of SO_2 emission reductions on $PM_{2.5}$ nitrate, *Atmospheric Environment*, **42**, 1287-1300.

Eder, B., D. Kang, R. Mathur, S.Yu and K. Schere, 2006: An operational evaluation of the Eta-CMAQ air quality model. *Atmospheric Environment*, **40**, 4894-4905.

Efron, B., 1983: Estimating the error rate of prediction rules: Improvement on cross-validation, *Journal of the American Statistical Association*, **78**, 316-331.

Emeis, S., C. Münkel, S.Vogt, W.J. Muller, and K.Schäfer, 2004: Atmospheric boundary-layer structure from simultaneous SODAR, RASS, and ceilometer measurements, *Atmospheric Environment*, **38**, 273-286.

Fine, J., L. Vuilleumier, S. Reynolds, P. Roth, N. Brown, 2003: Evaluating uncertainties in regional photochemical air quality modeling. *Annual Review of Environment and Resources*, **28**(1): 59-106.

Fox, D.G., 1984: Uncertainty in Air Quality Modeling. *Bulletin of American Meteorological Society,* **65**, 27-35.

Fox, D.G., 1981: Judging air quality model performance: A summary of the AMS Workshop on Dispersion Model Performance, *Bulletin of American Meteorological Society,* **62**, 599-609.

Frost, G.J., et al., 2006: Effects of changing power plant NO_x emissions on O_3 in the eastern United States: Proof of concept, *Journal of Geophysical Research*, **111**, D12306.

Fuentes, M. and A. Raftery, 2005: Model evaluation and spatial interpolation by Bayesian combination of observations with output from numerical models. *Biometrics*, **61**, 36-45.

Galmarini, S., R. Bianconi, W. Klug, T. Mikkelsen, R. Addis, S. Andronopoulos, P. Astrup, A. Baklanov, J. Bartniki, J.C. Bartzis, R. Bellasio, F. Bompay, R. Buckley, M. Bouzom, H. Champion, R. D'Amours, E. Davakis, H. Eleveld, G.T. Geertsema, H. Glaab, M. Kollax, M. Ilvonen, A. Manning, U. Pechinger, C. Persson, E. Polreich, S. Potemski, M. Prodanova, J. Saltbones, H. Slaper, M.A. Sofiev, D. Syrakov, J.H. Soerensen, L. Van der Auwera, I. Valkama and R. Zelazny, 2004a: Ensemble dispersion forecasting, part I: concept approach, and indicators. *Atmospheric Environment*, **38**, 4607-4617.

Galmarini, S., R. Bianconi, R. Addis, S. Andronopoulos, P. Astrup, J.C. Bartzis, R. Bellasio, R. Buckley, H. Champion, M. Chino, R. D'Amours, E. Davakis, H. Eleveld, H. Glaab, A. Manning, T. Mikkelsen, U. Pechinger, E. Polreich, M. Prodanova, H. Slaper, D. Syrakov, H. Terada, L. Van der Auwera, 2004b : Ensemble dispersion forecasting, part II: application and evaluation. *Atmospheric Environment*, **38**, 4619-4632.

Gégo, E., A. Gilliland, J. Godowitch, S.T. Rao, P.S. Porter, and C. Hogrefe, 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, *Journal of Air and Waste Management Association*, **58**, 580-588.

Gégo, E., P.S. Porter, V. Garcia, C. Hogrefe, J. Swall, A. Gilliland, and S.T. Rao, 2007: Enhanced ozone spatial fields: comparison of techniques. 29th NATO/SPS International Technical Meeting on Air Pollution Modelling and its Application, 24 - 28 September 2007, University of Aveiro, Aveiro, Portugal.

Gégo, E.L., P.S. Porter, J.S. Irwin, C. Hogrefe, and S. T. Rao, 2005: Assessing the comparability of ammonium, nitrate and sulfate concentrations measured by three air quality monitoring networks, *Pure and Applied Geophysics*, **162**, 1919-1939.

Gégo et al., 2003: Probabilistic assessment of regional scale ozone pollution in the eastern United States. In Air Pollution in Regional Scale. Proceedings of the NATO Advanced Research Workshop, Kallithea, Halkidiki, Greece, June 13–15, 2003. NATO Science Series: IV. Earth and Environmental Sciences. D. Melas, and D. Syrakov (Eds.). Kluwer Academic Publishers, 87–96.

Gertler, AW., J.C. Sagebiel, D.N. Wittorf, W.R. Pierson, W.A. Dippel, D. Freeman, L. Sheetz, 1997: Vehicle Emissions in Five Urban Tunnels, CRC Project No. E-5; Coordinating Research Council, Prepared by Desert Research Institute: Reno,NV.

Gilliland, A.B., C. Hogrefe, R.W. Pinder, J.M. Godowitch, K.L. Foley, and S.T. Rao, 2008: Dynamic evaluation of regional air quality models: Assessing changes in O_3 stemming from changes in emissions and meteorology, *Atmospheric Environment*, **42**, 5110-5123.

Gilliland, A.B., K.W. Appel, R. Pinder, S.J. Roselle, and R.L. Dennis, 2006: Seasonal NH_3 emissions for an annual 2001 CMAQ simulation: inverse model estimation and evaluation, *Atmospheric Environment*, **40**, 4986-4998.

Gilliland, A.B., R.L. Dennis, S.J. Roselle, T.E. Pierce, 2003: Seasonal NH_3 emission estimates for the Eastern United States using ammonium wet concentrations and an inverse modeling method, *Journal of Geophysical Research-Atmospheres*, **108**, NO. D15, 4477, 10.1029/2002JD003063.

Girardi, F, G. Graziani, D. van Velzen, S. Galmarini, S. Mosca, R. Bianconi, R. Bellasio, W. Klug, and G. Fraser, 1998: ETEX – The European Tracer Experiment, EUR 18143EN, ISBN 92-828-5007-2, JRC/European Commission, 106 pp.

Godowitch, J.L., A.B. Gilliland, R. Draxler, S.T. Rao, 2008: Modeling assessment of point source NO_x emission reductions on ozone air quality in the eastern United States, *Atmospheric Environment*, **42**, 87-100.

Hanna, S.R. and J.M. Davis, 2002: Evaluation of a photochemical grid model using estimates of concentration probability density functions, *Atmospheric Environment*, **36**, 1793-1798.

Hanna, S.R, and R. Yang, 2001: Evaluations of mesoscale model predictions of near-surface winds, temperature gradients, and mixing depths. *Journal of Applied Meteorology*, **40**, 1095-1104.

Hanna, S.R., Z. Lu, H.C. Frey, N. Wheeler, J. Vukovich, S. Arumachalam and M. Fernau, 2001: Uncertainties in predicted ozone concentration due to input uncertainties for the UAM-V photochemical grid model applied to the July 1995 OTAG domain. *Atmospheric Environment*, **35**, 891-903.

Hanna S.R., J.C. Chang, and M.E. Fernau, 1998: Monte Carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. *Atmospheric Environment,* **32**, 3619-3628.

Hanna, S.R., J.C. Chang and D.G. Strimaitis, 1993: Hazardous gas model evaluation with field observations. *Atmospheric Environment*, **27A**, 2265-2285.

Hanna, S.R., 1989: Confidence limits for air quality models, as estimated by bootstrap and jackknife resampling methods. *Atmospheric Environment*, **23**, 1385-1395.

Hanna S.R.and F.A. Gifford, 1971: Conference summary: Working meeting on mesoscale numerical modeling. *Bulletin of American Meteorological Society*, **52**:993.

Henderson, B., 2008: The Influence of Model Resolution on Ozone from Industrial VOC Releases, Master's Thesis, UNC Chapel Hill.

Hogrefe, C., K.L. Civerolo, W. Hao, J.-Y. Ku, E.E. Zalewsky, and G. Sistla, 2008: Rethinking the assessment of photochemical modeling systems in air quality planning applications, *Journal of Air & Waste Management Association*, **58**, 1086-1099.

Hogrefe, C., and S. T. Rao, 2001: Demonstrating attainment of the air quality standards: integration of observations and model predictions into the probabilistic framework. *Journal of Air & Waste Management Association*, **51**, 1060 – 1072.

Hogrefe, C., S.T. Rao, P. Kasibhatla, G. Kallos, C.J. Tremback, W. Hao, D. Olerud, A. Xiu, J. McHenry and K. Alapaty, 2001: Evaluating model performance of regional-scale photochemical modeling systems – meteorological predictions. *Atmospheric Environment*, **35,** 4159-4174.

Hogrefe, C., S. T. Rao, I. G. Zurbenko, and P. S. Porter, 2000: 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,* **81**, 2083 – 2106.

Hudman, R. C., D.J. Jacob, S. Turquety, E.M. Leibensperger, L.T. Murray, S. Wu, A.B. Gilliland, M. Avery, T.H. Bertram, W. Brune, R.C. Cohen, J.E. Dibb, F.M. Flocke, A. Fried, J. Holloway, J.A. Neuman, R. Orville, A. Perring, X. Ren, G.W. Sachse, H.B. Singh, A. Swanson, P.J. Wooldridge, 2007: Surface and lightning sources of nitrogen oxides over the United States: magnitudes, chemical evolution, and outflow, *Journal of Geophysical Research*, **112**, D12S05, doi:10.1029/2006JD007912.

Irwin, J.S., K. Civerolo, C. Hogrefe, W. Appel, K. Foley, and J. Swall, 2008: A procedure for inter-comparing the skill of regional-scale air quality model simulations of daily maximum 8-h ozone concentrations, *Atmospheric Environment*, **42**, 5403-5412.

Irwin, J.S., Rao, S.T., Petersen, W.B., and Turner, D.B., 1987: Relating error bounds for maximum concentration estimates to diffusion meteorology uncertainty, *Atmospheric Environment,* **21**, 1927–1937.

Jolliffe, I.T. and D.B. Stephenson, 2003: Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley, West Sussex, England, 240pp.

Jolliffe, I.T., 2002: Principal Component Analysis, Second Edition. Springer, New York, 487 pp.

Jones, J.M., C. Hogrefe, R.F. Henry, J.-Y. Ku, and G. Sistla, 2005: 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*, **55**, 13–19.

Jones, M., B. A. Colle, and J. Tongue, 2007: Evaluation of a short-range ensemble forecast system over the Northeast U.S., *Weather and Forecasting*, **22**, 36-55

Kang, D., R. Mathur, K. Schere, S. Yu and B. Eder, 2007: New categorical metrics for air quality model evaluation. *Journal of Applied Meteorology and Climatology*, **46**, 549-555.

Kleinman, L.I., P. Daum, J.H. Lee, Y.-N. Lee, L. Nunnermacker, S. Springston, J. Weinstein-Lloyd, L. Newman, and S. Sillman, 1997. Dependence of ozone production on NO and hydrocarbons in the troposphere, *Geophysical Research Letters*, **24**, 2299-2302.

Kleinman, L.I., 1994. Low- and high-NO_x tropospheric photochemistry, *Journal of Geophysical Research*, **99**, 16,831-16,838.

Lamb, R.G. and S.K. Hati, 1987: The representation of atmospheric motion in models of regional-scale air pollution. *Journal of Applied Meteorology*, **26**, 837-846.

Lewellen, W.S., R.I. Sykes and S.F. Parker, 1985: An evaluation technique which uses the prediction of both concentration mean and variance. Proceedings of the DOE/AMS Air Pollution Model Evaluation Workshop, Savannah River Lab Report Number DP-1701-1, Section 2, 24 pages.

Mallet, V., B. Sportisse, 2006a: Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: An ensemble approach applied to ozone modeling, *Journal of Geophysical Research*, **111**, D01302, doi:10.1029/2005JD006149.

Mallet, V., B. Sportisse, 2006b: Ensemble-based air quality forecasts: A multimodel approach applied to ozone, *Journal of Geophysical Research*, **111**, D18302, doi:10.1029/2005JD006675.

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 mode. *Journal of Geophysical Research* (in press).

McKeen, S., et al., 2005: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, *Journal of Geophysical Research*, **110**, D21307, doi:10.1029/2005JD005858.

Menut, L., 2003: Adjoint modeling for atmospheric pollution process sensitivity at regional scale, *Journal of Geophysical Research* **108** (2003) (D17), p. 8562.

Menut, L., R. Vautard, M. Beekmann and C. Honore, 2000. Sensitivity of photochemical pollution using the adjoint of a simplified chemistry-transport model, *Journal of Geophysical Research,* **105**, 15379–15402.

Morris, R.E., D.E. McNally, T.W. Tesche, G. Tonnesen, J.W. Boylan, and P. Brewer, 2005: Preliminary evaluation of the Community Multiscale Air Quality Model for 2002 over the southeastern United States. *Journal of Air & Waste Management Association*, **55**, 1694-1708.

Napelenok, S.L., R.W. Pinder, A.B. Gilliland, R.V. Martin, 2008: A method for evaluating spatially-resolved NO_x emissions using Kalman filter inversion, direct sensitivities, and space-based NO_2 observations, *Atmospheric Chemistry and Physics*, **8**, 6469–6499.

Napelenok, S.L., D.S. Cohan, Y.Hu, A.G. Russell, 2006: Decoupled direct 3D sensitivity analysis for particulate matter (DDM-3D/PM), *Atmospheric Environment*, **40**, 6112-6121.

NARSTO, 2006: Improving Emission Inventories for Effective Air Quality Management Across North America, A NARSTO Assessment, NARSTO-05-001

NRC, 2007: Models in Environmental Regulatory Decision Making, National Academies Press, Washington, DC, 267 pp.

Oreskes, N., K. Shrader-Frechette and K. Beitz, 1994: Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, **263**, 641-646.

Otte, T.L., 2008: 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*, **47**, 1868-1887.

Parrish, D., 2005: Top-down assessments and emission inventories. In Improving Emission Inventories for Effective Air Quality Management Across North America, NARSTO-05-001, NARSTO Emission Inventory Assessment Team (Eds.), NARSTO, 197-219.

Parrish, D.D., M. Trainer, D. Hereid, E.J. Williams, K.J. Olszyna, R.A. Harley, J.F. Meagher, F.C. Fehsenfeld, 2002: Decadal change in carbon monoxide to nitrogen oxide ratio in U.S.

vehicular emissions, *Journal of Geophysical Research,* **107** (D12), 4140, doi:10.1029/2001JD000720.

Parrish, D., M. Buhr, M. Trainer, R. Norton, J. Shimshock, F. Fehsenfeld, G. Anlauf, J. Bottemheim, Y. Tang, H. Wiebe, J. Roberts, R. Tanner, L. Newman, V. Bowersox, K. Olszyna, E. Bailey, M. Rodgers, T. Wang, H. Berresheim, U. Roychowdhury, K. Demerjian, 1993: The total reactive oxidized nitrogen levels and the partitioning between the individual species at six rural sites in eastern North America, *Journal of Geophysical Research*, **98**, 2927-2939.

Pinder, R.W., R.C. Gilliam, K.W. Appel, S.L. Napelenok, and A.B. Gilliland, 2008: Efficient probabilistic estimates of ozone concentration using an ensemble of model configurations and direct sensitivity calculations, *Environmental Science & Technology* (in review).

Pinder, R.W., P.J. Adams, S.N. Pandis, A.B. Gilliland, 2006: Temporally resolved ammonia emission inventories: Current estimates, evaluation tools, and measurement needs, *Journal of Geophysical Research-Atmospheres*, **111**, doi:10.1029/2005JD006603.

Porter, P.S., S.T. Rao, I.G. Zurbenko, A.M. Dunker, and G.T. Wolff, 2001: Ozone air quality over North America: Part II-An analysis of trend detection and attribution techniques, *Journal of Air & Waste Management Association*, **51**, 283-306.

Rao, S.T., Hogrefe, C., H. Mao, J. Biswas, I. G. Zurbenko, P. S. Porter, P. Kasibhatla, and D. A. Hansen, 2001: *How should the Photochemical Modeling Systems be Used in Guiding Emissions Management Decisions?* In: Air Pollution Modeling and Its Application XIV; Gryning, S. E. and F. Schiermeier, eds., Kluwer Academic/Plenum, New York, 25 – 34.

Rao, S.T., I.G. Zurbenko, R. Neagu, P.S. Porter, J.Y. Ku, and R.F. Henry, 1997: Space and time scales in ambient ozone data. *Bulletin of American Meteorological Society*, **78**, 2153 – 2166.

Russell, A. and R. Dennis, 2000: NARSTO Critical Review of photochemical models and modeling. *Atmospheric Environment*, **34**, 2283-2324.

Saltelli, A., S. Tarantola, F. Campolongo and M. Ratto, 2004: *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*. John Wiley & Sons.

Schere, K.L. and C.J. Coats, 1992: A stochastic methodology for regional wind-field modeling. *Journal of Applied Meteorology*, **31**, 1407-1425.

Seaman, N.L., 2000: Meteorological modeling for air-quality assessments, *Atmospheric Environment*, **34**, 2231-2259.

Sillman, S., D. He, M.R. Pippin, P.H. Daum, D.G. Imre, L.I Kleinman, J.H. Lee, and J. Weinstein-Lloyd, 1998: Model correlations for ozone, reactive nitrogen and peroxides for Nashville in comparison with measurements: Implications for O_3-NO_x-hydrocarbon chemistry, *Journal of Geophysical Research*, **103**, 22,629-22,644.

Sillman, S., 1995. The use of NO_y, H_2O_2, and HNO_3 as indicators for ozone-NO_x-hydrocarbon sensitivity in urban locations, *Journal of Geophysical Research*, **100**, 14175-14188.

Sistla, G., C. Hogrefe, W. Hao, J.-Y. Ku, E. Zalewsky, R. F. Henry and K. Civerolo, 2004: An operational assessment of the application of the relative reduction factors (RRF) in demonstration of attainment of the 8-hr ozone national ambient air quality standard (NAAQS*), Journal of Air & Waste Management Association*, **54**, 950-959.

Steyn, D. G. and S. Galmarini, 2008: Evaluating the Predictive and Explanatory Value of Atmospheric Numerical Models: Between Relativism and Objectivism. *The Open Atmospheric Science Journal*, **2***,* 38-45. doi: 10.2174/1874282300802010038.

Stohl, A., M. Hittenberger and G. Wotawa, 1998: Validation of the lagrangian particle dispersion model FLEXPART against large-scale tracer experiment data. *Atmospheric Environment*, **32**, 4245-4264.

Sykes, R.I., S. Parker, D. Henn, and B. Chowdhury, 2007: SCIPUFF Version 2.3 Technical Documentation. L-3 Titan Corp., POB 2229, Princeton, NJ 08543-2229, 336 pages.

Sykes, R.I., W.S. Lewellen, S.F. Parker and D.S. Henn, 1988: A hierarchy of dynamic plume models incorporating uncertainty. Vol. 4, Second-Order Closure Integrated Puff, EPRI EA-6095, 99 pages, available from ARAP/Titan, POB 229, Princeton, NJ 08543-2229.

Sykes, R.I., W.S. Lewellen and S.F. Parker, 1984: A turbulent transport model for concentration fluctuations and fluxes. *Journal of Fluid Mechanics*, **139**, 193-218.

Taylor, K.E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, 106(D7), 7183-7192.

Tonnesen, G.S. and R.L. Dennis, 2000a: Analysis of Radical Propagation Efficiency to Assess Ozone Sensitivity to Hydrocarbons and NO_x. Part 1: Local Indicators of Instantaneous Odd Oxygen Production Sensitivity, *Journal of Geophysical Research*, **105**, 9213-9225.

Tonnesen, G.S. and R.L. Dennis, 2000b: Analysis of Radical Propagation Efficiency to Assess Ozone Sensitivity to Hydrocarbons and NO_x. Part 2: Long-Lived Species as Indicators of Ozone Concentration Sensitivity, *Journal of Geophysical Research*, **105**, 9227-9241.

Trainer, M., D. Parish, M. Buhr, R. Norton, F. Fehsenfeld, G. Anlauf, J. Bottemheim, Y. Tang, H. Wiebe, J. Roberts, R. Tanner, L. Newman, V. Bowersox, J. Meagher, K. Olszyna, M. Rodgers, T. Wang, H. Berresheim, K. Demerjian, U. Roychowdhury, 1993: Correlation of ozone with NO_y in photochemically aged air, *Journal of Geophysical Research*, **98**, 2917-2925.

Tukey, J.W., 1977: *Exploratory Data Analysis*, Addison-Wesely, New York, 688 pp.

Uliasz, M., 1988: Application of the FAST method to analyze the sensitivity-uncertainty of a Lagrangian model of sulfur transport in Europe. *Water, Soil and Air Pollution*, **40**, 33-49.

U.S. EPA., 2007: Guidance on the Use of Models and Other Analyses for Demonstrating Attainment of Air Quality Goals for Ozone, $PM_{2.5}$, and Regional Haze, U.S. Environmental Protection Agency, Research Triangle Park, NC, EPA-454/B-07-002, 262 pp.

U.S. EPA, 2006: Revisions to Ambient Air Monitoring Regulations, CFR Parts 53 and 58, vol 71FR 61236, October 17, 2006.

Vautard, R., P.H.J. Builtjes, P. Thunis, C. Cuvelier, M. Bedogni, B. Bessagnet, C. Honoré, N. Moussiopoulos, G. Pirovano, M. Schaap, R. Stern, L. Tarrason, and P. Wind, 2007: Evaluation and intercomparison of Ozone and PM10 simulations by several chemistry transport models over four European cities within the CityDelta project. *Atmospheric Environment*, **41**, 173-188.

Vautard R., et al., 2006: Is regional air quality model diversity representative of uncertainty for ozone simulation?, *Geophysical Research Letters*, **33**, L24818, doi:10.1029/2006GL027610.

Venkatram, A., 1979: The expected deviation of observed concentrations from predicted ensemble means. *Atmospheric Environment*, **13**, 1547-1549.

Venkatram, A., 1988: Inherent uncertainty in air quality modeling. *Atmospheric Environment*, **22**, 1221-1227.

Vizuete, W., M. Kioumourtzoglou, H. Jeffries, B. Henderson, 2008a: Effects of radical source strengths on ozone formation in models for Houston, Texas, in review at *Atmospheric Environment*

Vizuete, W., B. Kim, H.E. Jeffries, Y. Kimura, D.T. Allen, M. Kioumourtzoglou, L. Biton, B. Henderson, 2008b: Modeling ozone formation from industrial emission events in Houston, Texas, in press at *Atmospheric Environment*, doi: ATMENV-D-07-01368R2.

Warner, S., Platt, N., and Heagy, J. F., 2004: User-oriented two-dimensional measure of effectiveness for the evaluation of transport and dispersion models. *Journal of Applied Meteorology*, **43**, 53-73

Warner, T.T., R.-S. Sheu, J.F. Bowers, R.I. Sykes, G.C. Dodd and D.S. Henn, 2002: Ensemble simulations with coupled atmospheric dynamic and dispersion models: Illustrating uncertainties in dosage simulations. *Journal of Applied Meteorology*, **41**, 488-504.

Weil, J.C., R.I. Sykes and A. Venkatram, 1992: Evaluating air quality models: review and outlook. *Journal of Applied Meteorology*, **31**, 1121-1145.

Wittig, A.E. and D.T. Allen, 2008: Improvement of the Chemical Mass Balance model for apportioning sources of non-methane hydrocarbons using composite aged source profiles,

Atmospheric Environment, **42**, 1319-1337.

Yu S., B. Eder, R. Dennis, S. Chu, S.E. Schwartz, 2006: New unbiased symmetric metrics for evaluation of air quality models. *Atmospheric Science Letters*, **7**, 26-34.

Zhang, F., N. Bei, J. W. Nielsen-Gammon, G. Li, R. Zhang, A. Stuart, A. Aksoy, 2007: Impacts of meteorological uncertainties on ozone pollution predictability estimated through meteorological and photochemical ensemble forecasts, *Journal of Geophysical Research*, **112**, D04304, doi:10.1029/2006JD007429.

Zhang, J. and S.T. Rao, 1999: The role of vertical mixing in the temporal evolution of ground-level ozone concentrations, *Journal of Applied Meteorology*, **38**, 1674-1691.

Figures

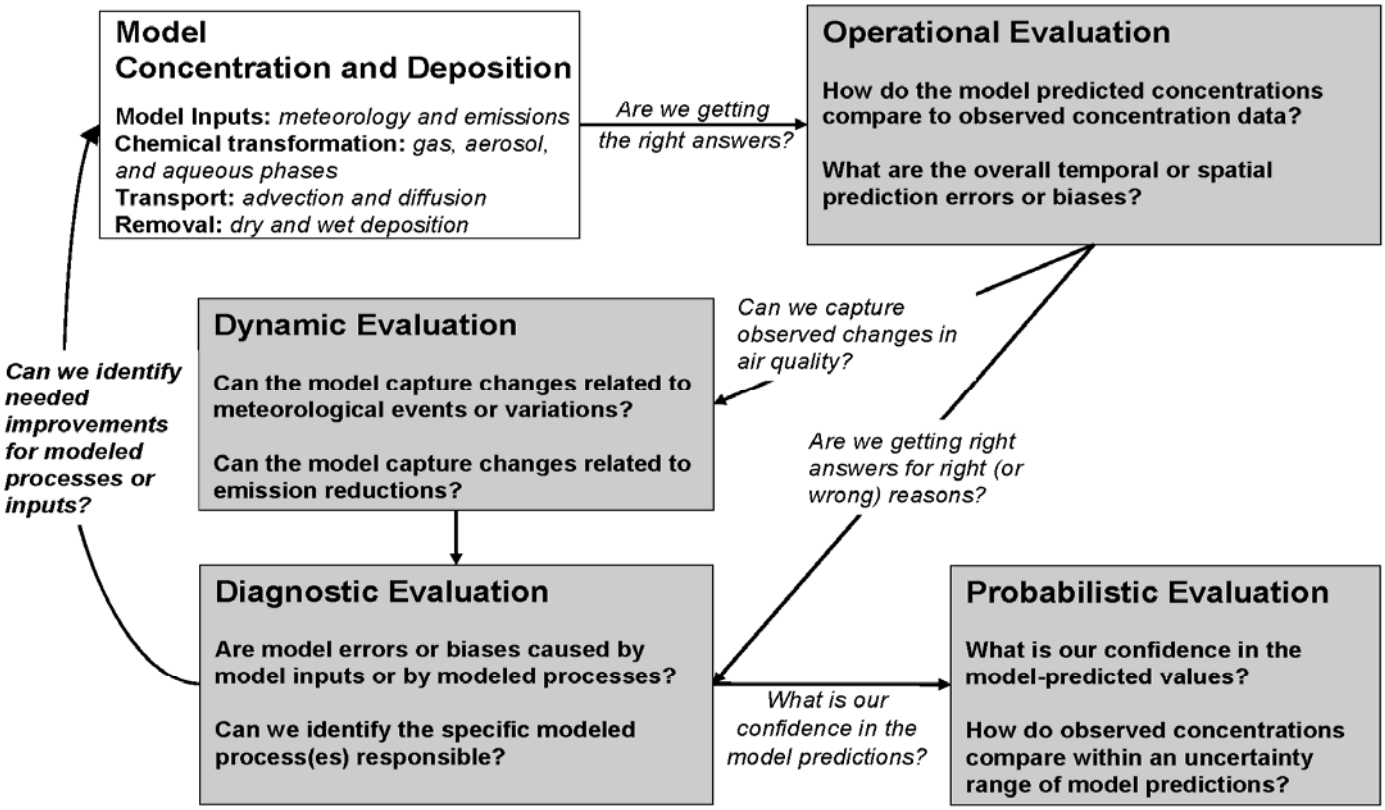

Model
Concentration and Deposition
Model Inputs: meteorology and emissions
Chemical transformation: gas, aerosol, and aqueous phases
Transport: advection and diffusion
Removal: dry and wet deposition
Are we getting the right answers?
Operational Evaluation
How do the model predicted concentrations compare to observed concentration data?
What are the overall temporal or spatial prediction errors or biases?
Can we capture observed changes in air quality?
Dynamic Evaluation
Can the model capture changes related to meteorological events or variations?
Can the model capture changes related to emission reductions?
Can we identify needed improvements for modeled processes or inputs?
Are we getting right answers for right (or wrong) reasons?
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?

Figure 1.

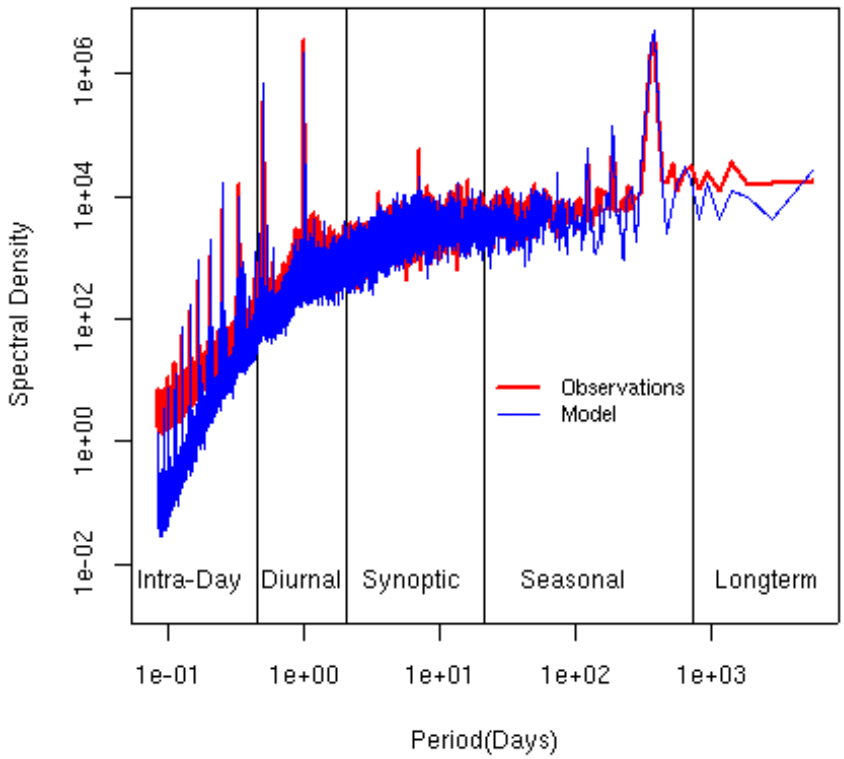
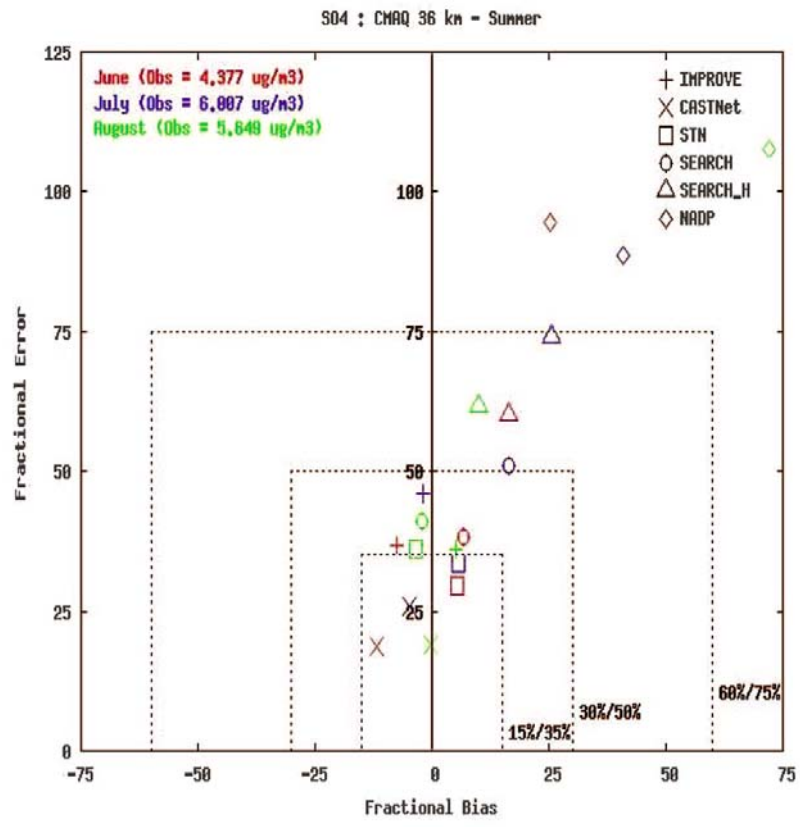


Figure 2.



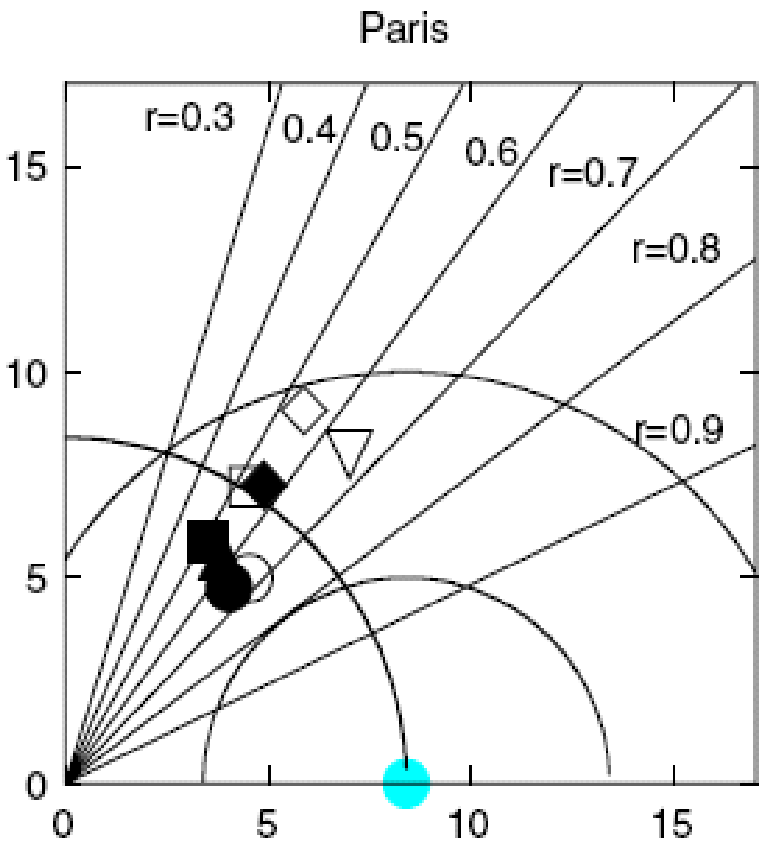


Fig 3. (a-top; b-bottom)

Figure 4.

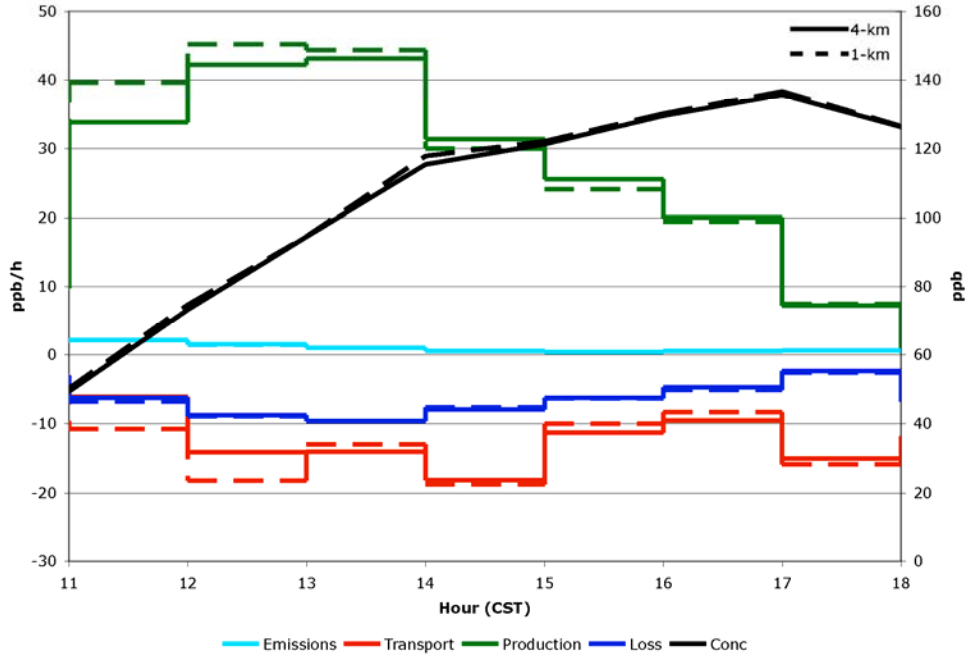
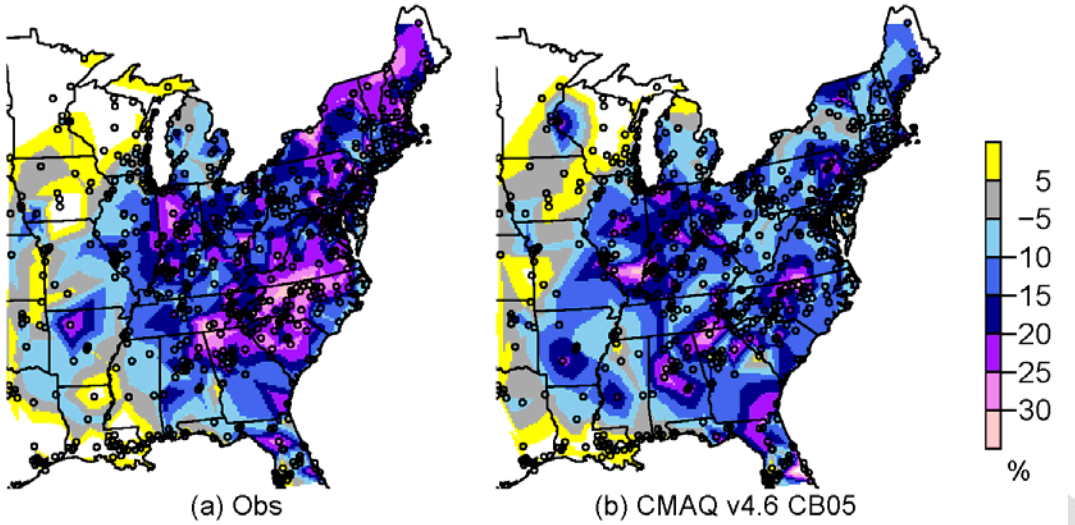


Figure 5.



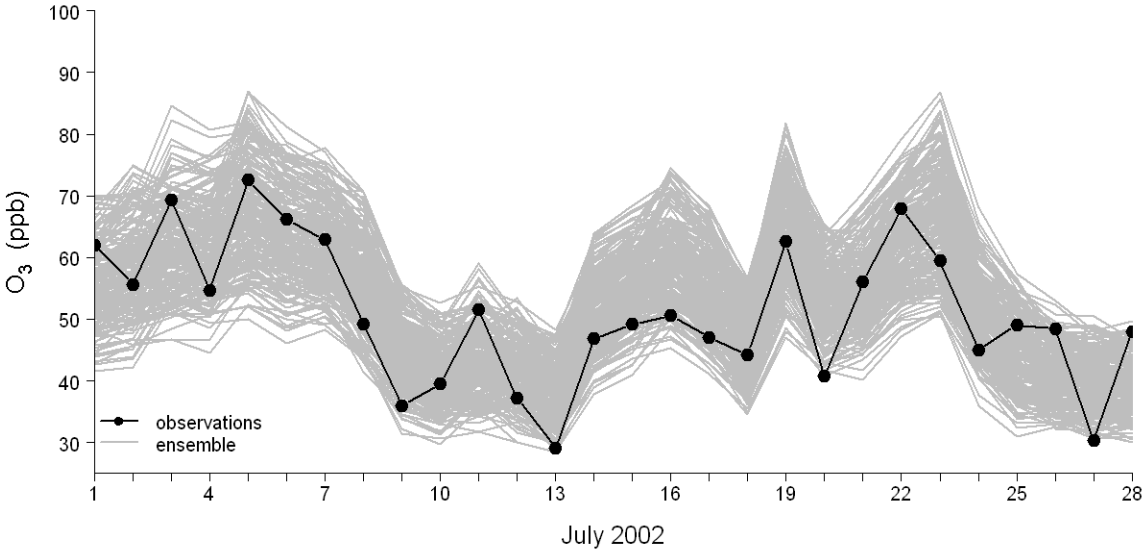


Figure 6.

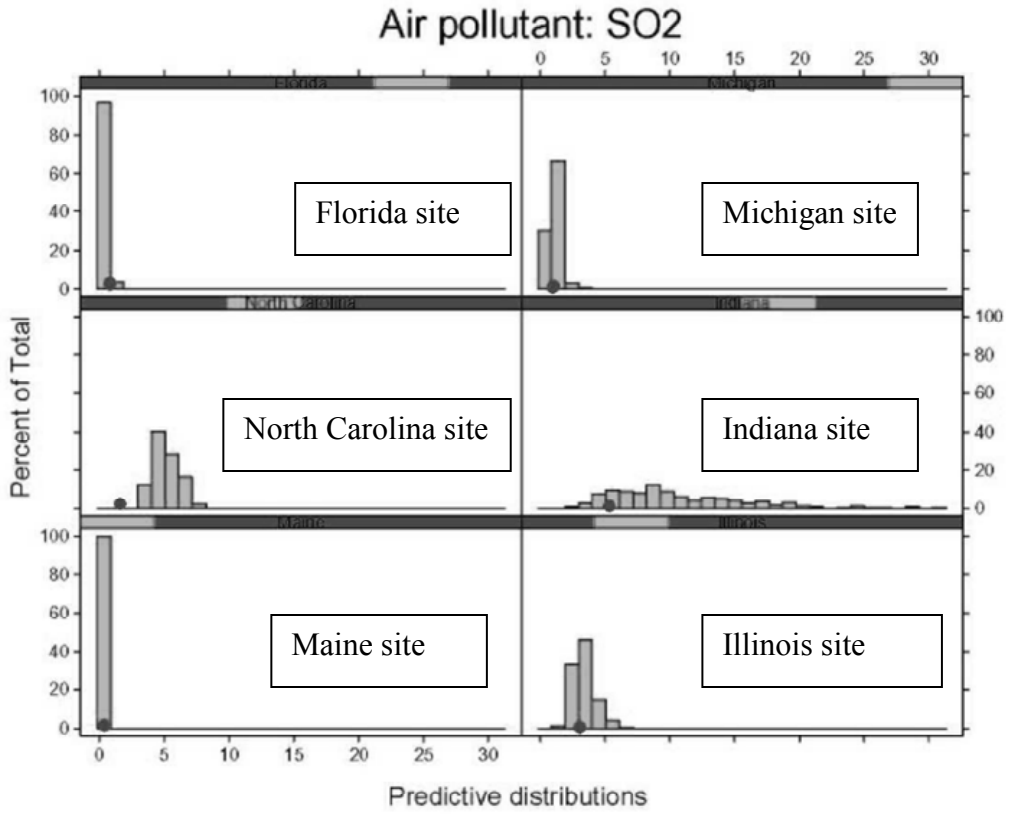


Figure 7.

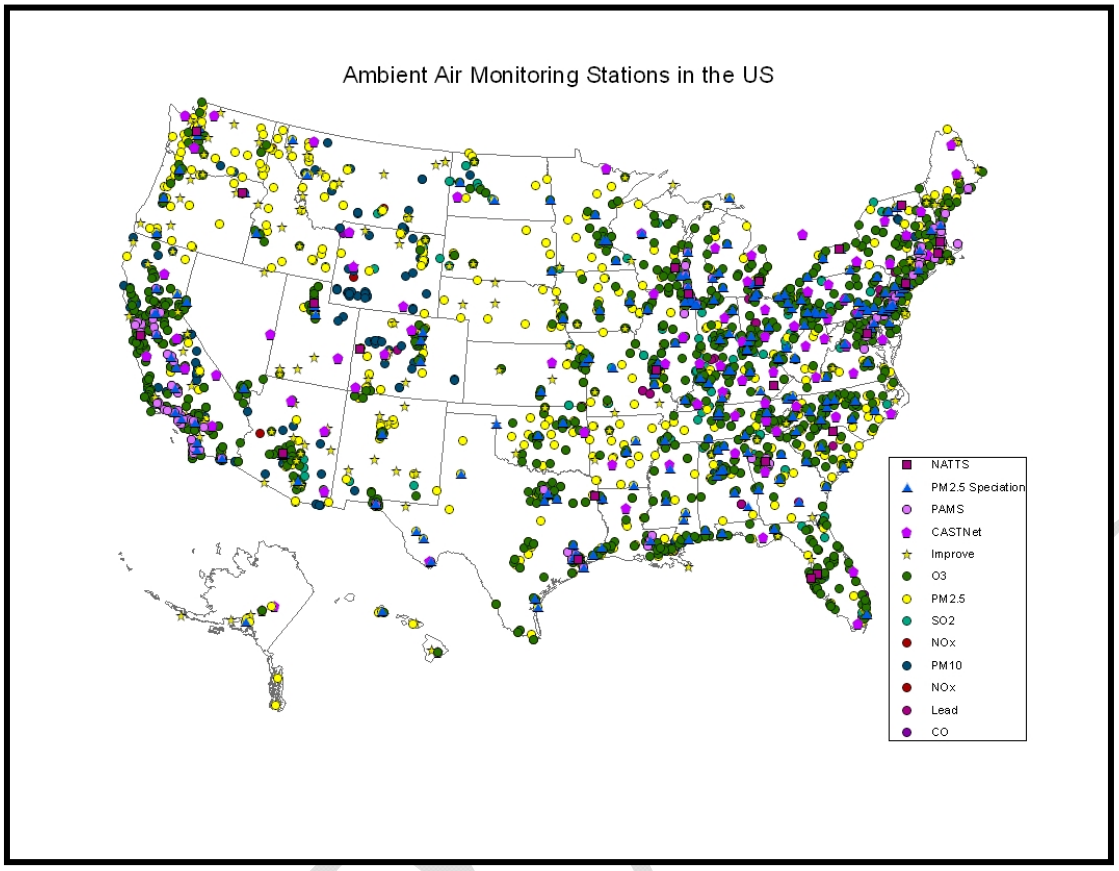


Figure 8.

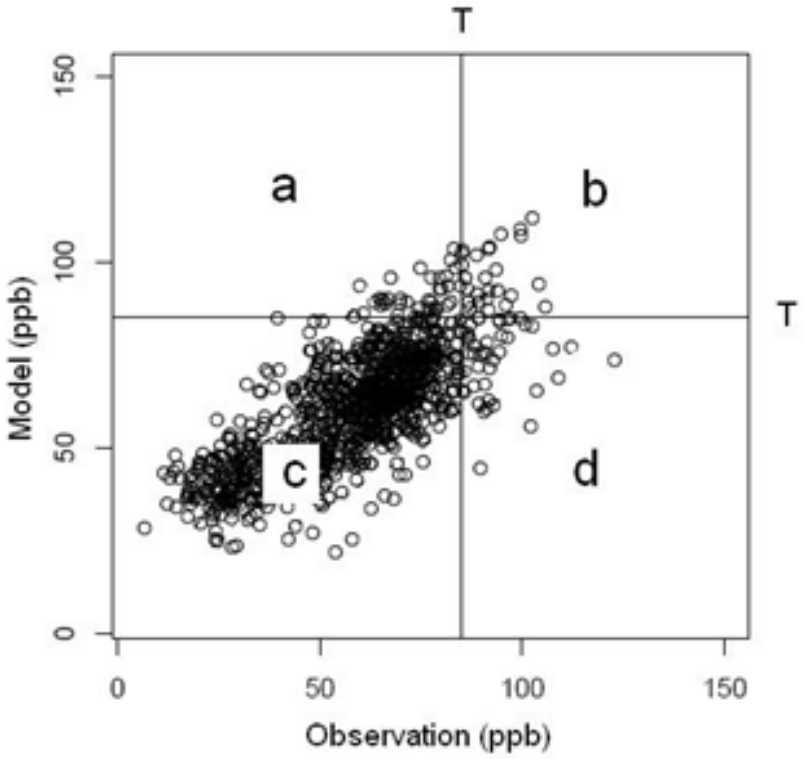


Figure A-1.