Canepa, E. and J. Irwin 2005. *Evaluation of Air Pollution Models*. Chapter 17 of AIR QUALITY MODELING - *Theories, Methodologies, Computational Techniques, and Available Databases and Software. Vol. II – Advanced Topics* (P. Zannetti, Editor). Published by The EnviroComp Institute (http://www.envirocomp.org/) and the Air & Waste Management Association (http://www.awma.org/).

# Chapter 17

# Evaluation of Air Pollution Models

Elisa Canepa [(1)] and John S. Irwin [(2)]

[(1)] *INFM (National Institute for the Physics of Matter), Department of Physics - University of Genova, Genova (Italy)*
*elisa.canepa@fisica.unige.it*
[(2)] *John S. Irwin and Associates[1], Raleigh, NC (USA)*
*jsirwinetal@nc.rr.com*

**Abstract:** Information is given about model evaluation, the overall system of procedures designed to measure model performance, and in particular, the process of statistical performance evaluations. Statistical performance evaluation is an assessment of model performance based on the comparison of model outputs with experimental data. Some performance measures, consisting of statistical indices and graphical methodologies, currently used are described. Problems related to uncertainty analysis are highlighted.

**Key Words:** model quality assurance, model evaluation, statistical model evaluation, uncertainty analysis, statistical indices, performance measures.

## 1 Introduction

Model quality assurance is a collection of activities one should perform in order to promote the development and application of good air quality simulation models (discussed in more detail in Section 8 below). One of the elements of model quality assurance is model evaluation. Model evaluation[2] is a collection of activities one should perform in order to understand how a model behaves and how a model compares with observations (discussed in more detail in Section 6).

[1] From 1975 to 2004, John Irwin was a NOAA employee, on assignment to the U.S. Environmental Protection Agency, Research Triangle Park, NC 27711.

[2] Readers will see that we have avoided the use of the term "validation". Fox (1981) and Olesen (1996) define "validation" as a conclusion resulting from detailed and copious evidence that leads to formal recognition, which might include several evaluations.

 

One of the elements of model evaluation is statistical model evaluation. Statistical model evaluation, also called "statistical performance evaluation", is an assessment of model performance based on the comparison of model outputs with experimental data (discussed in more detail in Section 7).

It is our experience and conclusion from a comprehensive review of past model evaluation exercises that it is not profitable to provide a "cookbook" with series of steps that one must accomplish in order to adequately implement a statistical model evaluation. Models are used in a variety of ways, many of which were never anticipated when the model was first developed and made available. Models are often used in situations that, in principle, they are incapable of handling as they are lacking characterization of relevant physical processes. For example, although most operational air quality models provide estimates of ensemble average concentrations, they are typically used to estimate maximum (peak) concentration values (which are extreme values within an ensemble).

Instead of a series of steps, we provide a framework (Section 4) within which one can understand why modeling results and observations differ. We believe that by following the ideas expressed in this framework, one can develop a successful evaluation of any model regardless of whether it is being applied in a manner consistent with its designed physics and modeling assumptions. In Section 5, we summarize those performance measures that are in common usage, and then in Section 7, we discuss concepts that can be employed in developing a statistical model evaluation.

## 2 Terminology

A review of recent evaluation exercises reveals[3] that various characteristics of atmospheric dispersion models have been tested, and often the methods have used application-specific schemes with various performance measures. In fact, in the literature it is possible to find widely diverse definitions concerning topics related to model evaluation. To avoid confusion and misunderstandings, we believe that it would be useful to achieve a harmonization about terminology and its use. The etymology of the terms we use is important for understanding by both scientists and decision-makers. Hence, we have tried not to depart too much from the etymological word meanings (Schlunzen, 1997).

[3] Of the large available number, we list a few for example. **Methods/Review:** U.S. Environmental Protection Agency (1992), Hanna et al. (1993), Poli and Cirillo (1993), Ward (1994), Weil et al. (1997), **Long Range Transport:** Bellasio et al. (1998), Brandt et al. (1998), Carhart et al. (1989), Mosca et al. (1998); **Complex Terrain:** Cox et al. (1998), Desiato (1991), Gronskei et al. (1993), Luhar and Rao (1994), Ross and Fox (1991), Thuillier (1992); **Plume Dispersion:** Brusasca et al. (1989), Carruthers et al. (1999), Hanna and Paine (1989), Hanna and Chang (1993), Olesen (1995), **Regional Grid:** Davis et al. (2000), Dennis (1986), Hanna et al. (1996), Hass et al. (1997), Kumar et al. (1994), **Low Winds/Street Canyons:** Kumar Yadav and Sharan (1996), Lanzani and Tamponi (1995), Okamoto et al. (1999).

**Atmospheric air quality model:** an idealization of atmospheric physics to calculate the magnitude and location of pollutant concentrations. This may take the form of an equation, algorithm, or series of equations/algorithms used to calculate average or time-varying concentration. They may take the form of a deterministic model or a statistical model. The model may involve process descriptions and numerical methods for solution.

**Calibration (or model calibration):** a procedure used to make, at the model development stage, estimates of the parameters of model equations, which best fit the general model structure to a specific observed data set.

**Data assimilation:** a numerical technique, which makes it possible to combine model results and observations in one integrated system, with the purpose of minimizing the discrepancy between model predictions and observations.

**Data quality assessment:** the scientific and statistical evaluation of data to determine if data obtained from environmental data operations are of the right type, quality and quantity to support their intended use.

**Data quality objective:** a range of acceptability for data used in modeling analyses for a specific application.

**Deterministic model:** a model is deterministic when it is assumed that all possible behaviors are determined by the set of equations comprising the model. These models are based on fundamental mathematical descriptions of atmospheric processes, in which effects (i.e., air pollution) are generated by causes (i.e., emissions).

**Diffusion, absolute:** the characterization of the spreading of material released into the atmosphere based on a coordinate system fixed in space.

**Diffusion, relative:** the characterization of the spreading of material released into the atmosphere based on a coordinate system that is relative to some local position of the dispersing material (e.g., center of mass).

**Dispersion:** the combined effects of eddy diffusion and advection (transport).

**Evaluation (or model evaluation):** the overall system of procedures designed to measure the model performance.

**Evaluation objective:** a feature or characteristic which can be defined through an analysis of the observed concentration pattern (e.g., maximum centerline concentration or lateral extent of the average concentration pattern as a function of downwind distance) for which one desires to assess the skill of the models to reproduce.

**Evaluation procedure:** the analytical steps to be taken to compute the value of the evaluation objective from the observed and modeled patterns of concentration values.

**Fate:** the destiny of a chemical or biological pollutant after release into the atmosphere.

**Model intercomparison:** a process where several models, all presumably appropriate for some chosen situations (idealized or real), simultaneously have their performances assessed and compared.

**Performance measures (or statistical comparison metrics):** evaluation tools (quantitative and/or qualitative) like statistical indices and graphical methodologies, used to compare model outputs with observed values.

**Process model:** an idealization of atmospheric physics envisioned as being composed of a series of inter-related processes to calculate the magnitude and location of pollutant concentrations based on fate, transport, and diffusion in the atmosphere. These models most often are deterministic models, but in principle, could attempt to characterize the stochastic process effects.

**Quality assurance:** all those planned and systematic actions necessary to provide adequate confidence that a product or service will satisfy given requirements for quality.

**Sensitivity analysis:** a process for identifying the magnitude, direction, and form (e.g., linear or non-linear) of the effect of the variation of one or more model parameters or model inputs on the model result.

**Statistical model:** a model of a stochastic process that represents the dependence of successive or neighboring events in response to variation in an external influence on the process. These models are parsimonious using the fewest number of parameters capable of explaining quantitative variation in some observed data. They are based upon semi-empirical statistical relations among available data and measurements.

**Statistical model evaluation:** the analysis of model performance based on the comparison of model outputs with experimental data (evaluation objectives). Statistical model evaluations involve summarizing model performance in an overall sense (typically called performance evaluation), and testing the simulation of specific processes within a model (typically called diagnostic evaluations).

**Stochastic process:** a continuous causal process in time, space, or both, responding to variation in an external influence, and producing a varying series of measured states or events.

**Uncertainty:** a difference (or differences) between what is modeled and what is observed. It is a consequence of a lack of knowledge in model formulation, and errors (or omissions) in data and observations. In principle, uncertainty can be reduced with either improved theory or observations; however, it is generally accepted that there is a limit to how much of the natural variability can be explicitly simulated by models. The portion of natural variability that is beyond the reach of modeling is referred to as inherent variability[4].

**Uncertainty analysis:** a process for estimating model uncertainty.

**Variability:** is what happens in the natural system; the observable variations.

**Verification:** is the checking of the computer code to ensure that it is a true representation of the conceptual model upon which it is based. This includes checking whether the mathematical equations involved have been solved correctly and comparing the numerical solutions with idealized cases for which an analytic solution exists.

## 3 Background

Air quality simulation models have been used for many decades to characterize the transport and diffusion of materials in the atmosphere (Pasquill, 1961; Randerson, 1984; Hanna et al., 1982). The wider use of atmospheric models in scientific studies for regulatory purposes and for describing air quality scenarios requires assessing the degree of reliability of model results. Generally, such an assessment is performed through the comparison of model outputs against field measurements. Tracer experiments are particularly helpful in evaluating the capability of these models to properly simulate transport and diffusion. Comparisons between model outputs and measurements are performed using both qualitative data analysis techniques and quantitative statistical methods.

Up until the early 1980s, comparing modeling results with observations was considered simple. The outputs of dispersion models were plotted against measurements (using traditional scatter plots of the values) and simple performance measures such as the correlation coefficient were computed (Clarke, 1964; Martin, 1971; Hanna, 1971). High correlation values were interpreted as an indication that the model was performing well; low correlation (a not uncommon case) was interpreted to mean that the model was performing poorly.

As air quality models came into more common use, concerns were raised that early statistical performance evaluation had been naive. Little consideration had been given to the consequences and sources of uncertainty and variability. As

[4] What we are labelling "inherent variability" is what Fox (1984) and others discuss as "inherent uncertainty", and what Hanna (1993) discuss as the "irreducible scatter caused by stochastic fluctuations".

discussed in Sections 4 and 6, the sources of uncertainty can be envisioned as being composed of: model formulation uncertainty, representativeness uncertainty, measurement uncertainty, and inherent variability[4].

- Model formulation uncertainty is composed of theory uncertainty (there may be more than one theory that adequately describes available data) and numerical uncertainty (conversion of mathematical algorithms to numerical code may involve approximations that could lead to spurious noise in the solutions if not well-treated);

- Representativeness uncertainty arises whenever there is a lack of agreement in the data used as model input or the data used for comparison with model output[5] to satisfy the spatial and temporal assumptions of the model;

- Measurement uncertainty results from errors in measurements, which can affect model inputs[6] and can affect observed concentrations used for comparison with model outputs;

- Inherent variability arises because models only characterize a portion of the naturally occurring variations.

In the early comparisons, measurement uncertainties were assumed to be small in comparison to the "real world fluctuations", when in fact that was not always a safe assumption. More importantly, even hypothetically error-free measurements possess space and time limitations that prevent them from being good approximations of the time and space assumptions used in the construction of the model. For instance, the comparison of measurements taken at an isolated receptor with grid-averaged model outputs is inappropriate (Davis et al., 2000).

The early statistical performance evaluations failed to address the fact that models provide estimates of ensemble means, whereas the observations are individual realizations from imperfectly defined ensembles (Lamb from Longhetto, 1980; Venkatram, 1988). Furthermore, reliance on linear regressions and correlation coefficient can provide misleading results (Zannetti and Switzer, 1979). Lastly, models rely on emission and meteorological inputs whose uncertainties could justify disagreements between predictions and observations (Irwin et al., 1987).

In the early eighties, several attempts were made to develop standard methodologies for judging air quality model performance (Bornstein and Anderson, 1979; Venkatram, 1982 and 1983; Willmot, 1982). The American

[5] Differences from not properly satisfying the model input assumptions are referred to by some as "data representativeness uncertainty", and by other as "input uncertainty"; differences in properly satisfying the model output assumptions are most often referred to as "data representativeness uncertainties".

[6] Uncertainties in emission data may result from measurement or formulation uncertainties since a wrong methodology might have been used for emission estimation.

Meteorological Society sponsored two workshops in an attempt to provide specific guidelines on the use of statistical tools in air quality applications. A summary of their recommendations is provided in two papers by Fox (1981, 1984).

The most interesting comments and recommendations from the above workshops were:

- the concern about the absolute, rather than statistical nature of U.S. air quality standards
- the possibility of computing statistics between measured data values and model predicted values even when these values are not coupled in time and/or in space
- the identification of reducible errors and inherent variability
- the recommendations to decision-makers to educate themselves and accept the challenge of decision making with quantified uncertainty

Following these two workshops, a series of studies were undertaken to continue to investigate the problem of statistically evaluating the performance of air quality models. Interesting methods were proposed at the DOE Model Validation Workshop, October 23-26, 1984, Charleston, South Carolina, and by Alcamo and Bartnicki (1987) and Hanna (1989a). Major operational evaluations of air quality models were sponsored by EPRI (Reynolds et al., 1984; Ruff et al., 1984; Moore et al., 1985; Reynolds et al., 1985).

Further development of the evaluation methodologies proposed in the early eighties was needed, as it was found that the rote application of performance measures, such as those listed in Fox (1981), was incapable of discerning differences in model performance (Smith, 1984). Whereas if the evaluation results were sorted by stability and distance downwind, then differences in modeling skill could be discerned (Irwin and Smith, 1984). It was becoming increasingly evident that the models were characterizing only a small portion of the observed variations in the concentration values (Hanna, 1988). To better deduce the statistical significance of differences seen in model performance in the face of small sample sizes and unknown uncertainties, investigators began to explore the use of bootstrap techniques (Hanna, 1989). By the late 1980s, most of the model evaluations involved the use of bootstrap techniques in comparing the maximum values of modeled and observed cumulative frequency distributions of the concentration values (Cox and Tikvart, 1990).

Even though the procedures and measures are still evolving to describe performance of models that characterize atmospheric fate, transport and diffusion (Weil et al., 1992; Dekker et al., 1990; Cole and Wicks, 1995), there has been a general acceptance for a need to address the large uncertainties inherent in atmospheric processes. There has also been a consensus reached on the philosophical reasons that models of earth science processes can never be verified

(in the sense of claiming that a model is truly representative of natural processes). General empirical proposition about the natural world cannot be certain since there will always remain the prospect that future observations may call the theory in question (Oreskes et al., 1994). It is seen that numerical models of air pollution are a form of a highly complex scientific hypothesis concerning natural processes that can be confirmed through comparison with observations, but never verified.

## 4 Framework

To set the context for the following discussion (Irwin, 2000), it is important to realize that most of the model evaluation results currently available in the literature are for applied air quality models that use ensemble average characterizations of the transport and diffusion, the chemical transformations, and the physical removal processes. Thus, these applied air quality models only provide a description of the average fate of pollutants to be associated with each possible ensemble of conditions (or "regime"). Natural variability that is not characterized by the model can result in large deviations when comparing individual observations (which are individual realizations from an ensemble of realizations) with modeling results (which are characterizing the ensemble average result).

The differences seen in comparison between model predictions and observations of atmospheric air concentrations may largely reflect inherent variability. Likely, this component of variability is inherent in that it cannot be simulated explicitly by improving the physics of the air quality models. At best, air quality models provide an unbiased estimate of the average concentration expected over all realizations of an ensemble. An estimate of an ensemble can be developed from a set of experiments having fixed external conditions (Lumley and Panofsky, 1964). To accomplish this, the available concentration values are sorted into classes characterizing ensembles. For each of the ensembles thus formed, the difference between the ensemble average and any observed realization (experimental observation) is then ascribed to inherent variability where its variance, $\sigma_n^2$, can be expressed as (Venkatram, 1988):

$$\sigma_n^2 = \overline{\left(C^o - \overline{C^o}\right)^2} \tag{1}$$

where $C^o$ is the observed concentration seen within a realization; the over-bars refer to an average over all available realizations within a given ensemble, so that $\overline{C^o}$ is the estimated ensemble average. In (1), the ensemble refers to the ideal infinite population of all possible realizations meeting the (fixed) characteristics of the chosen ensemble. In practice, we will only have a small sample from this ensemble. Measurement uncertainty in $C^o$ in most tracer experiments is typically

a small fraction of the measurement threshold, and when this is true, its contribution to $\sigma_n$ can usually be deemed negligible.

Defining the characteristics of the ensemble in (1) using the model's input values, $\alpha$, one can view the observed concentrations as:

$$C^o = C^o(\alpha,\beta) = \overline{C^o}(\alpha) + c(\Delta c) + c(\alpha,\beta) \tag{2}$$

where $\beta$ are the variables needed to describe the unresolved transport, fate and diffusion processes. The over-bar represents an average over all possible values of $\beta$ for the specified set of model input parameters $\alpha$; $c(\Delta c)$ represents the effects of concentration representativeness and measurement uncertainty, and $c(\alpha,\beta)$ represents ignorance in $\beta$, unresolved deterministic processes and stochastic fluctuations (Hanna, 1988; Venkatram, 1988). Since $\overline{C^o}(\alpha)$ is an average over all $\beta$, it is only a function of $\alpha$, and in this context, $\overline{C^o}(\alpha)$ represents the ensemble average that the model is ideally attempting to characterize.

The modeled concentrations, $C^s$, can be envisioned as:

$$C^s = C^s(\alpha) = \overline{C^o}(\alpha) + d(\Delta\alpha) + f(\alpha) \tag{3}$$

where $d(\Delta\alpha)$ represents the effects of uncertainty in specifying the model inputs and $f(\alpha)$ represents the effects of uncertainty in the model theory and numerical implementation.

The method we propose for performing an evaluation of modeling skill is separately averaging the observations and modeling results over a series of non-overlapping limited-ranges of $\alpha$, which are called "regimes". Averaging the observations provides an empirical estimate of what most of the current models are attempting to simulate, $\overline{C^o}(\alpha)$. A comparison of the respective observed and modeled averages over a series of $\alpha$-groups provides an empirical estimate of the combined deterministic error associated with input uncertainty and formulation errors.

Given this framework, designing a model evaluation can be envisioned as a two-step process. Step one, we analyze the observations to provide average patterns for comparison with modeled patterns. Step two, given the uncertainties in estimating the average patterns, we test to see whether differences seen in a comparison of performance of several models are statistically significant. In order to place confidence bounds on conclusions reached in step two, bootstrap resampling is recommended (see Section 5.4). Within the American Society for

Testing and Materials (ASTM), a standard guide[7] has been developed that outlines this strategy for designing statistical evaluations of dispersion model performance (a statistical evaluation of performance).

This process is not without problems, as grouping data together for analysis requires large data sets of which there are few. Sorting the data into groups requires sufficient knowledge of the experimental conditions to determine how the data collected on different days or during different time periods should be grouped together. In reality, the external forcing conditions are imperfectly known, and hence the groups are imperfectly composed.

Another problem is that air quality models only explain a small portion of the observed variations, and there are large uncertainties involved in any air quality modeling assessment. Earlier in this chapter, we mentioned that there are essentially four sources of uncertainty: formulation uncertainty, representativeness uncertainty, measurement uncertainty, and inherent variability. We now take a moment to provide some perspective as to the size and nature of inherent variability and model input uncertainty.

From Equation (2), we see that natural variation that is not explained by the model is the term $c(\alpha,\beta)$, and we have referred to this as the inherent variability. It has been estimated that the portion of natural variability that is not accounted for by atmospheric transport and diffusion models is of order of the magnitude of the regime averages (Weil et al., 1992; Hanna, 1993). Thus, small sample sizes in the groups, which are used in the statistical evaluation to form pseudo-ensembles, could lead to large uncertainties in the estimates of the ensemble averages.

An illustration of unexplained concentration variability is presented in Figure 1. Project Prairie Grass (Barad, 1958; Haugen, 1959) is a classic tracer dispersion experiment where sulfur-dioxide ($SO_2$) was released from a small tube placed 46 cm above the ground. Seventy 20-minutes releases were conducted during July and August 1956, in a wheat field near O'Neil, Nebraska. Sampling arcs were positioned on semicircles centered on the release, at downwind distances of 50, 100, 200, 400 and 800 m. The samplers were positioned 1.5 m above the ground, and provided 10-minute concentration values. For the purpose of illustrating concentration variability, two small ensembles of six experiments along the 400-m arc have been grouped together in Figure 1 using the inverse of Monin-Obukhov length, $L$, a stability parameter (as $1/L$ approaches zero, the surface layer of the atmosphere approaches neutral stability conditions). Concentration values from near-surface point sources are inversely proportional to the transport wind speed, $U$, and directly proportional to the emission rate, $Q$. To group the results of the six experiments together, the concentration values were normalized by multiplying the concentration values by $U/Q$, where $U$ was defined as the

[7] Standard Guide for the Statistical Evaluation of Atmospheric Dispersion Model Performance, D6589, Annual Book of Standards Volume 11.03, American Society for Testing and Materials, West Conshohocken, PA 19428 (http://www.astm.org).

value observed at 8 m above the ground. The solid line shown for each group is a Gaussian fit to the results for the six experiments in the group. The scatter of the normalized concentration values about this Gaussian fit can be statistically analyzed to provide an estimate of the concentration variability not characterized by the Gaussian fit. From this analysis and other tracer studies, the stochastic fluctuations (inherent variability) were investigated by analyzing the distribution of $C^o/\overline{C^o}$ for centerline concentration values. The distribution was found to be approximately lognormal, with a standard geometric deviation of order 1.5 to 2 (Irwin and Lee, 1997; Irwin, 1999). These results suggest that centerline concentration values from individual experiments may typically deviate from the ensemble average maximum by as much as a factor of two.

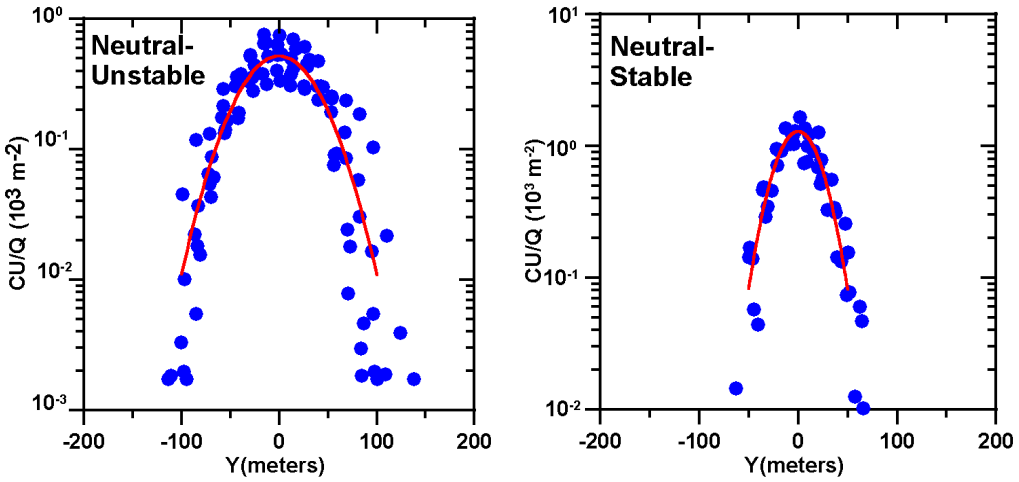


**Figure 1. Near-neutral unstable (left) and near-neutral stable (right) normalized concentration values at the 400-meter arc. The neutral-unstable experiments are 6, 11, 34, 45, 48 and 57, with Monin-Obukhov lengths ranging from -263 m to -82 m. The neutral-stable experiments are 21, 22, 23, 24, 42, and 55, with Monin-Obukhov lengths ranging from 164 m to 359 m.**

The use of wind tunnel measurements can be useful towards providing data for a model evaluation process (Schatzmann and Leitl, 1999). The work of Stein and Wyngaard (2000) investigates the relationship between inherent variability in laboratory and atmospheric boundary layer flows. For a given averaging time, they show that the inherent variability in laboratory flows is smaller than in the atmospheric boundary layer flows under the same stability and statistical conditions.

Characterizing the model input is another source of uncertainty. The variance in modeled concentration values due to input uncertainty can be quite large. Using a Gaussian plume model, Irwin et al. (1987) investigated the uncertainty in estimating the hourly maximum concentration from elevated buoyant sources during unstable atmospheric conditions due to model input uncertainties. A numerical uncertainty analysis was performed using the Monte-Carlo technique to propagate the uncertainties associated with the model input. Uncertainties were

assumed to exist in four model input parameters: wind speed, standard deviation of lateral wind direction fluctuations, standard deviation of vertical wind direction fluctuations, and plume rise. It was concluded that the uncertainty in the maximum concentration estimates is approximately double the uncertainty assumed in the model input. For instance, if half of the input values are within 30% of their error-free values, then half of the estimated maximum concentration values will be within 60% of their error-free values. Using a photochemical grid model, Hanna et al. (1998) investigated the uncertainty in estimating domain-wide hourly maximum ozone concentration values near New York City for July 7-8, 1988. Fifty Monte-Carlo runs were made in which the emissions, chemical initial conditions, meteorological input and chemical reaction rates were varied within expected ranges of uncertainty. The amount of uncertainty varied, depending on the variable. Those variables with the least assumed uncertainty (most of the meteorological inputs) were assumed to be within 30% of their error-free values 95% of the time. Larger uncertainties were generally assumed for the emissions and reaction rates. They found that the domain-wide maximum hourly averaged ozone ranged from 176 to 331 ppb (almost a factor of two range). These two investigations reveal that the sensitivity to model input uncertainties is quite large, regardless of whether the model is a Gaussian plume model, a photochemical grid model, or whether the specie being modeled is inert or chemically reactive.

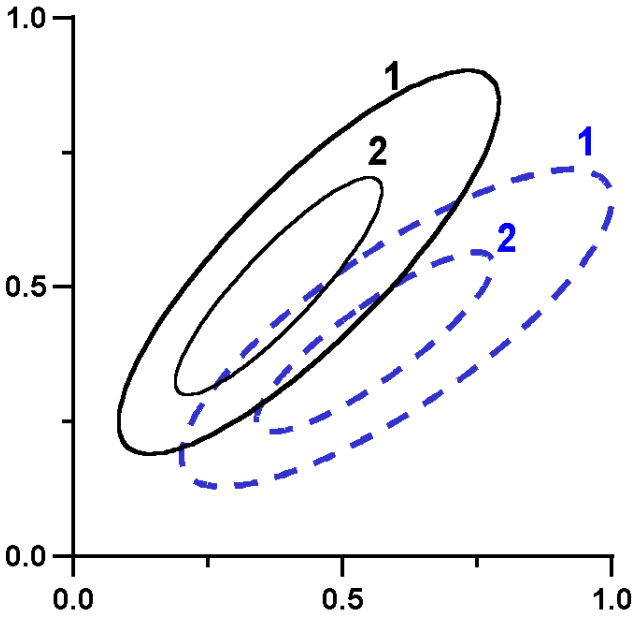


**Figure 2. Illustration of displacement of observed (solid lines) and predicted (dashed lines) ground-level concentration patterns. Isopleths represent points with the same concentration. The point-by-point correlation is poor, but the patterns are clearly similar (adapted from Hanna, 1988 [Reprinted with permission from the Air Pollution Control Association]).**

Irwin and Smith (1984) warned that the disagreement between the indicated wind direction and the actual direction of the path of a plume from an isolated point source is a major cause for disagreement between model predictions and observations. As a plume is transported downwind, it typically expands at an angle of approximately 10 degrees, and this angle is seldom larger than 20 degrees. With such narrow plumes, even a 2-degree error in estimating the plume transport direction can cause very large disagreement between modeled and observed surface concentration values. Weil et al. (1992) analyzed nine periods from the EPRI Kincaid experiments, where each period was about 4 hours long. They concluded that for short travel times (where the growth rate of the plume's width is nearly linear with travel time), the uncertainty in the plume transport direction is of the order of 1/4 of the plume's total width. Farther downwind, where the growth rate of the plume's width is less rapid, the uncertainty in the plume transport direction is larger than 1/4 of the plume's total width. Figure 2 illustrates that any point-to-point comparison of modeled and observed concentration values (e.g., correlation, bias, mean squared error) would suggest poor performance. It is clearly seen that the basic pattern is modeled well if the observed pattern is shifted over to better correspond with that modeled. In conclusion, the uncertainties of plume transport direction are substantial, and likely will preclude, especially for isolated source comparisons, a meaningful evaluation of modeled and observed concentration values paired in time and space.

This section presented a framework providing a means for understanding why modeled and observed values differ. The observations are envisioned as being composed of an ensemble mean about which there are deviations either resulting from representativeness and measurement uncertainty, or uncharacterized natural variability, $c(\alpha,\beta)$. Examples were provided that suggest that for maximum surface concentrations, uncharacterized variability, $c(\alpha,\beta)$, is on the order of the ensemble mean, (i.e., could easily account for factor of two deviations from the mean). The model values are envisioned as being composed of an ensemble mean about which there are deviations either resulting from input (representativeness and measurement) uncertainty, or model theory and numerical implementation errors. Examples were provided that suggest that the effects of input uncertainty can be amplified within the model (e.g., doubled), and can lead to variations on the order of a factor of two. As a pragmatic mean for assessing systematic errors in the model formulations, it was recommended that pseudo-ensembles be formed by grouping evaluation data into time periods where conditions can be assumed to be similar. Then it was recommended that comparisons be made of the group averages, because this insulates the comparisons from many of the sources of uncertainties. Other issues will be addressed, such as how to cope with uncertainties in the direction of transport, as discussed in Section 7. First, we will discuss in the next section the kinds of performance measures one might choose in developing an evaluation procedure.

# 5 Performance Measures

The preceding section described a philosophical framework for understanding why observations differ from model simulation results. This section provides definitions of the performance measures and methods often employed in current evaluations of air quality models. Proper model evaluation involves the application of both statistical indices and graphical methodologies. The list of possible performance measures is extensive (e.g., Fox, 1981), but it has been illustrated that a few well-chosen simple-to-understand performance measures can provide adequate characterization of a model's performance (e.g., Hanna, 1988). Therefore, the selection of performance measures to compare model outputs against observed values is a fundamental step. Statistical indices and graphical methodologies emphasize specific model characteristics (e.g., Canepa and Builtjes, 1999); therefore, outlining the characteristics of each performance measure is useful. The following discussion is not meant to be exhaustive. The key is not in how many performance measures are used, but is in the statistical design used when the performance measures are applied (e.g., Irwin and Smith, 1984).

For convenience, we discuss the comparison of the observed and modeled concentration values in the following discussion. In reality, model evaluation can involve comparisons of observed and modeled plume rise, building wake dimensions, etc. Any feature (evaluation objective) that can be deduced from an analysis of the concentration pattern and converted to a numeric value can be substituted for the word "concentrations" in the following discussion.

## 5.1 Basic Performance Measures

**MEAN** of both the observed and simulated concentrations is defined as:

$$MEAN_{observed} = \overline{C^o} = \sum_i \frac{C_i^o}{N} \qquad MEAN_{simulated} = \overline{C^s} = \sum_i \frac{C_i^s}{N} \qquad (4)$$

where N is the total number of the values being averaged, $C_i^o$ ($C_i^s$) is the i$^{th}$ observed (simulated) concentration value. A perfect model would give $MEAN_{observed} = MEAN_{simulated}$. Note, the values being averaged may be for the same time period (an average over a set of receptors), or the values being averaged may be at a fixed receptor location, either relative or absolute (an average over some time period).

**SIGMA (**standard deviation) of both the observed and simulated concentrations is defined as:

$$SIGMA_{observed} = \sigma^o = \sqrt{\frac{\sum_i (C_i^o - \overline{C}^o)^2}{N}} \tag{5}$$

$$SIGMA_{simulated} = \sigma^s = \sqrt{\frac{\sum_i (C_i^s - \overline{C}^s)^2}{N}} \tag{6}$$

a perfect model would give $SIGMA_{observed} = SIGMA_{simulated}$.

## 5.2 Description of Some Paired Performance Measures

Often the evaluation procedure involves a comparison that logically involves pairing of the observed and modeled values. This might be: the maximum concentration over a domain seen for an hour or a day, or the maximum concentration seen on receptor arcs centered on a tracer release location, etc.

It is not possible to assume that uncertainty in both the observations and the modeled values is small in comparison to the variations seen in their mean values (Irwin et al., 1987; Weil et al., 1992; Hanna, 1993; Hanna et al., 1998). Unless the uncertainties are small in comparison to the variations in their mean values, one cannot confidently make comparisons of raw observations with modeled values.

However, the paired comparison of group averages is meaningful, especially if the groups are well formulated and provide representative estimates of the ensemble average concentration for each group.

**BIAS** is defined as:

$$BIAS = \overline{C^s} - \overline{C^o} \tag{7}$$

A perfect model would give BIAS = 0. If BIAS > 0 (< 0), the model on average overestimates (underestimates) the observed concentrations. We have followed here and elsewhere the convention that a positive BIAS indicates a model over-prediction. This has been found to be better understood by decision-makers and users of model evaluation results (whereas, having to explain a negative BIAS as a model over-prediction was a constant problem). We mention this because one may find in some literature that the opposite convention is sometimes used.

**FB** (Fractional Bias) is defined as:

$$FB = \frac{\overline{C^s} - \overline{C^o}}{(\overline{C^s} + \overline{C^o})/2} \tag{8}$$

FB ranges between – 2 and + 2. For a perfect model, FB = 0. If FB > 0 (< 0), the model on average overestimates (underestimates) the observed concentration values.

The **MEAN**, **BIAS** and **FB** only characterize the "on average" model behavior. One can directly judge the average model performance by looking at the $MEAN_{observed}$ and $MEAN_{simulated}$ values simultaneously. From the value of the BIAS, one has an idea of whether the model underestimates (BIAS < 0) or overestimates (BIAS > 0) the observed values. However, the BIAS value does not convey any sense of how large the average difference is relative to the average magnitude of the observed values. For example, if one is dealing with two data sets, characterized by $\overline{C_A^o} = 10$ and $\overline{C_B^o} = 100$ (in appropriate units), and using a model obtains $\overline{C_A^s} = 20$ and $\overline{C_B^s} = 110$, the BIAS value, in both cases, is 10. However, the "on average" behavior of the model is better in case B, because the percentage difference is less in case B. To address this issue, the FB can be helpful. The FB is the BIAS normalized by the average value of $\overline{C^o}$ and $\overline{C^s}$. As far as the previous example is concerned, $FB_A = 0.67$ and $FB_B = 0.095$. Thus, better model performance is evident in case B.

**FS** (Fractional Standard deviation) is defined as:

$$FS = \frac{\sigma^s - \sigma^o}{(\sigma^s + \sigma^o)/2} \tag{9}$$

FS ranges between – 2 and + 2. For a perfect model, FS = 0. If FS > 0 (< 0), the spreading of the simulated concentration values is larger (smaller) than the spreading of the observed concentration values.

The **SIGMA** and **FS** provide information about the spread (variance) in the modeled and observed concentration values. One can directly judge the model performance looking at the $SIGMA_{observed}$ and $SIGMA_{simulated}$ values simultaneously. The FS index is analogous to the FB index, except it is only related to the relative difference in the variances.

**COR** (linear CORrelation coefficient) is defined as:

$$COR = \frac{\overline{\left(C^o - \overline{C^o}\right)\left(C^s - \overline{C^s}\right)}}{\sigma^o \sigma^s} \tag{10}$$

which ranges between – 1 and + 1 and a perfect model would give COR = + 1.

**COR** provides information on the strength of the linear correlation between the modeled and the observed concentration values. For a value of + 1, the so-called "complete positive correlation", there is correspondence between all pairs of modeled and observed concentration values ($C_i^o, C_i^s$). If the values were plotted against one another in a scatter diagram, all points would lay along a straight line with positive slope. The "complete negative correlation" corresponds to all the pairs on a straight line with negative slope, and COR = – 1. A value of COR near zero indicates the absence of linear correlation between the variables. A model will have COR = + 1 if $C_i^o - \overline{C^o} = C_i^s - \overline{C^s}$ for any ($C_i^o, C_i^s$). Because it is possible that $\overline{C^o} \neq \overline{C^s}$, the previous equality does not mean $C_i^o = C_i^s$ for any ($C_i^o, C_i^s$) as we should expect for a perfect model. Furthermore, it should also be pointed out that a high correlation coefficient does not necessarily indicate a direct dependence between the variables. Two variables may have no true relationship to one another, but may be correlated to a third variable ("spurious correlation").

**FA2** (fraction within a FActor of 2) is defined as:

$$\textit{fraction of data with } 0.5 \leq \frac{C_i^s}{C_i^o} \leq 2 \tag{11}$$

A perfect model would give FA2 = 1.

**NMSE** (Normalized Mean Square Error) is defined[8]:

$$NMSE = \frac{\overline{\left(C^s - C^o\right)^2}}{\overline{C^s}\ \overline{C^o}} \text{ or, if for every i, } C_i^o \neq 0 \text{ then, } NMSE = \frac{\sum_i s_i^2 (1 - k_i)^2}{\sum_i s_i k_i} \tag{12}$$

where $k_i = C_i^s / C_i^o$ and $s_i = C_i^o / \overline{C^o}$; a perfect model would give NMSE = 0. The value of this index is always positive.

**WNNR** (Weighted Normalized mean square error of the Normalized Ratios) is defined as:

---

[8] RMSE (Root Mean Square Error) is defined:

$RMSE = \overline{(C^s - C^o)^2}$. A perfect model would give a RMSE = 0, and the value of this index is always positive. Note, preference is given to using the RMSE rather than the NMSE when there are large uncertainties in $\overline{C^o}$ (which typically occurs when the observed concentration values are close to the measurement threshold).

$$WNNR = \frac{\sum_i s_i^2 \left(1 - \hat{k}_i\right)^2}{\sum_i s_i \hat{k}_i} \tag{13}$$

where $\hat{k}_i = 1/k_i$ (if $k_i > 1$) and $\hat{k}_i = k_i$ (if $k_i \leq 1$). A perfect model would give WNNR = 0. The value of this index is always positive.

**NNR** (Normalized mean square error of the distribution of Normalized Ratios) is defined as:

$$NNR = \frac{\sum_i \left(1 - \hat{k}_i\right)^2}{\sum_i \hat{k}_i} \tag{14}$$

A perfect model would give NNR = 0. The value of this index is always positive.

The **FA2**, **NMSE**, **WNNR** and **NNR** indices give information about the ratios between simulated and measured concentrations. Only the FA2 and NNR indices, out of all indices considered, depend solely on the ratios between simulated and measured concentrations, and not on the data set itself, so they are the only indices strictly usable to compare simulations of different experiments. NMSE attributes more weight to model errors concerning the estimates of the highest measured concentrations in some cases, of the lowest ones in other cases; WNNR attributes more weight to model errors concerning the estimates of the highest measured concentrations; and NNR attributes the same weight to model errors independently of the position of the data within the concentration range (Poli and Cirillo, 1993; Canepa and Modesti, 1997).

**SCATTER DIAGRAM**, **FOEX**, and **FAα** again give information about the ratios between simulated and measured concentrations. SCATTER DIAGRAM is a graph where predicted values are plotted versus measured ones (see Figure 3). The $y = x$ line represents the perfect agreement between predictions and measured values. A value above (below) the $y = x$ line indicates a situation of over-prediction (under-prediction). FOEX is defined as

$$FOEX = \left[ \frac{N_{\left(C_i^s > C_i^o\right)}}{N} - 0.5 \right] \cdot 100 \tag{15}$$

where $N_{\left(C_i^s > C_i^o\right)}$ is the number of over-predictions (i.e. the number of pairs where $C_i^s > C_i^o$). It ranges between – 50% and + 50%. If FOEX = – 50%, all the points are below the $y = x$ line and if FOEX = + 50%, all the points are above the $y = x$ line. The best value is 0%, which means that there are half under-predictions and half over-predictions. FOEX does not take into account the

magnitude of the over-predictions; it evaluates only the number of events of over-prediction. Representing the scatter diagram on logarithmic paper, the FA$\alpha$ band is the region between the two lines of equation $y - y_0 = (x - x_0) \pm \ln(\alpha)$, where $x_0$ and $y_0$ are the coordinates of the origin of the axes. If $\alpha = 2$, then FA$\alpha$ = FA2 (see above).

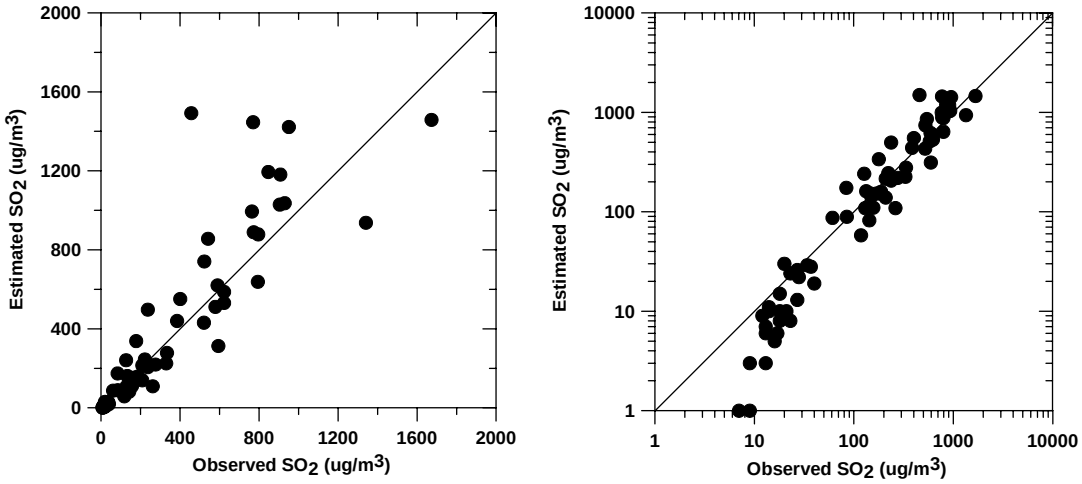


**Figure 3. Example of SCATTER DIAGRAM: linear axes (left), log axes (right).**

**PERCENTILES** and **BOX PLOT** give information about the cumulative probability. The $n^{th}$ percentile of a distribution of values is defined as the cumulative probability in percent, that is, the value that bounds the n% of values below and the (100 – n)% above it. Looking at the box plot (see Figure 4), the general features of the distribution of the considered values can be distinguished.

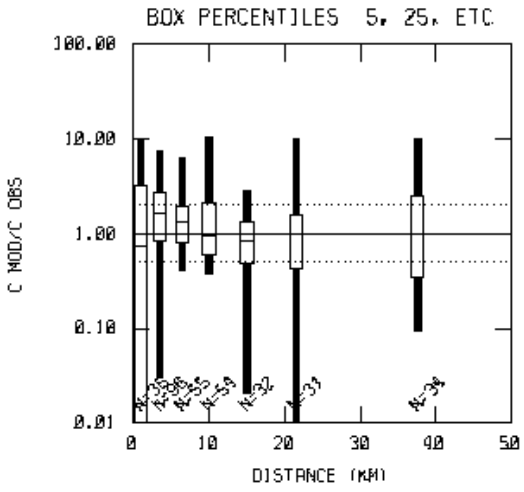


**Figure 4. Example of $C_i^s / C_i^o$ BOX PLOTS stratified with respect to the distance from the source.**

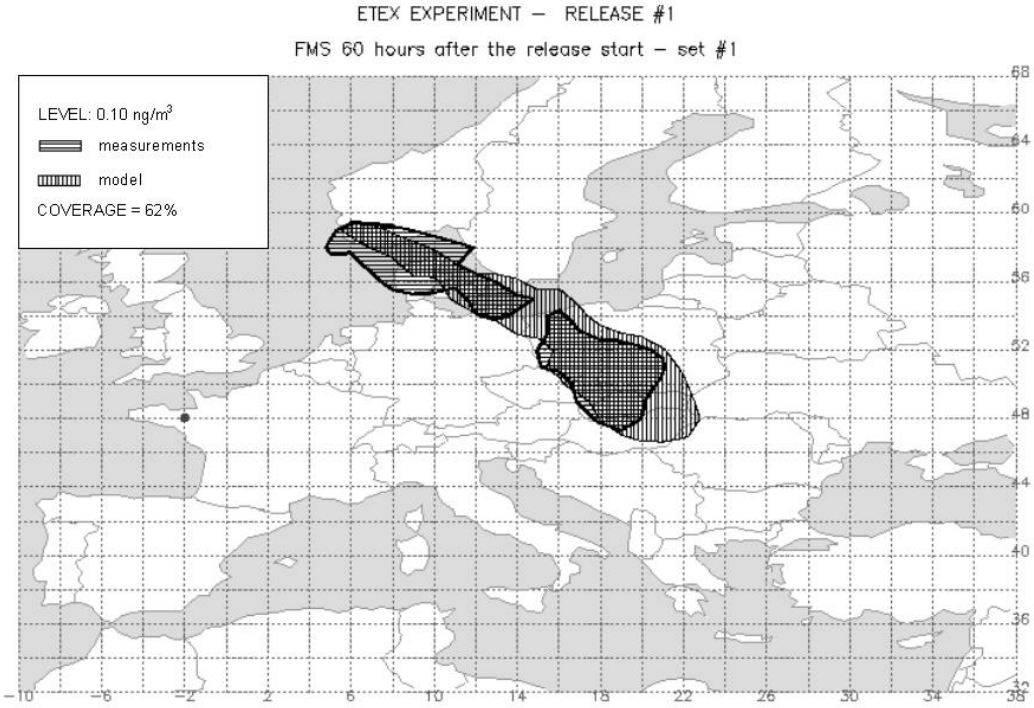


**Figure 5. Example of FMS [from Graziani et al. (1998), courtesy of EI/JRC].**

**FMS** (Figure of Merit in Space) gives information about the space analysis (see Figure 5), and is defined as

$$FMS = 100\frac{A_1 \cap A_2}{A_1 \cup A_2} \tag{16}$$

FMS is calculated at a fixed time for a fixed concentration level (significant level). FMS is the percentage of overlap between the measured ($A_1$) and predicted ($A_2$) areas. A shift in space of the concentration patterns can reduce the FMS significantly (e.g., Figure 2).

**FMT** (Figure of Merit in Time) gives information about the time analysis, and is defined as

$$FMT_{\bar{x}} = 100\frac{\sum_j \min\left\{C_{\bar{x}}^o(t_j), C_{\bar{x}}^s(t_j)\right\}}{\sum_j \max\left\{C_{\bar{x}}^o(t_j), C_{\bar{x}}^s(t_j)\right\}} \tag{17}$$

FMT is calculated at a fixed location $\bar{x}$ for a time series of data. FMT evaluates the overlap of the observed and predicted concentration patterns in time. A temporal shift of the time series can reduce the FMT significantly.

## 5.3 Description of Some Unpaired Performance Measures

An evaluation procedure often involves a comparison that logically involves unpairing of the observed and modeled values. An underlying assumption here is that we have two samples, presumably drawn from the same distribution. If the samples are representative and from the same distribution, then they should both have similar distributions. We have shown in Equations (2) and (3) that the observed and modeled concentration values have different sources of variance, and thus are not from the same distribution. However, if we have groups that are well-formulated and provide representative samples for a series of ensembles, then we can anticipate that the observed and modeled group averages are from the same underlying distribution, and hence have similar frequency distributions.

The **QUANTILE-QUANTILE PLOT** is constructed by plotting the ranked concentration values against one another (e.g., highest concentration observed versus the highest concentration modeled, etc.; see Figure 6). If the observed and modeled concentration frequency distributions are similar, then the plotted values will lie along the 1:1 line on the plot. By visual inspection, one can easily see if the respective distributions are similar, and whether the observed and modeled concentration maximum values are similar.

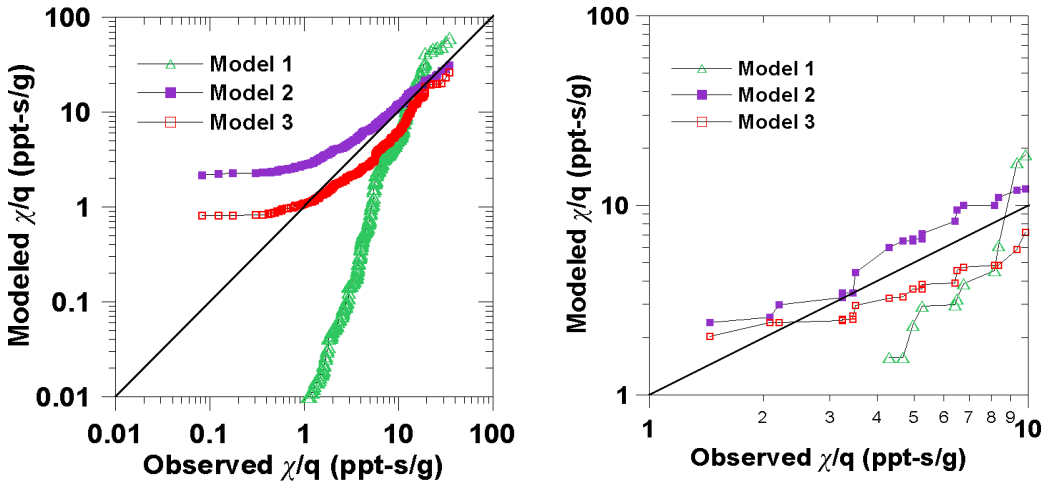


**Figure 6. Example of QUANTILE-QUANTILE PLOTS comparing: on the left side, observed and modeled centerline concentration values (not recommended); on the right side, observed and modeled regime average centerline concentration values (as recommended by the ASTM guide cited in Section 4).**

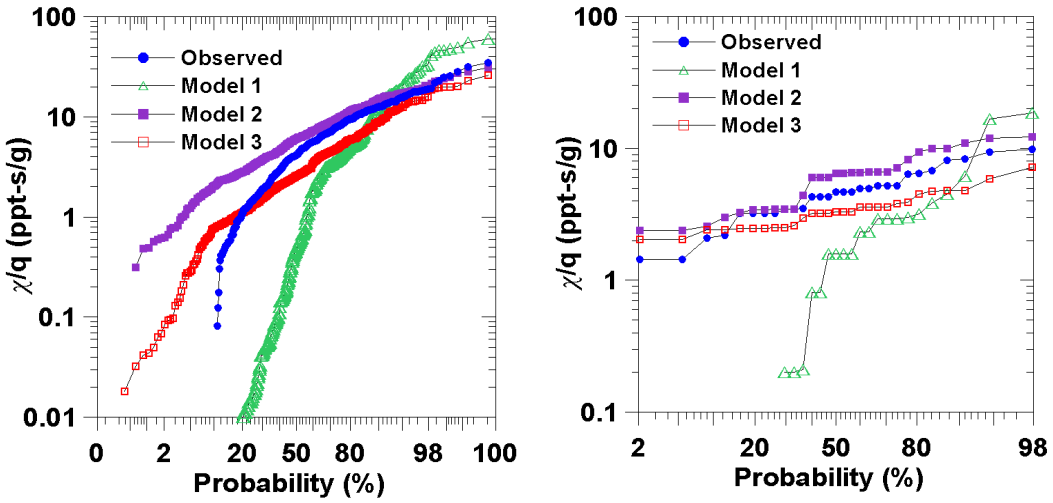


**Figure 7. Example of CUMULATIVE FREQUENCY PLOTS comparing: on the left side, observed and modeled centerline concentration values (not recommended); on the right side, observed and modeled regime average centerline concentration values (as recommended by the ASTM D6589 cited in Section 4).**

The **CUMULATIVE FREQUENCY PLOT** (Figure 7) is constructed by plotting the ranked concentration values (lowest to highest) against the plotting position frequency, $f$ (typically in percent), where $p$ is the rank (1 = lowest), $N$ is the number of values and $f$ is defined as (Larsen, 1969):

$$f = 100\% - 100\%(N - p + 0.6)/N \quad \text{for} \quad p > N/2$$
$$f = 100\%(p - 0.4)/N \qquad \text{for p} < N/2 \tag{18}$$

As with the QUANTILE-QUANTILE PLOT, a visual inspection of the respective CUMULATIVE FREQUENCY DISTRIBUTION PLOTS (observed and modeled) is usually sufficient to suggest whether the two distributions are similar, and whether there is a bias in the model to over- or under-estimate the maximum concentration values observed.

The **RHC** (Robust Highest Concentration) index is often used where comparisons are being made of the maximum concentration values, and is envisioned as a more robust statistical test than direct comparison of maximum values. The RHC is based on an exponential fit to the highest *R - 1* values of the cumulative frequency distribution, where *R* is typically set to 26 for frequency distributions involving a year's worth of values (averaging times of 24 hours or less) (Cox and Tikvart, 1990). The RHC is computed as:

$$RHC = C(R) + \Theta \cdot \ln\left(\frac{3R-1}{2}\right) \tag{19}$$

where $\Theta$ is the average of the *R-1* largest values minus *C(R)*, and *C(R)* is the $R^{th}$ largest value. The value of *R* may be set to a lower value when there are fewer values in the distribution to work with; the RHC of the observed and modeled cumulative frequency distributions are often compared using an FB index, see Cox and Tikvart (1990).

## 5.4 Bootstrap Resampling

The standard analytical formulas for confidence intervals on performance measures from statistics textbooks may be inappropriate (Fox, 1984) since air quality data and model performance measures are not necessarily normally-distributed nor can they always be transformed to a normal distribution. The bootstrap resampling procedure (Heidam, 1987; Hanna, 1989; Cox and Tikvart, 1990; Efron and Tibshirani, 1993) was suggested as an alternative method, since it did not depend on the form of the underlying distribution function.

Following the description provided by Efron and Tibshirani (1993), suppose one is analyzing a data set $x_1, x_2, \ldots x_n$, which for convenience is denoted by the vector $x = (x_1, x_2, \ldots x_n)$. A bootstrap sample, $x^* = (x_1^*, x_2^*, \ldots x_n^*)$, is obtained by randomly sampling *n* times with replacement from the original data points $x = (x_1, x_2, \ldots x_n)$. For instance, with $n = 7$ one might obtain $x^* = (x_5, x_7, x_5, x_4, x_7, x_3, x_1)$. From each bootstrap sample, one can compute some statistics *s* (median, average, RHC, etc.). By creating a number of bootstrap samples, *m*, one can compute the mean, $\bar{s}$, and standard deviation, $\sigma_s$, of the statistic of interest. For estimation of standard errors, *m* is typically on the order of 50 to 500.

Often, the bootstrap resampling procedure can be improved by blocking the data into two or more blocks or sets, with each block containing data having similar characteristics. This prevents the possibility of creating an unrealistic bootstrap sample where all the members are the same value (Hanna, 1989).

When performing model evaluations and model intercomparisons, for each hour there are not only the observed concentration values, but also the modeling results from all the models being tested. In such cases, the individual members, $x_i$, in the vector $x = (x_1, x_2, \ldots x_n)$ are vectors themselves, composed of the observed value and its associated modeling results (from all models, if there are more than one). Thus, the selection of the bootstrap sample $x^*$ also includes each model's estimate for this case.

For example, suppose confidence limits are desired on the NMSE calculated from a set of *n* couples ($C^s$, $C^o$), where $C^s$ is the model simulation estimate and $C^o$ is the corresponding observed value. In the bootstrap procedure, a new set of *n* couples ($C^s$, $C^o$) is randomly drawn from the original set. If a given ($C^s$, $C^o$) is

drawn, it is replaced before the next draw is made. Thus, it is possible (but not very probable) that all $n$ "draws" consist of the same couple ($C^S$, $C^O$). For each resample set of size $n$, the NMSE is calculated. If $m$ resamples are drawn, the cumulative distribution function of the $m$ values of NMSE will provide estimates of confidence limits on NMSE. For example, the 95% confidence interval on NMSE will range from 2.5% to 97.5% points on the NMSE distribution.

For assessing differences in model performance, one often wishes to test whether the differences seen in a performance measure computed between Model #1 and the observations (say, $NMSE_1$), is significantly different when compared to that computed for another model (say Model #2, $NMSE_2$) using the same observations. For testing whether the difference between performance measures is significant, the following procedure is recommended. Let each bootstrap sample be denoted $x^{*b}$, where "*" indicates this is a bootstrap sample and "$b$" indicates this is sample $b$ of a series of bootstrap samples (where the total number of bootstrap samples is $B$). From each bootstrap sample, $x^{*b}$, one computes the respective values for $NMSE_1^b$ and $NMSE_2^b$. The difference $\Delta^{*b} = NMSE_1^{*b} - NMSE_2^{*b}$ can then be computed. Once all $B$ samples have been processed, compute from the set of $B$ values of $\Delta^* = (\Delta^{*1}, \Delta^{*2}, \ldots \Delta^{*B})$, the average and standard deviation, $\overline{\Delta}$ and $\sigma_\Delta$. The null hypothesis is that $\overline{\Delta}$ is not equal to zero with a stated level of confidence, $\alpha$, and the $t$-value for use in a Student's $t$-test is:

$$t = \frac{\overline{\Delta}}{\sigma_\Delta} \tag{20}$$

For illustration purposes, assume the level of confidence is 90% ($\alpha = 0.1$). Then for large values of $B$, if the $t$-value from the above equation is larger than Student's $t_{\alpha/2}$ equal to 1.645, it can be concluded with 90% confidence that $\overline{\Delta}$ is not equal to zero, and hence there is a significant difference in the NMSE values for the two models being tested.

## 6 Model Evaluation

Model evaluation is one of the elements of model quality assurance (see Section 8). Model evaluation is itself a system of procedures designed to measure performance (Model Evaluation Group, 1994a, 1994b; U.S. Environmental Protection Agency, 1997). Following Borrego et al. (2001b), model evaluation is composed of: model algorithm verification, sensitivity analysis, uncertainty analysis, statistical model evaluation, and model inter-comparison. Therefore, statistical model evaluation is one of the fundamental steps to achieve model evaluation.

- **Model algorithm verification** is the process of checking the computer code to ensure the code is a true representation of the conceptual model on which it is based. This includes: checking that the mathematical equations involved have been solved correctly, and comparing numerical solutions with idealized cases for which an analytic solution exists ("verification of numerical solutions") to demonstrate that the two match over the particular range of conditions under consideration.

- **Sensitivity analysis** is a process of characterizing the response of a model to changes in input and parameter values. The purpose is to identify the magnitude, direction, and form (e.g., linear or non-linear) of the effect of such variations. Sensitivity tests can be performed with respect to: 1) uncertainty of physics/chemistry model parameters, and 2) uncertainty of emission and meteorological model input data. In either case, one can use two methods: a) systematically vary one or more of the model inputs to determine the effect on the modeling results (Hilst, 1970), or b) perform a Monte-Carlo study with random sampling (Irwin et al., 1987). In traditional sensitivity studies (item a), each input would be varied over a reasonable range likely to be encountered. These studies were routinely performed in the early years of air pollution modeling to develop a better understanding of the performance of plume dispersion models simulating the transport and diffusion of inert pollutants. Monte-Carlo studies (item b) are becoming more common, as they provide a sense of the overall response of the modeling system to known uncertainties throughout the system. They are especially useful for models simulating chemically reactive species where there are strong nonlinear couplings between the model input and the output (Hanna et al, 1998). Results from traditional sensitivity and Monte-Carlo studies provide useful guidance on which inputs should be most carefully prescribed because they account for the greatest sensitivity in the modeling output. Sensitivity analysis can also provide insight into how a model will behave when it is applied to conditions outside of the severely limited supply of available evaluation data.

- **Uncertainty analysis** (Section 4) is a process of estimating the model uncertainty. The total model uncertainty consists of three terms: 1) model formulation uncertainty in theoretical and numerical description of physics/chemistry parameters and processes (possibly systematic or random, assessable using traditional sensitivity or Monte-Carlo studies); 2) representativeness and measurement uncertainty in emission and meteorological model input data (both systematic and random, assessable using traditional sensitivity or Monte-Carlo studies); and 3) inherent variability associated with those physical processes which are not characterized within the model (both systematic and random, requires an extensive observational database and the estimation of ensemble averages). While the first and second contributions can in principle be reduced, it has

been recognized that inherent variability is not reducible, as it relates to the stochastic nature of the turbulent atmospheric motions (Fox, 1984).

- **Statistical model evaluation** (Section 7) is the comparison of model outputs with experimental data. It is also referred to as a statistical performance evaluation (ASTM D6589). It is preferred that the comparison data have not been used to develop the model.

- **Model intercomparison** is a process where several models, all presumably appropriate for some chosen situations (idealized or real), simultaneously have their performance assessed. This is a necessary step if one is to objectively select those models (from a list of possible candidate models), which perform best for some chosen situations. It is becoming increasingly more common for model intercomparisons to involve bootstrap resampling in order to arrive at an objective determination of whether differences seen in performance are statistically significant (see discussion in Section 5.4).

## 7 Statistical Model Evaluation

Statistical model evaluation is the analysis of model performance based on the statistical comparison of the model outputs with the experimental data (evaluation objectives). Although we can recommend specific steps one should accomplish, the details in how these steps are accomplished typically cannot be defined until:

1. the evaluation goal is defined,
2. the model is defined (or models if one is interested in performing model intercomparison), and
3. the databases are defined.

The sequence shown is a natural consequence that one cannot define which models to apply until the goal is defined. For instance, are we testing the performance of models to estimate the maximum concentrations near an industrial facility with one or several tall stacks next to buildings? Are we testing the performance of models to estimate the peak daily ozone concentration near a large city that is downwind of several even larger cities? The evaluation databases to be employed cannot be defined until one knows which model (or models) is selected and the task (objective) to be evaluated. Models require certain inputs, which may limit the usefulness of certain field data. Certain tasks require particular sampling plans (otherwise, one cannot evaluate the model's performance).

## 7.1 Before Evaluation

### 7.1.1 Defining the Evaluation Goal

To statistically assess model performance, one must define an overall evaluation goal or purpose. This will suggest features (evaluation objectives, see Section 7.2.1) within the observed and modeled concentration patterns to be compared (e.g., maximum surface concentrations, lateral extent of a dispersing plume). The selection and definition of evaluation objectives are typically tailored to the model's capabilities and intended uses. The very nature of the problem of characterizing air quality and the way models are applied make it impossible to define one single or absolute evaluation objective suitable for all purposes. The definition of evaluation objectives will be restricted by the limited range of conditions in the available comparison data. A procedure needs to be defined that allows definition of an evaluation objective from available observations of concentration values.

The evaluation goal can be process oriented (diagnostic); in this case one will have to make a selection of the model characteristics/modules to be validated. The evaluation goal may concern the overall model (integrated). It can be episodic or climatological depending on the time scale. The goal should be specific enough that it can be converted into one or more objective comparisons, which allows construction of null hypotheses that can be tested.

The evaluation goal may be to assess the performance of models to characterize what they were intended to characterize, namely ensemble estimates. Alternatively, the goal may be to assess the performance of models to characterize something different from their design capabilities, like maximum values as seen in the observations. There are consequences in choosing the latter, as good correspondence in this case may be indicative of a systematic flaw in a model, rather than a well-performing model. We recommend including, in all model evaluations, an assessment of how well models perform their designed capabilities.

When the intent is to select, among several models, a model able to perform as intended, the goal can be to determine which of several models has the lowest combination of bias and scatter, when modeling results are compared with observed values of the evaluation objectives. For this assessment, we recommend using the **NNR** or the **NMSE** (other performance measures may also provide useful insights). We first define the model having the lowest value for the **NNR** as the base-model. Then to assess the relative skill of the other models, the null hypotheses would be that the **NNR** values computed for the other models are significantly different when compared to that computed for the base-model (see Section 5.4).

### 7.1.2 Understanding Models to be Validated

Another part of statistical model evaluation is in having a fundamental understanding of what a model is capable of estimating (what physical processes are included or excluded from explicit treatment). All models are a compromise in what physical processes are chosen for explicit treatment. If the objective is to estimate the pattern of concentration values in the near vicinity of one (or several) source, then typically chemistry is of little importance. For such situations, the travel times from the sources to the receptor locations of interest are too short for chemistry formation and destruction to greatly affect the results. However, such situations demand that the air quality model properly treat near-field source emission effects such as: building wakes, initial characterization of source release conditions and size, rates of diffusion of pollutants released as they transport downwind, terrain and land use effects on plume transport, etc. If chemistry is to be explicitly treated, then initial source release effects are typically unimportant, as the pollutants are well-mixed over some volume of the atmosphere by the time the chemistry of interest has greatly affected the results. First attempts to treat both near-field dispersion effects and chemistry have been found to be inefficient and slow on today's computers that are available for routine use.

One might ask why more than one model is often involved in a statistical model evaluation exercise. There are several pragmatic reasons. Often, there is already an "accepted" model, and the purpose of statistical model evaluation is to prove whether a candidate model's performance is significantly better than the "accepted" model. Models differ in the characterized physical processes, the sophistication of input data required, and the numerical processor required. If several models can be shown to have statistically similar performance, then one might select from these a model for use that best meets available resources in input data, computer expertise, processing time, etc. Parsimony (economy or simplicity of assumptions) is a desired trait in modeling. As illustrated in Figure 8, as the model formulation increases in complexity (to explicitly treat more physical processes), we increase the number of input variables, which increases the likelihood of degrading the model's performance due to input data and model parameter uncertainty. Underlying the model formulation and input uncertainty, there is the inherent variability that the model does not characterize (represented as the line labeled "noise").

Alternatively, one might select from a group of models having similar performance, a model that is known to handle a specific process (deposition, sulfate chemistry, etc.). For testing certain specific processes, there may be very few databases suitable for use in an evaluation. This is not a desirable situation, but one often faces less evaluation data than is needed. Thus, part of model evaluation is an artful use of sparsely available field data. A corollary is to have a working knowledge of the field data that possibly could be used, and knowing the strengths and weaknesses of each field experiment's data.

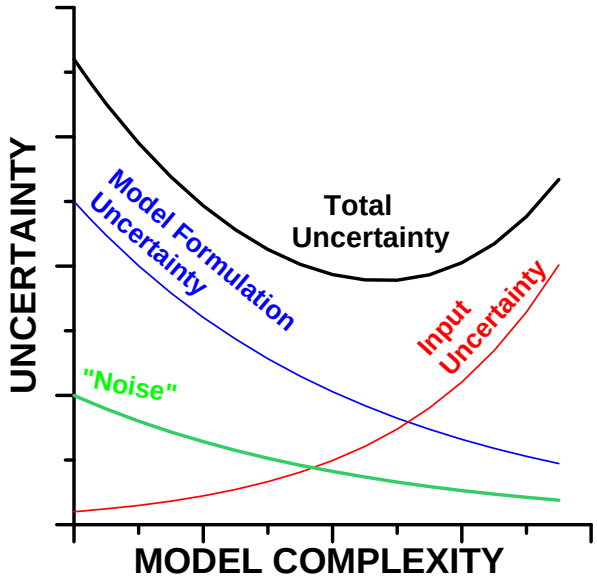


**Figure 8. Illustration of relationship of model formulation uncertainty and input uncertainty, and the combined effect on total model uncertainty. [Adapted from Hanna (1989b). Reprinted with permission from the Butterworth-Heinemann Publishing Company].**

### 7.1.3 Selecting Field Data for Use in the Model Evaluation

Model evaluation is mostly constrained by the amount and quality of available observational data for comparison with modeling results. The simulation models are capable of providing estimates of a larger set of conditions than for which there are observations. Furthermore, most models do not provide estimates of directly measurable quantities. For instance, even if a model provides an estimate of the concentration at a specific location, it is most likely an estimate of an ensemble average result which has an implied averaging time; for grid models, it represents an average over some volume of air (e.g., grid average). Hence, in establishing what abilities of the model are to be tested, one must first consider whether there are sufficient data available that can provide (either directly or through analysis) observations of what is being modeled.

Some fundamental understanding of the sampler limitations (operational range), background concentration values, and stochastic nature of the atmosphere is necessary for developing effective evaluation methodologies. All samplers have a detection threshold; below this threshold, observed values either are not provided, or are considered suspect. It is possible that there is a natural background of the tracer, which either has been subtracted from the observations, or needs to be considered in using the observations. Some samplers have a saturation point that limits the maximum value that can be observed. The user of concentration observations should address these limitations, as needed, in designing the evaluation procedures.

It is often worthwhile to perform a preliminary data set review in order to learn the "structure" of the data, and thereby identify appropriate strengths and limitations within a field experiment (U.S. Environmental Protection Agency QA/G-9, 1998). This review could include calculations of basic statistical quantities - number of observations and data capture, average, median (P50), range, standard deviation, coefficient of variation, P99, and P95 - and graphical representation of the data. The preliminary data set review should include considerations about quality of the data as well (for more details see Section 8.2).

We recommend viewing a model's performance, in relative terms, in comparison to several available models over a variety of circumstances. As new field data becomes available, the selection of the best performing model may change, as the models may be validated for new conditions and in new circumstances. This argues for using a variety of field data sets to provide hope for developing robust conclusions as to which of several models can be currently deemed to perform best.

The following series of steps should be considered in choosing data sets for model evaluation studies:

- select field data sets appropriate for the applications for which the model is to be evaluated, taking the quality of the data into account
- note the model input values that require estimation for the selected data sets
- determine the required levels of temporal detail (e.g., minute-by-minute or hour-by-hour) and spatial detail (e.g., vertical or horizontal variation in the meteorological conditions) for the models to be evaluated, as well as the existence and variations of other sources of the same material within the modeling domain
- ensure that the samplers are sufficiently close to one another and in sufficient numbers for definition of the evaluation objectives
- find or collect appropriate data to estimate the model inputs and to compare with model outputs

## 7.2 Evaluation Strategy

### 7.2.1 Defining Evaluation Procedures

Performing a statistical model evaluation involves defining those evaluation objectives (features or characteristics) within the pattern of observed and modeled concentration values that are of interest to compare. As yet, no single feature or characteristic has been found, that can be defined within a concentration pattern, that can fully test a model's performance. For instance, the maximum surface concentration may appear unbiased through a compensation of errors in estimating the lateral extent of the dispersing material and in estimating the vertical extent of the dispersing material. Considering that other biases may exist (e.g., in treatment of the chemical and removal processes during transport, in

estimating buoyant plume rise, in accounting for wind direction changes with height, in accounting for penetration of material into layers above the current mixing depth, and in systematic variation in all of these biases as a function of atmospheric stability), one can appreciate that there are many ways that a model can falsely give the appearance of good performance.

In principle, modeling diffusion involves characterizing the size and shape of the volume into which the material is dispersing as well as the distribution of the material within this volume. Volumes are three dimensional, so a model evaluation will be more complete if it tests the model's ability to characterize diffusion along more than one of these dimensions. In practice, there are more observations available on the downwind and crosswind concentration profiles of the dispersing material than are available on vertical concentration profiles of the dispersing material.

Developing evaluation objectives involves having a sense of what analytical procedures might be employed. This involves a combination of understanding the modeling assumptions, knowledge of possible comparison measures, and knowledge of the success of previous practices. For example, to assess the performance of the skill of a model to simulate the areal extent of a dispersing puff of tracer emissions from a comparison of isolated measurements with the estimated concentration pattern, Brost (1988) used evaluation objectives and procedures developed for measuring the skill of mesoscale meteorological models to forecast the areal extent of a tropical cyclone from a comparison of isolated pressure measurements to the estimated pressure pattern (Anthes, 1983). In particular, the surface area where concentrations were predicted to be above a certain threshold was compared to the surface area deduced from the available monitoring data. The lesson here is that evaluation objectives and procedures developed in other earth sciences can often be adapted for use in evaluating air dispersion models.

### 7.2.2 Developing Evaluation Procedures

Having selected evaluation objectives for comparison, the next step would be to define an evaluation procedure (or series of procedures), which defines how each evaluation objective will be derived from the available information. Development of statistical model evaluation procedures begins by providing technical definitions of the terminology used in the goal statement. In the following discussion, we use a plume dispersion model example, but, as discussed in Section 7.4, the thought process is also valid for grid models.

For instance, suppose that the evaluation goal is to test the ability of models to replicate the average centerline concentration as a function of transport downwind and as a function of atmospheric stability. The stated goal involves several items which require definition, namely: 1) what is an "average centerline

concentration", 2) what is "transport downwind", and 3) how will "stability" be defined?

What questions arise in defining the average centerline concentration? Given a sampling arc of concentration values, a decision is needed of whether the centerline concentration is the maximum value seen anywhere along the arc, or whether the centerline concentration is that seen near the center of mass of the observed lateral concentration distribution. If one chooses the latter concept, then a definition is needed of how "near" the center of mass one has to be, in order to be representative of a centerline concentration value. One might decide to select all values within a specific range (nearness to the center of mass). In such a case, either a definition or a procedure will be needed to define how this specific range will be determined. A decision will have to be made on the treatment of observed zero (and near measurement threshold) concentrations. Discarding such values is saying that low concentrations cannot occur near a plume's center of mass, which is a dubious assumption. One might test to see if conclusions reached regarding "best performing model" are sensitive to the decision made on the treatment of near-zero concentrations.

What questions arise in defining "transport downwind"? During near-calm wind conditions when transport may have favored more than one direction over the sampling period, "downwind" is not well described by one direction. If plume models are being tested, one might exclude near-calm conditions since plume models are not meant to provide meaningful results during such conditions. If puff models or grid models are being tested, one might sort the near-calm cases into a special regime for analysis.

What questions arise in defining the "stability"? For surface releases, surface-layer Monin-Obukhov length, $L$, has been found to adequately define stability effects, whereas, for elevated releases, $Z_i/L$, where $Z_i$ is the mixing depth, has been found to be a useful parameter for describing stability effects. Each model likely has its own meteorological processor. It is likely that different processors will have different values for $L$ and $Z_i$ for each of the evaluation cases. There is no one best way to deal with this problem. One solution might be sorting the data into regimes using each of the model's input values, and seeing whether or not the previous conclusions as to the best performing model are affected.

What questions arise if one is grouping data together? If one is grouping data together for which the emission rates are different, one might choose to resolve this by normalizing the concentration values by dividing by the respective emission rates. Dividing by the emission rate requires either a constant emission rate over the entire release, or the downwind transport must be sufficiently obvious that one can compute an emission rate based on travel time that is appropriate for each downwind distance.

We discussed earlier the difficulty in properly characterizing the plume transport direction. A decision will have to be made as to how one will compare a feature

(or characteristic) in a concentration pattern, when uncertainties in transport direction are large. Will the observed and modeled patterns be shifted, and if so, in what manner?

Even defining the "observed" pattern is problematic, because one must decide where the "edge" of the pattern occurs. Will the reported concentration be used, even though it is near (or below) the measurement threshold? If one includes for analysis only concentration greater than zero, the testing may favor models that overestimate the extension in space and/or in time of the pollution episode. On the contrary, if the statistic includes all data (including zeros), the performance of a model, that in general underestimates the extension in space and/or in time of the pollution episode, is improved. Furthermore, one can imagine that adding a number of receptor points far from the area of interest of the pollutant, obviously measuring zero concentration, would artificially improve the performance of any model. An approach might consist of including all the points where either measured or simulated values give non-zero concentration. However, this criterion generates different ensembles of selected data that are dependent on each model's results. To try to overcome the outlined difficulties, a filter like that used by Mosca et al. (1998) dealing with the ETEX (European Tracer EXperiment) can be applied. They selected pairs ( $C^o$ , $C^s$ ) showing a non-zero measured concentration that occur not earlier than two time intervals (6 h) before the model predicts the arrival of the cloud, and not later than two time intervals after the model predicts the departure of the cloud.

This discussion is not meant to be exhaustive, but to illustrate how the thought process might evolve. By defining terms, other questions arise and, when resolved, will eventually develop an analysis that will compute the evaluation objective from the available data. There likely is more than one answer to the questions that develop. This may cause different people to develop different objectives and procedures for the same goal. If the same set of models is chosen as the best performing, regardless of which path is chosen, one can likely be assured that the conclusions reached are robust.

## 7.3 Summarizing Evaluation Results

Summarizing model evaluation results usually involves both performance and diagnostic evaluations, and both are needed to establish credibility within the client and scientific community. Performance evaluations allow determination of relative model precision and accuracy in comparison with data and alternative modeling systems. Performance evaluations allow us to answer the question, how well does the model simulate the temporal and spatial patterns seen in the observations, and typically employ large spatial/temporal scale data sets (e.g., large field experiments, national data sets). A performance evaluation might involve a summary of one or more evaluation objectives over all conditions experienced within a particular field experiment. Performance evaluations can be done with or without stratification of the evaluation data into regimes; however,

we have recommended the use of modeled and observed regime averages, as this improves the likelihood of detecting bias in the models' ability to perform as intended. Diagnostic evaluations allow determination of the model precision and accuracy in simulating intermediate processes that affect the final results. Diagnostic evaluations allow us to answer the question, do we get the right answer for the right reason, and usually employ smaller spatial/temporal scale date sets (e.g., field studies). A diagnostic evaluation might involve comparison of observed and modeled values (evaluation objectives) as a function of one or more model input variables, with a focus on a particular process (e.g., plume rise, production of chemical species).

### 7.3.1 Detecting Trends in Modeling Bias

In this discussion, references to observed and modeled values refer to the observed and model evaluation objectives (e.g., regime averages). A plot of the observed and modeled values as a function of one of the model input parameters is a direct means for detecting model bias. Such comparisons have been recommended and employed in a variety of investigations (e.g., Fox 1981; Weil et al., 1992; and Hanna, 1993). In some cases, the comparison is the ratio formed by dividing the modeled value by the observed value, plotted as a function of one or more of the model input parameters. If the data have been stratified into regimes, one can also display the standard error estimates on the respective modeled and observed regime averages. If the respective averages are encompassed by the error bars (typically plus and minus two times the standard error estimates), one can assume that the differences are not significant (Irwin, 1998). As described by Hanna (1988), this is "seductive" inference. A more robust assessment of the significance of the differences would be to use the analysis discussed in Section 5.4.

### 7.3.2 Overall Summary of Performance

As an example of overall summary of performance, we will discuss a procedure constructed using the scheme introduced by Cox and Tikvart (1990) as a template. The design for statistically summarizing model performance over several regimes is envisioned as a five-step procedure.

1. Form a replicate sample using concurrent sampling of the observed and modeled values for each regime. Concurrent sampling associates results from all models with each observed value so that selection of an observed value automatically selects the corresponding estimates by all models.
2. Compute the average of observed and modeled values for each regime.
3. Compute the **NNR** using the computed regime averages, and store the value of the **NNR** computed for this pass of the bootstrap sampling.
4. Repeat steps 1 through 3 for all B bootstrap sampling passes.

5. Implement the procedure described in Section 5.4 to detect: a) which model has the lowest computed **NNR** value (call this the "base" model); b) which models have **NNR** values that are significantly different from the "base" model.

In the Cox and Tikvart (1990) analysis, the data were sorted into regimes (defined in terms of Pasquill stability category and low/high wind speed classes), and bootstrap resampling was used to develop standard error estimates on the comparisons. The performance measure was the **RHC** (computed from the raw observed cumulative frequency distribution), which is a comparison of the highest concentration values (maxima), which most models do not contain the physics to simulate. This procedure can be improved if the performance measure is the **NNR** computed from the modeled and observed regime averages of centerline concentration values.

The data demands are much greater for using regime averages than for using individual concentrations. Procedures that analyze groups (regimes) of data require intensive tracer field studies, with a dense receptor network, and many experiments. Whereas, Cox and Tikvart (1990) devised their analysis to make use of very sparse receptor networks having one or more years of sampling results. With dense receptor networks, attempts can be made to compare average modeled and "observed" centerline concentration values, but there are only a few of these experiments that have sufficient data to allow stratification of the data into regimes for analysis. With sparse receptor networks, there are more data for analysis, but there is insufficient information to define the observed maxima relative to the dispersing plume's center of mass. Thus, there is uncertainty as to whether or not the observed maxima are representative of centerline concentration values. As discussed earlier, observed concentrations for inert gas can easily vary by a factor of two in magnitude about their respective ensemble averages. It is not obvious that the average of the N (say 25) observed maximum hourly concentration values (for a particular distance downwind and narrowly defined stability range) is the ensemble average centerline concentration the model is predicting. In fact, one might anticipate that the average of the N maximum concentration values is likely to be higher than the ensemble average of the centerline concentration. Following the testing procedure outlined by Cox and Tikvart (1990) may favor selection of poorly formed models that routinely underestimate the lateral diffusion (and thereby overestimate the plume centerline concentration). This in turn may bias the performance of such models in their ability to characterize concentration patterns for longer averaging times. We see evaluations, using field data from sparse networks, as a useful extension to further explore the performance of a well-formulated model for other environs and for use of the model for other purposes.

## 7.4 Evaluation of Eulerian Grid Models

For the most part, the preceding discussion and the examples provided were explicitly discussed from the viewpoint that the models being validated were for inert species (e.g., sulfur dioxide, primary emissions of particulate, carbon monoxide, etc.). In addition, the examples were discussed in terms of plume and puff modeling concepts. Evaluation of grid models is not governed by different principles. All of the philosophy and principles discussed in the previous sections apply equally to grid models.

The problems and uncertainties of characterizing the inert pollutant patterns and transport are just as severe for a grid model as for plume or puff models. In recent years, more attention has been given to assessing the performance of Eulerian grid models in characterizing concentrations of primary pollutants. Studies such as Kumar et al. (1994) suggest that large differences are seen when comparisons are made involving primary pollutants. Differences seen in comparisons involving primary pollutants are typically an order of magnitude larger than those seen for reactive (secondary formed) pollutants. The surface concentration values of primary pollutants are typically one of localized maxima or minima, surrounded by strong gradients. The observed pattern is one stochastic realization from some imperfectly defined ensemble. The simulation results are strongly dependent on proper characterization of the emissions, and on the sophistication brought to bear on the analysis and characterization of the time and space varying three-dimensional wind field. To further complicate the problem for grid models, the spatial and temporal characterization of the precursor emissions are highly uncertain (e.g., Hanna et al., 1998), as they are most often deduced from assumptions of land use, activity patterns, traffic flows, etc., rather than on direct measurements of emissions.

Unlike primary pollutants, the spatial pattern for surface concentration values of secondary pollutants (like ozone) is typically a broad flat maximum with weak spatial gradients. Localized areas with strong gradients in ozone concentration are found in the near vicinity of sources emitting large amounts of nitrogen oxides, which can locally deplete the ozone. Given a reasonably good precursor inventory, one would expect the ozone pattern to be well simulated. Sources with large emissions of nitrogen oxide should be easy to identify. As discussed in Hogrefe et al. (2001b), the production of ozone within and downwind of a large urban area correlates with the diurnal course of available sunlight and the precursor emissions are often correlated with the diurnal course of the surface temperatures, so the model estimates are forced somewhat to show good correlation in time.

Photochemical grid models should be validated using extensive and detailed field data (e.g., see Section 7.1.3) to determine if the models adequately represent the ozone processes. However, there are few (if any) field studies that have collected ozone data over an extensive length of time with a reasonably dense network of

receptors. Without such data, the formal statistical approaches where null hypotheses are constructed and tested are of little value.

Photochemical model intercomparison is of interest. In order to perform such intercomparisons, it is necessary to design the base runs for the various models so that they are as comparable as possible (i.e. using the same grid domain, the same emissions files, the same meteorological inputs, and the same initial and boundary conditions) while still preserving the advanced features available in the technical components of each of the models. In any case, if the database is not extensive and detailed, it is difficult to discern significant differences between models (e.g., Hanna et al., 1996).

It is also of interest to determine how well models simulate important variables, such as $NO_2$, VOC and other precursors, at the surface and aloft. Uncertainties involving the initial conditions and boundary conditions should be assessed, and it should be determined whether models perform better with initial and boundary conditions provided by larger scale models, or with values derived from intensive observations. It is of interest to know whether the models respond differently to changes in VOC and $NO_x$ emissions, or whether the predictions are improved by using prognostic rather than diagnostic meteorological model input.

### 7.4.1 The "Threshold" Methods

Traditionally, Tesche et al. (1990) and the U.S. Environmental Protection Agency (1991) have recommended statistical analysis of the residuals to evaluate photochemical models. The final acceptance criteria are arbitrary, requiring the calculated model biases and variances to be within certain bounds or "thresholds". For completeness, we shall review these methods in this section. However, as described by Arnold et al. (1998), recent analyses have shown that "acceptable" performance has been determined using these bias and threshold criteria in spite of the existence of fundamental errors in the model inputs of emissions and meteorology. Currently, an effort is underway to develop a new generation of model evaluation methods for assessing the performance of chemical grid models, and we have summarized some of the methods being examined in Section 7.4.2.

Hanna et al. (1996) photochemical model evaluation exercise was founded on two steps, suggested by Tesche et al. (1990) and U.S. Environmental Protection Agency (1991):

- the first step involved visual inspections of the various contour plots, vertical profiles, and time series, to look for obvious signs of correlation, trends, biases, and scatter
- the second step made use of the average normalized bias

$$\textit{average normalized bias} = \overline{\left(C^s - C^o\right) \Big/ C^o} \tag{21}$$

and average normalized absolute bias

$$average\ normalized\ absolute\ bias = \overline{\left|C^s - C^o\right| \Big/ C^o} \qquad (22)$$

It is worth noting that it is not possible to deal with zero observed concentrations using these indices. U.S. Environmental Protection Agency (1991) recommends the average normalized bias be less than about 10-15%, and the average normalized absolute bias be less than about 30-35%, for data sets in which the daily maximum ozone predictions and observations are paired in time and space.

Hanna et al. (1996) performed a statistical analysis concerning: 1) peak 1 h averaged ozone concentration for a given day anywhere in the domain; 2) 1 h averaged ozone concentrations larger than 60 ppb at all monitors and hours (i.e. paired in space and time).

Following U.S. Environmental Protection Agency (1991), U.S. Environmental Protection Agency (1996) presents a compilation of a series of photochemical model simulations and evaluation exercises conducted within the United States. These evaluations focused on the models' ability to predict the domain-wide peak ozone concentration, and the concentrations at all locations with observed ozone concentrations above 60 ppb. The performance measures used are:

- the normalized accuracy of domain-wide maximum 1-hour concentration unpaired in space and time

$$A_u = 100\left(\frac{C^s{}_{domain-wide\ peak} - C^o{}_{domain-wide\ peak}}{C^o{}_{domain-wide\ peak}}\right) \qquad (23)$$

- mean normalized bias of all simulated and observed concentration pairs with $C_i^o > 60\,\text{ppb}$

$$NBIAS_{60} = \frac{100}{N}\sum_{i=1}^{N}\frac{\left(C_i^s - C_i^o\right)}{C_i^o} \qquad (24)$$

where N includes all the simulated and observed concentration pairs with $C_i^o > 60\,\text{ppb}$

- mean normalized error of all simulated and observed concentration pairs with $C_i^o > 60\,\text{ppb}$

$$NERROR_{60} = \frac{100}{N}\sum_{i=1}^{N}\frac{\left|C_i^s - C_i^o\right|}{C_i^o} \qquad (25)$$

Again following U.S. Environmental Protection Agency (1991), Lurmann and Kumar (1997) and SAIC (1997) presented an evaluation of the ability of the UAM-V model (Morris et al., 1993) to estimate 1-hour and 8-hour average ozone concentrations, respectively.

### 7.4.2 Advanced Methodologies

**Data representativeness problem**

Davis et al. (2000) deal with the problem of the data representativeness using spatial statistical techniques to compare observed ozone fields with the surface level ozone forecast fields from grid models. The 8-hour average daily observed ozone at the monitoring sites was interpolated to the model grid cells using a spatial statistical method, and then differences between the model output fields and the spatially interpolated observed fields were compared.

**Scale analysis methodologies**

Hogrefe et al. (2001a, 2001b) and Biswas et al. (2001) suggest that there are several shortcomings in using traditional performance measures, such as: if data assimilation is applied, the required statistical independence of the observed and simulated data sets is violated; traditional statistics provide little insight into the physical behavior of the model (i.e. they do not give any insight into the model's ability to reproduce the spatial and temporal correlation structures embedded in the observations on various scales). Therefore, to analyze meteorological input parameters, ozone predictions, predictions of ozone precursors, and predictions of ozone-precursors relationship, they introduced additional model evaluation methods based on the concept of scale analysis. To this end, a spectral decomposition technique is applied. It then becomes evident that model performance is time-scale specific and, therefore, the outcome of model evaluation on different time scales can be tied to the model formulation of the relevant processes on these time scales.

Time series of ozone observations contain fluctuations occurring on many different time scales (e.g., Vukovich, 1997; Sebald et al., 2000). Since ozone observations are taken at discrete intervals, the highest and lowest frequencies that can be estimated for any particular time series are determined by the sampling interval and the length of data record, respectively. The choice of the different frequency bands used by Hogrefe et al. (2001a, 2001b) and Biswas et al. (2001) was performed both on the recorded power spectrum and on *a priori* knowledge about different physical processes of interest to the simulation of air quality. They choose: the intra-day (ID) range (periods less than 12 hours), the diurnal (DU) range (periods of 24 hours), the synoptic (SY) range (periods of 2-21 days), and long-term baseline (BL) fluctuations (containing periods greater than 21 days).

The intra-day fluctuations are determined by the effects of turbulent horizontal and vertical mixing, and ozone response to fast-changing emissions patterns during traffic rush hours. Diurnal fluctuations are associated with the diurnal variation of the solar flux, and the resulting differences between the daytime photochemical production and the nighttime removal of ozone as well as the diurnal cycle of boundary layer evolution and decay. The variations of ozone on the synoptic scale are caused by changing meteorological conditions such as the presence of a nearly stagnant high pressure system or the passage of frontal systems. Fluctuations in baseline are caused by seasonal variations of the solar flux, changing large-scale flow patterns, and change in vegetation coverage and biogenic emissions.

Hogrefe et al. (2001a, 2001b) and Biswas et al. (2001) used the Kolmogorov-Zurbenko (KZ) filter (Zurbenko, 1986) because of its powerful separation characteristics, simplicity, and ability to handle missing data. This technique is described in more detail in Eskridge et al. (1997), Rao et al. (1997) and Hogrefe et al. (2000). In the following, we give only an outline.

The temporal components mentioned previously are estimated as follows:

$$ID(t) = \ln[O_3(t)] - KZ_{3,3}\{\ln[O_3(t)]\} \tag{26}$$

$$DU(t) = KZ_{3,3}\{\ln[O_3(t)]\} - KZ_{13,5}\{\ln[O_3(t)]\} \tag{27}$$

$$SY(t) = KZ_{13,5}\{\ln[O_3(t)]\} - KZ_{103,5}\{\ln[O_3(t)]\} \tag{28}$$

$$BL(t) = KZ_{103,5}\{\ln[O_3(t)]\} \tag{29}$$

where $KZ_{m,k}$ is the KZ filter with a window size of m hours and k iterations. Thus, by adding all components as defined in Equations (26), (27), (28), and (29), the ozone process is represented as

$$\ln[O_3(t)] = ID(t) + DU(t) + SY(t) + BL(t) \tag{30}$$

where the intra-day, diurnal and synoptic components are zero-mean processes. The actual ozone concentration in the ppb scale can be obtained as

$$O_3(t) \approx e^{ID(t)} \times e^{DU(t)} \times e^{SY(t)} \times e^{BL(t)} \tag{31}$$

As far as the model's ability to simulate ozone fields is concerned, Hogrefe et al. (2001b) first compared the relative importance of the individual components to the overall ozone process for both observations and model predictions. For this purpose, the variance of each component is computed and compared to the overall variance for both observations and model predictions. Then, to compare the

absolute amount of energy on different time scales between observation and model predictions, Hogrefe et al. (2001b) listed the ratios of the variances of the modeled to the observed time series for different time scales. They concluded that the models characterize best those variations having time scales longer than several days. They suggested that to increase confidence in the regulatory modeling process, the modeling period should be several synoptic cycles in duration rather than the 2-3 days of a single ozone pollution episode.

**Process oriented methodologies**

As we have already said, the evaluation goal may concern the overall model (integrated), or can be process oriented (diagnostic). In this case, one will have to make a selection of the model characteristics/modules to be validated. Several authors (e.g., Hass et al., 1997; Dennis et al., 1999; Tonnesen and Dennis, 2000a and 2000b; Luecken et al., 1999) have been asking that model evaluation of complex grid models be process oriented. In other words, they want to know if the model is describing how things happen correctly without too much emphasis on whether the magnitude of the changes predicted are correct. Therefore, they do not believe that a photochemical model has been fully evaluated if comparisons are only observed versus predicted ozone. They are worried that a model can give the correct result for the wrong reason. So, they want the evaluation to "look inside" the model, and to analyze modules to see if the model has modeled the right causes for the effects seen.

### 7.4.3 Final Remarks

It is concluded that validating the performance of Eulerian grid models is not philosophically different than validating the performance of plume or puff models. The pattern for inert species is just as difficult to characterize for any of the various model types. To validate performance for characterizing the inert species pattern, the same thought process would be followed, regardless of the model being validated. For reactive species, the pattern appears to have fewer anomalies, but characterization of the chemistry, initial and boundary conditions, and characterization of the precursor emission rates are very uncertain. The logistics of running several Eulerian grid models for the same field studies are found to have their own sets of problems and constraints. Furthermore, field studies of reactive species rarely provide a time series long enough for developing confidence bounds to formally test whether differences seen in comparing different models is significant.

## 8 Model Quality Assurance

Confidence in using air quality models in scientific studies, as well as in operational decision-making applications is founded on a program of quality assurance. The definition of quality assurance can be inferred from different sources. From ISO 14000 (International Standards on Environmental

Management), the definition of quality assurance is all those planned and systematic actions necessary to provide adequate confidence that a product or service will satisfy given requirements for quality. From U.S. Environmental Protection Agency QA/G-5 (1998) and EUROTRAC (http://www.gsf.de/eurotrac/organisation/g-qa-qc.htm), quality assurance is defined as an integrated system of management activities involving planning, implementation, assessment, reporting, and quality improvement to ensure that a process, item or service is of the type and quality needed and expected by the user.

## 8.1 Overview of Model Quality Assurance

Model quality assurance can be envisioned as documentation of the following items (e.g., Borrego and Tchepel, 1998; Borrego et al., 2001b): definition of purpose and scope of the modeling, model description, database description, selection of performance measures, model evaluation, scientific peer review, and user oriented assessment. Therefore, model evaluation (see Section 6), which in turn includes statistical model evaluation (see Section 7), is a basic component of model quality assurance.

- **Definition of purpose and scope of the modeling.**
  - Identification of the type of model under evaluation: long range transport models, photochemical models at continental scale, photochemical models at urban scale (without obstacles), street canyon models (urban scale with obstacles), stack models, concentration fluctuation models, dense gas models, indoor pollution, and other models[9].
  - Identification of the purpose of the modeling: air quality assessment (determine impact on human health, ecosystems), regulatory purpose (e.g., calculation of a minimum stack height for new installation), policy support (e.g., scenario studies on effect of emission abatement measures), emergency planning (estimation of hazardous gas concentration), public information (e.g., online information on the possible occurrence of smog episodes), and scientific research (better understanding of physical/chemical processes involving air pollution).

- **Model description.** Availability of extended description of the model is important for quality assurance procedures. The model description should include a detailed description of the physics and chemistry contained in the model. The description should include a summary of the model characteristics (e.g., model approximations, time and space resolution, modeling scale). Furthermore, it should contain details of the model such as: model name, version number, date of first release, area of application,

[9] For example, mesoscale flow models (that are the necessary support for dispersion in complex terrain), chemical modules, chemical heterogeneous reactions, cloud formation models, models for aerosol transformation and growth, model for turbulence, etc. Some of these models are for the purposes mentioned, others are only used to understand physical phenomena and eventually are inserted as sub-models in some of the mentioned models.

originating organization, source of model (from the originators, through third parties, or in particular whether it is an improved version of an earlier model), model type, hardware requirements, software requirements, and references.

- **Database description.** Database description first identifies the data used to construct the model parameters at the development stage. Then it identifies the data used during the process of both "model algorithm verification" and "model evaluation". It contains details as data ownership/accessibility and origin of the data (from analytic results, simulated by higher-order models, laboratory experiments, field experiments, incident reports). Database selection includes consideration of factors such as: data quality assurance, completeness, appropriateness, model features/parameters covered, data uncertainties (which concern both data used as model inputs, e.g., emission and meteorological data, and data used to make a comparison against model outputs, e.g., pollutant concentrations), and data representativeness.

- **Selection of performance measures.** This would include selection of evaluation tools (quantitative or/and qualitative) as statistical indices and graphical methodologies to compare model outputs with observed values. Performance measures reflect the ability of a model to simulate real world phenomena; it helps in understanding a model's limitations and provides support for model inter-comparisons.

- **Model evaluation.** Model evaluation is the overall system of procedures designed to measure the model performance, and includes: model algorithm verification, sensitivity analysis, uncertainty analysis, statistical model evaluation, and model inter-comparisons (see Section 6).

- **Scientific peer review.** Scientific peer review includes: an assessment of the appropriateness of the scientific content; the limits of applicability of the model; limitations and advantages of the model; and possible improvements. A further objective of a scientific peer review is to guarantee that all steps of model evaluation were implemented in agreement with a model's requirements. For example, good models will likely exhibit poor correlation with observations if applied in a manner inconsistent with their physics assumptions. For instance, the modeled concentration values from a mesoscale photochemical model will compare poorly with observations from an urban station directly affected by traffic emissions. Scientific peer review may involve expert external analysis.

- **User oriented assessment.** User oriented assessment provides information on: availability of the model, associated documentation, installation procedures, user interface, ease of use, guidance in selecting model options and input data, limitations on the applicability of the model, explanations concerning the output, clarity of warnings and error messages, computational costs, and possible improvements.

## 8.2 Related Topics

The following items describe some topics related to model quality assurance.

- **Quality assurance of emission inventories.** Harmonization of the methodology for the compilation of emission inventories is needed. Furthermore, an effort should be focused on developing objective estimates of the uncertainties in emission inventories.

- **Data quality assessment and data quality objectives.** Data quality assessment (DQA) is the scientific and statistical evaluation of data to determine if the data obtained from environmental data operations are of the right type, quality and quantity to support their intended use (U.S. Environmental Protection Agency QA/G-9, 1998). A data quality objective (DQO) is a range of acceptability of measured data for a specific application. The definition of DQO depends on the project scientific objectives and on intended use of the data. Different monitoring programs have distinct DQOs. To estimate the quality of measurements, the data quality indicators (DQI) are used. DQI are (Borrego et al., 2001a): bias (systematic error); precision (random error); accuracy (combination of systematic and random errors); and completeness (percent of valid measurements). The uncertainties of measurements have to be reported and considered in data application.

- **Model calibration.** Model calibration is a procedure used to make, at the model development stage, estimates of parameters within the model equations, which best fit the general model structure to a specific observed data set. Note that successful model calibration only indicates that the structure of the model includes the important variables that influence behavior (or correlated well with variables that influence behavior) under the conditions prevailing for the calibration data set. Model calibration does not ensure that the model will predict well under conditions that are quite different than were used in the calibration. For this reason, as new data become available, models almost invariably need additional calibration. When updating calibrated values within a model, one should consider previously used data, as well as the newly acquired data. Finally, use of any model beyond its proven range of application will involve expert judgment, knowledge of the physical processes being modeled, and an awareness of the sensitivity of the model's output to changes in input.

- **Data assimilation**. Data assimilation is a numerical technique, which makes it possible to combine model results and observations in one integrated system. Observations are input to the numerical system, which consists of the model combined with the data assimilation technique. To several parameters (either internal model parameters or input data), noise factors are added. The system will attempt to minimize the discrepancy between calculated concentrations and observations. An essential consideration in this process is

to balance the "accepted range of disparity" (noise factors) with the data representativeness of the observations in time and space.

# 9 Guidelines for Model Evaluation: Towards Harmonization in Model Evaluation Methodology

Although presently the evaluation methodology is generally left to the user to define, it is important to realize that efforts are underway to standardize evaluation methodologies. This would allow comparison of model evaluation results performed by different users. It would also provide a standard manner for gaining acceptance of models for various operational uses. As experience increases, it is hoped that consensus will be reached in certain evaluation goals, evaluation objectives and associated evaluation methodologies, and data sets to be employed. The ultimate goal will be to define a standard evaluation methodology for each evaluation goal.

## 9.1 The USA Effort

Within the United States, the emphasis has been on the development, evaluation and application of air quality simulation models that allow development of air quality management plans to achieve defined national air quality goals. These plans involve development of emission control strategies sometimes for individual sources ("primary" impacts associated with pollutants emitted directly into the atmosphere) and sometimes for classes of sources ("secondary" impacts associated with pollutants formed during transport). Part of the decision on which model to select is dictated by ensuring that the appropriate physical processes are addressed by the model. However, another part of the decision in model selection is the recognition that every model is a compromise in that not all processes are included or else the computational demands would become excessive. Hence, model selection often involves expert judgment based on actual experience in the use and application of the various models available.

The American Society for Testing and Materials (ASTM) has published a "Standard guide for statistical evaluation of atmospheric dispersion models" (ASTM D6589). This guide provides a general philosophy that can be used to design statistical model evaluation procedures, either for the comparison of modeled concentrations with observations, or to assess one model's performance relative to other candidate models.

Founded in 1995, NARSTO[10] (http://www.cgenv.com/Narsto/) is a public/private partnership, whose membership spans the government, utilities, industry, and academy throughout Mexico, the United States, and Canada. Its primary mission is to coordinate and enhance policy-relevant scientific research, and assess

[10] Formerly an acronym for "North American Research Strategy for Tropospheric Ozone".

tropospheric pollution behavior; its activities provide input for science-based decision-making and determination of workable, efficient, and effective strategies for local and regional air-pollution management. NARSTO has an ongoing activity to evaluate regional air-pollution models by comparing output from multiple models as well as by testing against data obtained from NARSTO field intensives.

## 9.2 The European Effort

During the last few years in Europe, many insights have been given about the need to improve model evaluation quality. Excellent examples are the ETEX campaigns on the real-time assessment of the long-range atmospheric dispersion of harmful releases (Mosca et al., 1998, http://rem.jrc.cec.eu.int/etex/) and the RTMOD exercises (Bellasio et al., 1998, http://rtmod.jrc.it/rtmod/). The need to understand the differences between operational uses of air quality models and the desire to reduce the disparity between different models when applied to the same problem was highlighted in recent International Conferences. These conferences (there have been seven so far) were organized with the aim of "Harmonization within Atmospheric Dispersion Modeling for Regulatory Purposes" (http://www.harmo.org/), and attracted an increasingly large scientific participation (Olesen, 1996, 2001). A central activity of the "Harmonization" initiative is the distribution of a Model Validation Kit[11]. The Model Validation Kit is a collection of three experimental data sets accompanied by software for model evaluation. It is a practical tool meant to serve as a common frame of reference for modelers. The experience gained from these conferences together with that in the field of long-range dispersion for accidental releases points in the direction of inter-comparing mesoscale flow models. This is the content of MESOCOM (http://java.ei.jrc.it/Projects/MESOCOM), which is currently ongoing. The EUROTRAC-2 subprojects GLOREAM [12] and SATURN (http://aix.meng.auth.gr/saturn/) are aimed at the formulation of suitable evaluation methodologies for regional and urban scale air pollution models, respectively. The German organization BWPLUS (http://bwplus.fzk.de/) is presently promoting an inter-comparison of methods for the prediction of the air pollutant concentrations in a specific street canyon using usually available input data.

Furthermore, it is useful to recall the Data Sets for Atmospheric Modeling (DAM) initiative of the JRC (http://java.ei.jrc.it/Projects/DAM). DAM's objective is to facilitate the accessibility of datasets, presently available to the Scientific Community for atmospheric model evaluation, to any model developer or user that intends to validate his/her modeling tool. DAM is intended as an interface between modelers and the information available through existing web sites or other contact points.

---

[11] http://www.dmu.dk/atmosphericenvironment/Harmoni/M_V_KIT.htm.

[12] http://www.dmu.dk/AtmosphericEnvironment/gloream/.

**Acknowledgements**

Prof. Carlos Borrego, Prof. Peter Builtjes, Dr. Ana Cristina Carvalho, Dr. Jerry Davis, Dr. Giovanni Graziani, Mr. Ariel Stein, Dr. Oxana Tchepel, Dr. Steve Warner, and WIT press (Southampton) are gratefully acknowledged. Dr. Canepa's contribution to this work was supported by the MURST COFIN 99 project ACME CUE, and the INFM PA project GEPAIGG01. The information in this document has been funded in part by the United States Environment Protection Agency under an Interagency Agreement (139-384-83) to the National Oceanic and Atmospheric Administration. It has been subjected to Agency review for approval for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

**References**

Alcamo, J., and J. Bartnicki (1987): A framework for error analysis of a long-range transport model with emphasis on parameter uncertainty. *Atmos. Environ*., **21**(10): 2121-2131.

Anthes, R.A. (1983): Review - regional models of the atmosphere in middle latitudes. *Monthly Weather Review*, **111**: 1306-1335.

Arnold, J.R., Dennis, R.L., and Tonnesen, G.S., (1998): Advanced techniques for evaluating Eulerian air quality models: background and methodology. Proceeding of the 10th Joint Conference on the Application of Air Pollution Meteorology with the A&WMA, Phoenix, AZ., pp. 1-5.

Barad, M.L., Editor (1958): Project Prairie Grass, A Field Program in Diffusion. Geophysical Research Paper, No. 59, Vol I and II, Report AFCRC-TR-58-235, Air Force Cambridge Research Center, 439 pp.

Bellasio, R., R. Bianconi, G. Graziani, and S. Mosca (1998): RTMOD: An Internet based system to analyze the predictions of long-range atmospheric dispersion models. *Computer Geosciences* **25**: 819-833.

Biswas, J., S.T. Rao, P. Kasibhatla, W. Hao, G. Sistla, R. Mathur, and J. McHenry (2001): Evaluating the performance of regional-scale photochemical modeling systems: part III – precursor predictions and ozone-precursor relationship. *Atmos. Environ.,* **35**: 6129-6149.

Bornstein, R.D., and S.F. Anderson (1979): A survey of statistical techniques used in validation studies of air pollution prediction models. Technical Report No. 23, Stanford University, Stanford, California.

Borrego, C., and O. Tchepel (1998): Quality assurance/quality control in SATURN: a first approach. Hamburg Workshop, August 1998.

Borrego, C., N. Barros and O. Tchepel (2001a): Data Quality Assessment: Lisbon experimental field campaign LisbEx 96. Transport and Chemical Transformation in the Troposphere. Proceedings from the EUROTRAC-2 Symposium 2000, 27-31 March 2000, Garmisch-Partenkirchen, Germany. P.M. Midgley, M. Reuther, M. Williams (Eds.). Springer-Verlag Berlin, Heidelberg, Germany, on CD-ROM.

Borrego, C., O. Tchepel and A.C. Carvalho (2001b): Model Quality Assurance. Transport and Chemical Transformation in the Troposphere. Proceedings from the EUROTRAC-2 Symposium 2000, 27-31 March 2000, Garmisch-Partenkirchen, Germany. P.M. Midgley, M. Reuther, M. Williams (Eds.). Springer-Verlag Berlin, Heidelberg, Germany, pp. 21-26.

Brandt, J., A. Bastrup-Birk, J.H. Christensen, T. Mikkelsen, S. Thykier-Nielsen, and Z. Zlatev (1998): Testing the importance of accurate meteorological input fields and parameterizations in atmospheric transport modeling using DREAM – validation against ETEX-1. *Atmos. Environ*., **32**(24): 4167-4186.

Brost, R.A., P.L. Haagenson, and Y.H. Kuo (1988): The effect of diffusion on tracer puffs simulated by a regional scale Eulerian model. *Journal of Geophysical Research*, **93**(D3): 2389-2404.

Brusasca, G., G. Tinarelli, and D. Anfossi (1989): Comparison between the results of a Monte Carlo atmospheric diffusion model and tracer experiments. *Atmos. Environ.,* **23**(6): 1263-1280.

Canepa, E., and F. Modesti (1997): Considerations about statistical indices used in the evaluation of air quality models. Air Pollution 97, 16-18 September 1997, Bologna, Italy, *Air Pollution V*, pp. 607-616, H. Power, T. Tirabassi and C.A. Brebbia editors, Computational Mechanics Publications, Southampton, UK.

Canepa, E., and P.J.H. Builtjes (1999): Methodology of model testing and application to dispersion simulation above complex terrain, *Conference Proceedings on CD-ROM*, 6th International Conference on Harmonization within Atmospheric Dispersion Modeling for Regulatory Purposes, 11-14 October 1999, INSA de Rouen, France. [to be published in *Int. J. Environ. Pollut*.]

Carhart, R.A., A.J. Policastro, M. Wastag, and L. Coke (1989): Evaluation of eight short-term long-range transport models using field data. *Atmos. Environ*., **23**(1): 85-105.

Carruthers, D.J., A.M. Mckeown, D.J. Hall, and S. Porter (1999): Validation of ADMS against wind tunnel data of dispersion from chemical warehouse fires. *Atmos. Environ*., **33**: 1937-1953.

Clarke, J.F. (1964): A Simple Diffusion Model for Calculating Point Concentrations from Multiple Sources. *Journal of the Air Pollution Control Association*, **14**(9): 347-352.

Cole, S.T., and P.J. Wicks, Editors (1995): Model Evaluation Group: Report of the Second Open Meeting. EUR 15990 EN, European Commission, Directorate-General XII, Environmental Research Programme, L-2920 Luxembourg, 77 pp.

Cox, W.M., and J.A. Tikvart (1990): A statistical procedure for determining the best performing air quality simulation model. *Atmos. Environ*., **24A**(9):2387-2395.

Cox, R.M., J. Sontowski, R.N. Fry Jr., C.M. Dougherty, and T.J. Smith (1998): Wind and diffusion modeling for complex terrain. *J. Appl. Meteor*., **37**: 996-1009.

Davis, J.M., D. Nychka and B. Bailey (2000): A comparison of regional oxidant model (ROM) output with observed ozone data. *Atmos. Environ*., **34**: 2413-2423.

Dekker, C.M., A. Groenendijk, C.J. Sliggers, and G.K. Verboom (1990): Quality Criteria for Models to Calculate Air Pollution. Lucht (Air) 90, Ministry of Housing, Physical Planning and Environment, Postbus 450, 2260 MB Leidschendam, The Netherlands, 52 pp.

Dennis R.L. (1986): Issue, design and interpretation of performance evaluations: ensuring the emperor has clothes. *Air Pollution Modeling and Its Application V*, edited by DeWispelaere V.C., Schiermeier F.A. and Gillani N.V., pp. 411-424, Plenum Press, New York.

Dennis, R.L., Arnold, J.R., Tonnesen, G.S., Y. Li (1999): A new response surface approach for interpreting Eulerian air quality model sensitivities. *Computer Physics Communications*, **117**: 99-112.

Desiato, F. (1991): A dispersion model evaluation study for real-time application in complex terrain. *J. Appl. Meteor*., **30**: 1207-1219.

Efron, B., and R.J. Tibshirani (1993): *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57, Chapman & Hall, New York, 1993, 436 pp.

Eskridge, R.E., J.Y. Ku, S.T. Rao, P.S. Porter, and I.G. Zurbenko (1997): Separating different scales of motion in time series of meteorological variables. *Bull. Amer. Meteor. Soc*., **78**: 1473-1483.

Fox, D.G. (1981): Judging air quality model performance: a summary of the AMS workshop on dispersion model performance. *Bull. Amer. Meteor. Soc.,* **62**: 599-609.

Fox, D.G. (1984): Uncertainty in air quality modeling. *Bull. Amer. Meteor. Soc*., **65**(1): 27-36.

Graziani G., W. Klung and S. Mosca (1998): Real-time long-range dispersion model evaluation of ETEX first release, EUR 17754 EN, ISBN 92-828-3657-6.

Gronskei, K.E., S.E. Walker, and F. Gram (1993): Evaluation of a model for hourly spatial concentration distributions. *Atmos. Environ*., **27B**(1):105-120.

Hanna, S.R. (1971): A simple method of calculating dispersion from urban area sources. *Journal of the Air Pollution Control Association*, **21**(22): 774-777.

Hanna, S.R. (1988): Air quality model evaluation and uncertainty. *Journal of the Air Pollution Control Association*, **38**: 406-412.

Hanna, S.R. (1989a): Confidence limits for air quality model evaluation as estimated by bootstrap and jackknife resampling methods. *Atmos. Environ*., **23**(6): 1385-1398.

Hanna, S.R. (1989b): Plume dispersion and concentration fluctuations in the atmosphere. P.N. Cheremisinoff (Editor). *Encyclopedia of Environmental Control Technology, Volume 2, Air Pollution Control*. Gulf Publishing Company, Houston, TX, pp. 547-582.

Hanna, S.R. (1993): Uncertainties in air quality model predictions. *Boundary-Layer Met*. **62**: 3-20.

Hanna, S.R., and R.J. Paine (1989): Hybrid Plume Dispersion Model (HPDM) development and evaluation. *J. Appl. Meteor.,* **28**: 206-224.

Hanna, S.R., and J.C. Chang (1993): Hybrid plume dispersion model (HPDM) improvements and testing at three field sites. *Atmos. Environ.*, **27A**(9): 1491-1508.

Hanna, S.R., G.A. Briggs, and R.P. Hosker (1982): *Handbook on Atmospheric Diffusion*. DOE/TIC-11223 (NTIS Document DE82002045), National Technical Information Service, U.S. Department of Commerce, Springfield, VA, 102 pp.

Hanna, S.R., J.C. Chang, and D.G. Strimaitis (1993): Hazardous gas model evaluation with field observation. *Atmos. Environ.,* **27A**(15): 2265-2285.

Hanna, S.R., G.E. Moore, and M.E. Fernau (1996): Evaluation of photochemical grid models (UAM-IV, UAM-V, and the ROM/UAM-IV couple) using data from the Lake Michigan Ozone Study (LMOS). *Atmos. Environ.*, **30**(19): 3265-3279.

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. *Atmos. Environ.*, **32**: 3617-3628.

Hass, H., P.J.H. Builtjes, D. Simpson, and R. Stern. (1997): Comparison of model results obtained with several European regional air quality models. *Atmos. Environ*., **31**(19): 3259-3279.

Haugen, D.A. Editor (1959): Project Prairie Grass, A Field Program in Diffusion. Geophysical Research Paper, No. 59, Vol III, Report AFCRC-TR-58-235, Air Force Cambridge Research Center, 673 pp.

Heidam, N.Z. (1987): Bootstrap estimates of factor model variability. *Atmos. Environ.*, **21**(5): 1203-1217.

Hilst, G.R. (1970): Sensitivities of Air Quality Prediction to Input Errors and Uncertainties. Proceedings of Symposium on Multiple-Source Urban Diffusion Models, Air Pollution Control Office Publication No. AP-86, U.S. Environmental Protection Agency, Research Triangle Park, NC, Chap. 8, 41 pp.

Hogrefe, C., S.T. Rao, I.G. Zurbenko, and P.S. Porter (2000): Interpreting the information in ozone observation and model predictions relevant to regulatory policies in the Eastern United States. *Bull. Amer. Meteor. Soc*., **81**: 2083-2106.

Hogrefe, C., S.T. Rao, P. Kasibhatla, G. Kallos, G.J. Tremback, W. Hao, D. Olerud, A. Xiu, J. McHenry, and K. Alapaty (2001a): Evaluating the performance of regional-scale photochemical modeling systems: part I – meteorological predictions. *Atmos. Environ.,* **35**: 4159-4174.

Hogrefe, C., S.T. Rao, P. Kasibhatla, W. Hao, G. Sistla, R. Mathur, and J. McHenry (2001b): Evaluating the performance of regional-scale photochemical modeling systems: part II – ozone predictions. *Atmos. Environ.,* **35**: 4175-4188.

Irwin, J.S., (1998): "Statistical evaluation of atmospheric dispersion models," 5-th International Conference on Harmonization with Atmospheric Dispersion Modeling for Regulatory Purposes, May 18-21, 1998, Rhodes, Greece (Preprints pp. 46-53, to be published in International Journal of Environment and Pollution), 8 pp.

Irwin, J.S. (1999): Effects of concentration fluctuations on statistical evaluation of centerline concentration estimates by atmospheric dispersion models, *Conference Proceedings on CD-ROM*, 6th International Conference on Harmonization within Atmospheric Dispersion Modeling for Regulatory Purposes, 11-14 October 1999, INSA de Rouen, France. [to be published in *Int. J. Environ. Pollut*.]

Irwin, J.S. (2000): Modeling Air Quality Pollutant Impacts. Air Quality Management In Urban Areas In the Light Of EU Legislation, November 6-7, Krakow, Poland. (EURASAP Newsletter, Number 40, pages 2-28, ISSN-1026-2172, available at http://www.meteo.bg/EURASAP/Newsletter.html).

Irwin, J.S. and M.E. Smith (1984): Potentially Useful Additions to the Rural Model Performance Evaluation, Bulletin American Meteorological Society, Vol 65, pp. 559-568.

Irwin, J.S., and R.F. Lee (1997): Comparative evaluation of two air quality models: within-regime evaluation statistic. *Int. J. Environ. Pollut*., **8**(3-6): 346-355.

Irwin, J.S., S.T. Rao, W.B. Peterson, and D.B. Turner (1987): Relating error bounds for maximum concentration estimates to diffusion meteorology uncertainty. *Atmos. Environ.,* 21(9): 1927-1937.

Kumar, N., A.G. Russel, T.W. Tesche, and D.E. McNally (1994): Evaluation of CALGRID using two different ozone episodes and comparison to UAM results. *Atmos. Environ*., **28**(17): 2823-2845.

Kumar Yadav, A., and M. Sharan (1996): Statistical evaluation of sigma schemes for estimating dispersion in low wind conditions. *Atmos. Environ*., **30**(14): 2595-2606.

Lanzani, G., and M. Tamponi (1995): A microscale Lagrangian particle model for the dispersion of primary pollutants in a street canyon: sensitivity analysis and first validation trials. *Atmos. Environ*., **29**(23): 3465-3475.

Larsen, R.I. (1969): A new mathematical model of air pollution concentration averaging time and frequency. *Journal of the Air Pollution Control Association*, **19**: 24-30.

Longhetto, A., Editor (1980): *Atmospheric Planetary Boundary Layer Physics.* Elsevier, New York.

Luecken, D.J, Tonnesen, G.S., and Sickles, J.E., II (199): Differences in NOx speciation predicted by three photochemical mechanisms. *Atmos. Environ.,* (33): 1073-1084.

Luhar, A.K., and K.S. Rao (1994): Lagrangian stochastic dispersion model simulations of tracer data in nocturnal flows over complex terrain. *Atmos. Environ.*, **28**(21): 3417-3431.

Lumley, J.L, and H.A. Panofsky (1964): *The structure of atmospheric turbulence*. Wiley Interscience, New York, 239 pp.

Lurmann, F.W., and N. Kumar (1997): Evaluation of the UAM-V model performance in OTAG simulations phase I: summary of performance against surface observations. Final Report STI-996120-1605-FR, Sonoma Technology, Inc., Santa Rosa, CA. Prepared for: Science Applications International Corporation, McClean, VA.

Martin, D.O. (1971): An Urban Diffusion Model for Estimating Long Term Average Values of Air Quality. *Journal of the Air Pollution Control Association*, **21**(1): 16-19.

Model Evaluation Group (1994a): Guidelines for model developers. Can be requested from Dr. S. Cole, DG XII/D1, Rue de la Loi 200, B-1049 Belgium. Fax +32 2 296 3024.

Model Evaluation Group (1994b): Model Evaluation Protocol. Can be requested from Dr. S. Cole, DG XII/D1, Rue de la Loi 200, B-1049 Belgium. Fax +32 2 296 3024.

Moore, G.E., M.K. Liu, and R.J. Londergan (1985): Diagnostic validation of Gaussian and first-order closure plume models at a moderately complex terrain site. Systems Applications, Inc., Final Report EA-3760, San Rafael, California.

Morris, R.E., M. Yocke and T.C. Myers (1993): Application of the nested grid urban airshed model to the Lake Michigan region. Presented at AW&MA International Conference and Course, Tropospheric Ozone: Nonattainment and Design Value Issues, 27-30 October, Boston, Massachusetts. (UAM-V is available at: http://uamv.saintl.com/).

Mosca, S., G. Graziani, W. Klug, R. Bellasio, and R. Bianconi (1998): A statistical methodology for the evaluation of long-range dispersion models: an application to the ETEX exercise. *Atmos. Environ*., **32**(24): 4307-4324.

Okamoto, S., H. Ohnishi, T. Yamada, T. Mikami, S. Momose, H. Shinji, and T. Itohiya (1999): A model for simulating atmospheric dispersion in a low-wind condition. *Conference Proceedings on CD-ROM*, 6th International Conference on Harmonization within Atmospheric Dispersion

Modeling for Regulatory Purposes, 11-14 October 1999, INSA de Rouen, France. [to be published in *Int. J. Environ. Pollut*.]

Olesen, H.R., (1995): The model validation exercise at Mol: overview of results. Workshop on Operational Short-range Atmospheric Dispersion Models for Environmental Impact Assessment in Europe, Mol, November 1994, published in: *Int. J. Environ. Pollut*., **5**(4-6): 761-784.

Olesen, H.R. (1996): Toward the establishment of a common framework for model evaluation. In: Air Pollution Modeling and Its Application XI, pp. 519-528. Edited by S-E. Gryning and F. Schiermeier, Plenum Press, New York.

Olesen, H.R. (2001): Ten years of harmonization: past, present, and future. Presentation at the 7th International conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, May, 2001, Belgirate, Italy. 10 pp.
(http://www.dmu.dk/atmosphericenvironment/harmoni.htm)

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

Pasquill, F. (1961): The estimation of the dispersion of windborne material. *Meteorological Magazine*, **90**: 33-49.

Poli, A.A., and M.C. Cirillo (1993): On the use of the normalized mean square error in evaluating dispersion model performance. *Atmos. Environ*., **15**: 2427-2434.

Randerson, D., Editor (1984): Atmospheric Science and Power Production. DOE/TIC-27601 (NTIS Document DE84005177). National Technical Information Service, U.S. Department of Commerce, Springfield, VA, 850 pp.

Rao, S.T., I.J. Zurbenko, R. Neagu, P.S. Porter, J.Y. Ku, and R.F. Henry (1997): Space and time scales in ambient ozone data. *Bull. Amer. Meteor. Soc*., **78**: 2153-2166.

Reynolds, S.D., C. Seigneur, T.E. Stoeckenius, G.E. Moore, R.G. Johnson, and R.J. Londergan (1984): Operational validation of Gaussian plume models at a plain site. System Applications, Inc., Final Report EA-3076, San Rafael, California.

Reynolds, S.D., T.C. Myers, J.E. Langstaff, M.K. Liu, G.E. Moore, and R.E. Morris (1985): Operational validation of Gaussian and first-order closure plume models at a moderately complex terrain site. System Application, Inc., Final Report EA-3759, San Rafael, California.

Ross, D.G., and D.G. Fox (1991): Evaluation of an air pollution analysis system for complex terrain. *J. Appl. Meteor.,* **30**(7): 909-923.

Ruff, R.E., K.C. Nitz, F.L. Ludwig, C.M. Bhumralkar, J.D. Shannon, C.M. Sheih, I.Y. Lee, R. Kumar, and D.J. McNaughton (1984): Regional air quality model assessment and evaluation. SRI International Final Report EA-3671, Menlo Park, California.

SAIC (1997): Summary of the UAM-V model performance in OTAG simulations phase II: 8-hour performance statistics. Science Applications International Corporation, Final Report, Raleigh, North Carolina.

Schatzmann, M., and B. Leitl (1999): Quality assurance of urban dispersion models. *Conference Proceedings on CD-ROM*, 6th International Conference on Harmonization within Atmospheric Dispersion Modeling for Regulatory Purposes, 11-14 October 1999, INSA de Rouen, France (to be published in *Int. J. Environ. Pollut*.).

Schlunzen, K.H. (1997): On the validation of high-resolution atmospheric mesoscale models. *J. Wind Eng. Ind. Aerod.*, **67**&**68**: 479-492.

Sebald, L., R. Treffeisen, E. Reimer, and T. Hies (2000): Spectral analysis of air pollutants. Part 2: ozone time series. *Atmos. Environ.*, **34**: 3503-3509.

Smith, M.E. (1984): Review of the attributes and performance of 10 rural diffusion models. *Bull. Amer. Meteor. Soc.*, 65: 554-558.

Stein, A.F., and J.C. Wyngaard (2000): Fluid modeling and the evaluation of inherent uncertainty. GMU Transport and Dispersion Modeling Workshop, 11-12 July 2000, George Mason University, Fairfax, Virginia.

Stein, A.F., and J.C. Wyngaard (2001): Fluid modeling and the evaluation of inherent uncertainty. Journal of Applied Meteorology. Vol. 40(10): 1769-1774.

Tonnesen, G.S., and R.L. Dennis (2000a): Analysis of radical propagation efficiency to assess ozone sensitivity to hydrocarbons and NOx 1. local indicators of instantaneous odd oxygen production sensitivity. *J. of Geophys. Res.*, 105, D7: 9213-9225.

Tonnesen, G.S., and R.L. Dennis (2000b): Analysis of radical propagation efficiency to assess ozone sensitivity to hydrocarbons and NOx 2. long-lived species as indicators of ozone concentration sensitivity. *J. of Geophys. Res.*, 105, D7: 9227-9241.

Tesche, T.W., P. Georgopoulos, J.H. Seinfeld, F. Lurmann, and P.M. Roth (1990): Improvements in procedures for evaluating photochemical models. Report A832-103, California Air Resources Board, Sacramento, California.

Thuillier, R.H. (1992): Evaluation of a puff dispersion model in complex terrain. *J. Air Waste & Manage. Assoc.*, **42**: 290-297.

U.S. Environmental Protection Agency (1991): Guideline for regulatory applications of the Urban Airshed Model. EPA-450/4-91-013, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711.

U.S. Environmental Protection Agency (1992): Protocol for Determining the Best Performing Model, EPA-454/R-92-025, Office of Quality Planning and Standards

U.S. Environmental Protection Agency (1996) Compilation of Photochemical Models' Performance Statistic for 11/94 Ozone SIP Applications. EPA-454/R-96-004. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, 156 pp.

U.S. Environmental Protection Agency (1997): Evaluation of Modeling Tools for Assessing Land Use Policies and Strategies. EPA 420/R-97-007. U.S. Environmental Protection Agency, Ann Arbor, MI, 60 pp. (available at http://www.epa.gov/orcdizux/transp/publicat/pub_sust.htm).

U.S. Environmental Protection Agency QA/G-5 (1998): Guidance for Quality Assurance Project Plans. http://www.epa.gov/quality/qa_docs.html.

U.S. Environmental Protection Agency QA/G-9 (1998): Guidance for the Data Quality Assessment: Practical Methods for Data Analysis. http://www.epa.gov/quality/qa_docs.html.

Venkatram, A. (1982): A framework for evaluating air quality models. *Boundary-Layer Meteor.*, **24**: 371-385.

Venkatram, A. (1983): Uncertainty in predictions from air quality models. *Boundary-Layer Meteor.*, **27**: 185-196.

Venkatram, A. (1988): Inherent uncertainty in air quality modeling. *Atmos. Environ.*, **22**(6): 1221-1227.

Vukovich, F.M. (1997): Time scales of surface ozone variations in the regional, non-urban environment. *Atmos. Environ*., **31**: 1513-1530.

Ward, A.C. (1994): A simple procedure for ranking the performance of several air-quality models across a number of different sites. *Atmos. Environ*., **28**(18): 2909-2915.

Weil, J.C., R.I. Sykes, and A. Venkatram (1992): Evaluating air-quality models: review and outlook. *J. Appl. Meteor.*, **31**: 1121-1145.

Weil, J.C., L.A. Corio, and R.P. Brower (1997): A PDF dispersion model for buoyant plumes in the convective boundary layer. *J. Appl. Meteor*., **36**: 982-1003.

Willmot, C.J. (1982): Some comments on the evaluation of model performance. *Bull. Amer. Meteor. Soc*., **63**: 1309-1313.

Zannetti, P., and P. Switzer (1979): Some problems of validation and testing of numerical air pollution models. Proceedings, Fourth Amer. Meteor. Soc. Symp. on Turbulence, Diffusion, and Air Pollution, January, Reno, Nevada, pp. 405-410.

Zurbenko, I.G. (1986): *The spectral analysis of time series.* North Holland.