Meteorol Atmos Phys 87, 167–196 (2004)
DOI 10.1007/s00703-003-0070-7

Meteorology and Atmospheric Physics
Printed in Austria

School of Computational Sciences, George Mason University, Fairfax, VA, USA

# Air quality model performance evaluation

**J. C. Chang** and **S. R. Hanna**

Received April 16, 2003; accepted September 22, 2003
Published online: June 2, 2004 

**Summary**

This paper reviews methods to evaluate the performance of air quality models, which are tools that predict the fate of gases and aerosols upon their release into the atmosphere. Because of the large economic, public health, and environmental impacts often associated with the use of air quality model results, it is important that these models be properly evaluated.

A comprehensive model evaluation methodology makes use of scientific assessments of the model technical algorithms, statistical evaluations using field or laboratory data, and operational assessments by users in real-world applications. The focus of the current paper is on the statistical evaluation component. It is important that a statistical model evaluation exercise should start with clear definitions of the evaluation objectives and specification of hypotheses to be tested. A review is given of a set of model evaluation methodologies, including the BOOT and the ASTM evaluation software, Taylor's nomogram, the figure of merit in space, and the CDF approach. Because there is not a single best performance measure or best evaluation methodology, it is recommended that a suite of different performance measures be applied. Suggestions are given concerning the magnitudes of the performance measures expected of "good" models. For example, a good model should have a relative mean bias less than about 30% and a relative scatter less than about a factor of two.

In order to demonstrate some of the air quality model evaluation methodologies, two simple baseline urban dispersion models are evaluated using the Salt Lake City Urban 2000 field data. The importance of assumptions concerning details such as minimum concentration and pairing of data are shown. Typical plots and tables are presented, including determinations of whether the difference in the relative mean bias between the two models is statistically significant at the 95% confidence level.

## 1. Introduction and background review

Air quality models are powerful tools to predict the fate of pollutant gases or aerosols upon their release into the atmosphere. The models account for the dilution effects of wind speed and turbulent diffusion. Pollutants can be contaminants routinely emitted by industrial sources (such as the sulfur dioxide emissions from power plants), hazardous chemicals released due to accidents (such as the rupture of a railroad car containing chlorine), or chemical and biological warfare agents disseminated by weapon systems. It is imperative that these dispersion models be properly evaluated with observational data before their predictions can be used with confidence, because the model results often influence decisions that have large public-health and economic consequences.

There can be three components to the evaluation of air quality models: scientific, statistical, and operational. In a scientific evaluation, the model algorithms, physics, assumptions, and codes are examined in detail for their accuracy, efficiency, and sensitivity. This exercise usually requires in-depth knowledge of the model. For statistical evaluation, model predictions (such as concentrations and cloud widths) are examined to see how well they match observations. It is possible for a model to produce the right answers, but as a result of compensating errors.

The operational evaluation component mainly considers issues related to the user-friendliness of the model, such as the user's guide, the user interface, error checking of model inputs, diagnostics of interim (or internal) model calculations, and processing and display of model outputs. The focus of this paper is mainly on statistical model evaluation.

Dispersion is primarily controlled by turbulence in the atmospheric boundary layer. Turbulence is random by nature and thus cannot be precisely described or predicted, other than by means of basic statistical properties such as the mean and variance. As a result, there is spatial and temporal variability that naturally occurs in the observed concentration field. On the other hand, uncertainty in the model results can also be due to factors such as errors in the input data, model physics, and numerical representation. Because of the effects of uncertainty and its inherent randomness, it is not possible for an air quality model to ever be ''perfect'', and there is always a base amount of scatter that cannot be removed.

Decision makers should be able to make use of available information on air quality model performance. An example of a decision maker is a person in a regulatory agency who uses the results of photochemical models to decide emission control strategies which clearly will have large economic and social impacts. As another example, for an accidental release of hazardous materials, emergency responders need to use the model results to decide which neighborhoods to evacuate. Or, when under attack by chemical or biological weapons in a battlefield, a military commander would need to decide whether to order troops to put on protective gear, or to order evacuation based on the model results.

Model evaluation methods and the issue of model uncertainty are further reviewed in the following. Most of the materials presented in this paper, except for the test case application discussed in Sect. 3, are based on a chapter of the lead author's Ph.D. thesis (Chang, 2002).

### *1.1 Model evaluation definitions and methods*

The goal of a model evaluation study must first be well-defined. Different goals could lead to different variables to evaluate and different performance measures to use. For example, for the Environmental Protection Agency (EPA) regulatory applications to standard pollutants such as $SO_2$, the highest and second highest hourly-average concentrations at ground level, for all hours of a year, are typically of interest. In this case, air quality models are evaluated to find out whether they can correctly predict the high end of the concentration distribution, preferably for the right reasons of course. Whether the locations of high concentrations are correctly predicted is usually less important for these regulatory applications.

For military applications, on the other hand, it is typically the hazard areas, as defined by the contours at ground level of certain concentration or dosage (concentration integrated over time) thresholds, that are of interest. In this case, a model should correctly predict both the location and the size of a hazard area. In contrast to regulatory applications, it may be more important for a model in this case to correctly predict the middle or even the low end of the concentration or dosage distribution.

Dosages are also important in EPA (2002) assessments of the effects of toxic pollutants such as benzene. These assessments look at annual average concentrations over broad areas and integrate over the population distribution to calculate health effects such as numbers of excess cancers. The relative impacts of various toxic chemicals are then compared to enable decision makers to prioritize nation-wide emissions control strategies.

The terms *verification* and *validation* (*V&V*) are often used, especially by the military community, to describe activities aimed to demonstrate the credibility of numerical models. However, according to Oreskes et al (1994), verification and validation of numerical models of natural systems are impossible, because natural systems are never closed and because model solutions are always non-unique. The random nature of the process leads to a certain irreducible inherent uncertainty. Oreskes et al (1994) suggest that models can only be confirmed or evaluated by the demonstration of good agreement between several sets of observations and predictions. Following this guidance, the term *evaluation* is used instead of *verification* throughout this paper.

Different model evaluation methodologies have been recommended and developed for various disciplines (e.g., air quality models, water quality models, weather and climate prediction models). Depending on the intended goals associated with each discipline, the focus and approach are also different.

Fox (1984), Hanna (1989), Hanna et al (1991; 1993) and ASTM (2000) propose some comprehensive model performance measures for air quality and dense-gas dispersion models. Measures such as the fractional bias, the normalized mean square error, the geometric mean, the geometric variance, the correlation coefficient, and the fraction of predictions within a factor of two of observations are suggested (see Sect. 2 for more details). The bootstrap resampling method (Efron, 1987) is used by Hanna (1989) and ASTM (2000) to estimate the confidence limits on the performance measures. The methodology has been traditionally used to study whether air quality models can correctly predict quantities such as the maximum concentration over the entire receptor network, the maximum concentration across a given sampling arc, the concentration integrated on a given sampling arc, or the cloud width on a given sampling arc. However, the methodology is generic and is also applicable to other quantities such as wind speed, temperature, or even stock price, as long as predictions and observations are paired (e.g., in space only, in time only, or in both space and time). The Hanna et al (1991; 1993) methodology has been widely used in the air quality modeling community. For example, in its *Initiative on Harmonization Within Atmospheric Dispersion Modeling for Regulatory Purposes*, the National Environmental Research Institute (NERI) of Denmark developed the Model Validation Kit (MVK) based on the methodology (see Olesen, 2001, and www.harmo.org).

Traditionally, model predictions are directly compared to observations. As described by John Irwin, who is the lead author of the American Society of Testing and Materials (ASTM, 2000) model evaluation guidelines, this direct comparison method may cause misleading results, because (1) air quality models almost always predict ensemble means, but observations represent single realizations from an infinite ensemble of cases under the same conditions; and (2) the uncertainties in observations and model predictions arise from different sources. For example, the uncertainty in observations may be due to random turbulence in the atmosphere and measurement errors, whereas the uncertainty in model predictions may be due to input data errors and model physics errors. Therefore, an alternative approach has been proposed by the ASTM to compare observations and model predictions for dispersion models (ASTM, 2000). The approach calls for properly averaging the observations before comparison.

Evaluation methods for regional Eulerian grid models for photochemical pollutants or for fine particles have developed along a slightly different path. For example, Seigneur et al (2000) describe guidance for the performance evaluation of regional Eulerian modeling systems for particulate matter (PM) and visibility, where four levels of evaluation efforts are suggested, including *operational*, *diagnostic*, *mechanistic*, and *probabilistic*. An operational evaluation uses statistical performance measures to test the ability of the model to estimate PM concentrations or other quantities used to characterize visibility. A diagnostic evaluation tests the ability of the model to predict the components of PM or visibility, including PM precursors and associated oxidants, particle size distribution, temporal and spatial variations, and light extinction. A mechanistic evaluation tests the ability of the model to predict the response of PM concentrations or visibility to changes in meteorology and emissions. A probabilistic evaluation takes into the account the uncertainty associated with model predictions and observations of PM and visibility. A set of statistical measures for model evaluation is also recommended by Seigneur et al (2000). In a similar effort, McNally and Tesche (1993) developed the Model Performance Evaluation, Analysis, and Plotting Software (MAPS) to evaluate three-dimensional urban- and regional-scale meteorological, emissions, and photochemical models.

The discipline of numerical weather prediction (NWP) also has a long history of evaluating NWP models (e.g., Pielke and Pearce, 1994; Pielke, 2002; Seaman, 2000). Hamill et al (2000) suggest that a suite of performance measures should be considered to judge the performance of an analysis or forecast made by

mesoscale meteorological models. A single evaluation metric is generally not adequate since it provides only unique information on model performance.

Many sophisticated dispersion modeling systems now routinely use outputs from NWP models. For example, the Third-Generation Air Quality Modeling System (Models-3; EPA 1999) of the Environmental Protection Agency (EPA) is coupled to the Pennsylvania State University-National Center for Atmospheric Research (PSU-NCAR) Fifth-Generation Mesoscale Model (MM5; Grell et al, 1994). Similarly, the National Atmospheric Release Advisory Center (NARAC) system (Nasstrom et al, 2000) of the Department of Energy uses the outputs from the Naval Research Laboratory (NRL) Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS; Hodur, 1997).

This close integration between meteorological and dispersion models is quite promising for the future development of air quality modeling. However, there is a noteworthy irony that mesoscale models have been traditionally evaluated when there are significant weather systems to speak of, in which case the winds tend to be moderate to strong. On the contrary, dispersion or air quality modeling is usually more concerned with the so-called worst-case scenario (i.e., leading to the highest pollutant concentrations) when the winds tend to be light and variable. The study by Hanna and Yang (2001) is among the first to systematically evaluate mesoscale meteorological model predictions of near-surface winds, temperature gradients, and mixing depths, quantities that are most crucial to dispersion modeling applications.

### *1.2 Model uncertainty definitions and methods*

As previously mentioned, because of random turbulence in the atmospheric boundary layer, observed concentrations are expected to fluctuate around some sort of average (e.g., Wilson, 1995). The uncertainty will increase for a smaller spatial scale or a shorter averaging period, and is approximately equal to the mean concentration for averaging times of 10 seconds or less. Venkatram (1979) developed an empirical formula for the expected deviation of observed concentrations from ensemble means, and estimated the deviation to be larger for shorter sampling time and under unstable conditions. Some advanced dispersion models such as SCIPUFF (Sykes et al, 1998) were developed based on higher-order turbulence closure schemes, so that basic statistical properties (e.g., mean and variance) of concentration predictions can be estimated.

According to Fox (1984), Anthes et al (1989), Hanna et al (1991), and Beck et al (1997), uncertainty in air quality modeling generally can be due to (1) variability because of random turbulence, (2) input data errors, and (3) errors and uncertainties in model physics. It has already been mentioned that there are natural fluctuations in the concentration fields due to random turbulence in the atmosphere. Input data errors can be due to uncertain source terms, instrument errors, unrepresentative instrument siting, and many other factors. Furthermore, there might be errors in the physical formulations of a model. Even if these physical formulations were correct, there are still uncertainties in the parameters used in the formulations.

The current paper does not go into detailed analyses of air quality model uncertainty. This rapidly-growing field is following guidelines developed for other types of environmental models. For example, Morgan and Henrion (1990) and Helton (1997) suggest formal statistical frameworks for uncertainty analysis. Helton (1997) recommends that there are two kinds of uncertainty that should be differentiated, *stochastic* and *subjective*. Cullen and Frey (1998) focus on the uncertainty issue in the field of environmental exposure assessment. Saltelli et al (2000) provide a comprehensive collection of methodologies for sensitivity analysis, which is closely related to uncertainty analysis. In traditional literature on meteorology, uncertainty is often synonymous to predictability, i.e., the sensitive dependence of the solution of a nonlinear system on its initial conditions (Lorenz, 1963).

## 2. Description of statistical model performance evaluation methods

This section describes several statistical approaches used in air quality model performance evaluation, and is mainly based on a chapter of the lead author's Ph.D. thesis (Chang,

2002). The approaches and the resulting BOOT model evaluation software package represent the work previously developed by the authors, but with additional methods suggested for interpretation of the results (Sects. 2.1–2.4). Four innovative evaluation methodologies (Sects. 2.5–2.8) developed by other researchers are also described, and the existing BOOT software is further upgraded to include three of these new capabilities (Sects. 2.5–2.7).

It is important that air quality models be properly evaluated in order to demonstrate their fidelity in simulating the phenomena of interest. In the following, a set of procedures for the statistical evaluation of air quality models is described. Even though the emphasis here is somewhat biased towards the air quality branch of the environmental sciences, the same procedures and approaches are also applicable to other disciplines and models. For example, the BOOT model evaluation software could just as well be applied to the question of whether a certain drug produces a significant improvement in health, or to whether a certain new type of corn seed produced significantly higher corn yields that the old type of seed.

For statistical evaluation, model predictions are evaluated against some reference states, which in most cases are simply ''observations.'' Observations can be directly measured by instruments, or are themselves products of other models or analysis procedures. It is important to recognize that different degrees of uncertainty are associated with different types of observations. Furthermore, it is important to define how predictions are to be compared with observations. For example, should observations and predictions be paired in time, in space, or in both time and space? Different conclusions can be reached depending on the type of pairing chosen.

### *2.1 Definition of evaluation objective*

The evaluation objective must be clearly defined before doing any model performance evaluation. For example, for EPA regulatory applications, the primary objective might be how well an air quality model simulates the maximum one-hour averaged concentration anywhere on the sampling network. In this case, the location of the maximum impact is less important. For environmental justice applications, it might be important to evaluate model predictions of 24 hour averaged PM at specific locations such as heavily populated areas with poor housing. For military applications, the location of the dosage footprint of a chemical agent cloud is an important piece of information. For flammable substances, the instantaneous maximum concentration is more important than the one-hour average concentration. For a forensic study involving decisions concerning population impacts, it might be of interest to correctly estimate the cloud arrival and departure times. Moreover, for any actual field experiment, there are some practical constraints that lead to only a limited number of the above evaluation objectives that can actually be considered.

When conducting a statistical test, a null hypothesis must first be defined. Depending on the goals and emphases of the study, there are a number of potential outputs of an air quality model that could be evaluated, such as

(1) For a given averaging time:
   - The overall maximum concentration over the entire domain,
   - The maximum concentration along a sampling line,
   - The cross-line integrated concentration along a sampling line,
   - The location of a contour (i.e., cloud footprint) for a certain concentration threshold (e.g., toxicity limit or flammability limit),
   - The cloud width along a sampling line,
   - The cloud height along a vertical tower.

(2) Dosage (integration of concentration over time):
   - The maximum dosage (concentration integrated with time) along a sampling line,
   - The cross-wind integrated dosage along a sampling line,
   - The total dosage over an area,
   - The location of a contour for a certain dosage.

(3) Cloud timing:
   - The cloud arrival and departure times and the effective velocity.

Once the model output that is to be evaluated is decided, it is useful to pose a hypothesis to be

tested by the statistical evaluation. Such a hypothesis might be, for example, ''Can we say with 95% confidence that the model's mean bias is not significantly different from zero?'' Or, ''Can we say with 95% confidence that the Normalized Mean Square Error of Model 1 is not significantly different from that for Model 2?'' Statistical model evaluation software such as BOOT or the ASTM method are capable of addressing these questions.

### *2.2 Exploratory data analysis*

Before beginning the calculation of various statistical performance measures, it is extremely useful to perform exploratory data analysis by simply plotting the data in different ways. The human eye can often glean more valuable insights from these plots than pure statistics, especially since many of the statistical measures depend on linear relations. These plots can also provide clues as to why a model performs in a certain way. Some of the commonly-used plots are scatter plots, quantile–quantile plots, residual box plots, and simply plots of predictions and observations as a function of time or space. Some of these plots are demonstrated in Sect. 3 with data from Intensive Operating Period 9 (IOP09) from the Salt Lake City Urban 2000 field data.

Scatter plot: In a scatter plot, the paired observations and predictions are plotted against each other. Visual inspection can reveal the magnitude of the model's over or under-predictions. Also, as implied by its name, the scatter of the points can be quickly seen and estimated by eye (factor of 2, 5, or 10?). Because of the obvious impacts on the public health due to high pollutant concentrations or dosages, the high end of the plots can be studied. On the other hand, correct predictions of low concentrations may sometimes also be important for highly toxic chemicals.

Quantile–quantile plot: The quantile–quantile plot begins with the same paired data as the scatter plots, but removes the pairing and instead ranks each of the observed and predicted data separately from lowest to highest. Thus the 3rd lowest predicted concentration would be plotted versus the 3rd lowest observed concentration. It is often of interest to find out whether a model can generate a concentration distribution that is similar to the observed distribution. Biases at low or high concentrations are quickly revealed in this plot.

Residual plots employing box diagrams: The scatter and quantile–quantile plots mentioned above clearly do not provide a complete understanding of the physical reasons why a model performed in a certain way. The issue can be addressed using residual analyses, and combined with box diagrams if necessary. In this plot, model residuals, defined as the ratio of predicted ($C_p$) to observed ($C_o$) concentrations (or dosages or other outputs) are plotted, in the form of a scatter plot, versus independent variables such as hour of day, downwind distance, ambient wind speed, mixing height, atmospheric stability. If there are many points, it is not effective to plot all of them, and instead the residuals are binned according to different ranges of independent variables, and the distribution of all data points in each bin is represented by a box diagram. The significant points for each box diagram represent the 2nd, 16th, 50th, 84th, and 98th percentiles of the cumulative distribution of the n points considered in the box. A good performing model should not show any trend of the residuals when they are plotted versus independent variables.

### *2.3 Quantitative performance measures implemented in BOOT software*

Hanna et al (1991; 1993) recommended a set of quantitative statistical performance measures for evaluating models, and implemented the procedures in a software package called BOOT. The performance measures have been widely used in many studies (e.g., Ichikawa and Sada, 2002; Nappo and Essa, 2001; Mosca et al, 1998), and have been adopted as a common model evaluation framework for the European Initiative on ''Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes'' (Olesen, 2001).

In addition to the standard statistical procedures, the BOOT software includes the capability to produce the scatter, quantile–quantile, and residual plots used in the exploratory data analysis described in the previous subsection. The quantitative performance measures used in the BOOT procedure are described below, together with some recent enhancements.

### 2.3.1 Definitions and properties of performance measures

In order to evaluate the predictions of a model with observations, Hanna et al (1991; 1993) recommend the use of the following statistical performance measures, which include the fractional bias (FB), the geometric mean bias (MG), the normalized mean square error (NMSE), the geometric variance (VG), the correlation coefficient (R), and the fraction of predictions within a factor of two of observations (FAC2):

$$\mathrm{FB} = \frac{(\overline{C_o} - \overline{C_p})}{0.5(\overline{C_o} + \overline{C_p})}, \tag{1}$$

$$\mathrm{MG} = \exp(\overline{\ln C_o} - \overline{\ln C_p}), \tag{2}$$

$$\mathrm{NMSE} = \frac{\overline{(C_o - C_p)^2}}{\overline{C_o}\,\overline{C_p}}, \tag{3}$$

$$\mathrm{VG} = \exp[\overline{(\ln C_o - \ln C_p)^2}], \tag{4}$$

$$\mathrm{R} = \frac{\overline{(C_o - \overline{C_o})(C_p - \overline{C_p})}}{\sigma_{C_p}\sigma_{C_o}}, \tag{5}$$

$$\mathrm{FAC2} = \text{fraction of data that satisfy } 0.5 \le \frac{C_p}{C_o} \le 2.0, \tag{6}$$

where:

| | |
|---|---|
| $C_p$: | model predictions, |
| $C_o$: | observations, |
| overbar($\overline{C}$): | average over the dataset, and |
| $\sigma_C$: | standard deviation over the dataset. |

A perfect model would have MG, VG, R, and FAC2 = 1.0; and FB and NMSE = 0.0. Of course, as noted earlier, because of the influence of random atmospheric processes, there is no such thing as a perfect model in air quality modeling. Note that since FB and MG measure only the systematic bias of a model, it is possible for a model to have predictions completely out of phase of observations and still have FB = 0.0 or MG = 1.0 because of canceling errors.

The six performance measures defined above are by no means exhaustive. Depending on the purpose and emphasis of a study, other measures can be defined and can be easily incorporated into the BOOT software. The confidence limits on these alternate performance measures can be calculated by the same algorithms in BOOT that are used to calculate the confidence limits on the six standard performance measures.

In the above discussions, it is simply assumed that the "dataset" contains pairs of $C_p$ and $C_o$ (or dosages or other model outputs), and that they represent averages over an averaging time, $T_a$. BOOT allows sets of $C_p$ for several alternate models. The pairing is completely generic, and can be:

- Pairing in time only, such as the time series of the maximum pollutant concentrations anywhere in the domain of interest (i.e., no penalty is given if the model predicts the maximum concentration at a wrong place);
- Pairing in space only, such as the spatial distribution of the maximum pollutant concentrations over a time period (i.e., no penalty is given if the model predicts the maximum concentration at a wrong time);
- Pairing in both time and space.

Pairing in both time and space is clearly most stringent. Concentration fields often exhibit complex spatial patterns. As Weil et al (1992) point out, because of typical variations in wind direction of 20 to 40 degrees (or more in light winds), predicted plumes often completely fail to overlap observed plumes, even though the magnitudes and patterns may be similar. This difficulty of separating the effects of winds and plume dispersion is a common challenge for model evaluation.

Multiple performance measures should be applied and considered in any model evaluation exercise, as each measure has advantages and disadvantages and there is not a single measure that is universally applicable to all conditions. The relative advantages of each performance measure are partly determined by the characteristics and distributions of the model predictions and observations. The distribution is close to lognormal for most atmospheric pollutant concentrations. In this case, the linear measures FB and NMSE may be overly influenced by infrequently occurring high observed and/or predicted concentrations, whereas the logarithmic measures MG and VG may provide a more balanced treatment of extreme high and low values. Therefore, for a dataset where both

predicted and observed concentrations vary by many orders of magnitude, MG and VG may be more appropriate. FAC2 is the most robust measure, because it is not overly influenced by outliers.

MG and VG may be overly influenced by extremely low values, near the instrument thresholds, and are undefined for zero values. These low and zero values are not uncommon in dispersion modeling, where a low concentration value might occur at a receptor that the plume has missed. Therefore, when calculating MG and VG, it is recommended that a minimum threshold be assumed for data values. For example, the instrument threshold, such as the limit of quantitation (LOQ), could be used as the lower bound for both $C_p$ and $C_o$. The sensitivity of MG and VG to various assumptions regarding minimum values should be determined as part of the evaluation exercise, which will be later demonstrated with the Urban 2000 field data.

FB and MG are measures of mean relative bias and indicate only systematic errors, whereas NMSE and VG are measures of mean relative scatter and reflect both systematic and unsystematic (random) errors. For FB, which is based on a linear scale, the systematic bias refers to the arithmetic difference between $C_p$ and $C_o$; and for MG, which is based on a logarithmic scale, the systematic bias refers to the ratio of $C_p$ to $C_o$. Because FB is based on the mean bias, it is possible for a model whose predictions are completely out of phase with observations to still have a $FB = 0$ because of compensating errors. An alternative is to consider a slightly modified version of FB where the two error components (i.e., overprediction and underprediction) are separately treated. This approach is described in detail in Sect. 2.6.

The correlation coefficient, R, reflects the linear relationship between two variables and is thus insensitive to either an additive or a multiplicative factor. Because of the linear implication, if there is a clear non-linear relation (say, a parabolic relation) in the scatter plots, it is not revealed by R. Also, R is sensitive to a few aberrant data pairs. For example, a scatter plot might show generally poor agreement; however, the presence of a good match for a few extreme pairs will greatly improve R. As a result, Willmott (1982) discourages the use of R, because it does not consistently relate to the accuracy of predictions. It is often suggested that the more robust ranked correlation, $R_{rank}$ (or often called the Spearman correlation coefficient), be considered, where the ranks of $C_p$ and $C_o$ are correlated instead of their values. The $R_{rank}$ is not influenced by extreme outliers.

Moreover, it is typical for short-range dispersion field experiments to have concentration data measured along concentric arcs at several discrete distances from the source, and it is also customary to evaluate model performance based on the maximum concentration or the cross-line integrated concentration along each sampling arc. In this case, the value of R can be deceivingly high, mainly reflecting the fact that concentration decreases with downwind distance, which any reasonable dispersion model will simulate. Therefore, the correlation coefficient is less useful in a typical evaluation exercise for dispersion models. On the other hand, it might be more useful when gridded fields are involved (e.g., McNally and Tesche, 1993).

Since NMSE accounts for both systematic and random errors, it is helpful to partition NMSE into the component due to systematic errors, $NMSE_s$, and the unsystematic component due to random errors, $NMSE_u$. It can be shown that

$$NMSE_s = \frac{4FB^2}{4 - FB^2}. \tag{7}$$

The above expression gives the minimum NMSE, i.e., without any unsystematic errors, for a given value of FB (Hanna et al, 1991). The total NMSE is the sum of $NMSE_s$ and $NMSE_u$.

Similarly, VG can also be partitioned into the systematic component, $VG_s$, and the random (unsystematic) component, $VG_u$. The systematic component of VG is given by

$$VG_s = \exp(\overline{\ln C_o} - \overline{\ln C_p})^2 = \exp\left((\ln MG)^2\right) \tag{8}$$

or,

$$\ln VG_s = (\ln MG)^2. \tag{9}$$

The above equation gives the minimum possible VG, without any unsystematic errors, given a value of MG (Hanna et al, 1991). The total VG is the product of $VG_s$ and $VG_u$. Or, the total $(\ln VG)$ is the sum of $(\ln VG_s)$ and $(\ln VG_u)$.

#### 2.3.2 Relations between FB and NMSE, and MG and VG

The FB, MG, NMSE, and VG defined in Eqs. (1)–(4) can be used to quantitatively define a certain aspect of model performance. However, direct quotations of their values are often not that informative. For example, without experience, it will be difficult for a user to discern the meaning of, say, NMSE = 9 and VG = 13. It is shown below how the values of FB, NMSE, MG, and VG can be further interpreted in terms of a measure that is more easily comprehended, such as the equivalent ratio of $C_p$ to $C_o$.

For example, Eq. (1) can be expressed as

$$\frac{\overline{C_p}}{\overline{C_o}} = \frac{1 - \frac{1}{2}FB}{1 + \frac{1}{2}FB}. \tag{10}$$

Therefore, FB = 0.67 would imply a factor of two mean underprediction, and FB = −0.67 would imply a factor of two mean overprediction.

MG is simply the ratio of the geometric mean of $C_o$ to the geometric mean of $C_p$:

$$\frac{\langle C_p \rangle}{\langle C_o \rangle} = \frac{1}{MG}, \tag{11}$$

where the angle brackets indicate a geometric mean. Consequently, a factor of two mean bias would imply that MG = 0.5 or 2.0, and a factor of four mean bias would imply that MG = 0.25 or 4.0.

To interpret NMSE, assume that the mean of the observed concentrations equals the mean of the predicted concentrations. Then NMSE = 1.0 implies that the root-mean-square-error is equal to the mean. As NMSE becomes much larger than 1.0, it can be inferred that the distribution is not normal but is closer to log-normal (e.g., many low values and a few large values).

VG expresses the scatter of a log-normal distribution, which can be expressed as, say, ''plus or minus 20%'', or ''plus or minus a factor of 10''. For example, a factor of 2 scatter would imply a VG = 1.6 and a factor of 5 scatter would imply VG = 12.

### *2.4 Confidence limits on performance measures*

A model might appear to have a certain amount of skill as reflected by, for example, a small fractional bias, FB. A model might also appear to have a better performance than other models based on, for example, a smaller FB. However, these results may not be significant in a statistical sense. To investigate the significance question, there are several hypotheses that could be tested. Two examples are:

- When compared to observations, is a model's relative mean bias performance measure, FB, significantly different from zero at the 95% confidence levels? For the case of geometric measures such as MG, it is necessary to consider whether MG is significantly different from 1.0, since a perfect model has MG = 1.0.
- When comparing the performance of two models, are the differences in the performance measures for the two models (e.g., FB for Model 1 minus FB for Model 2) significantly different from zero at the 95% confidence levels?

If the distribution of the quantity follows a normal distribution or can be transformed to a normal distribution, then significance tests such as the student-t test can be applied to the problem. However, in the case of general distributions, then random resampling methods such as the Bootstrap method (Efron, 1987; Efron and Tibshirani, 1993) can be used to estimate the mean, $\mu$, and standard deviation, $\sigma$, of the distribution of each performance measure. Each random sample will yield one estimate of the performance measure. After, say, 1000 samples, there will be 1000 estimates of any given performance measure. These 1000 estimates can be used to estimate the $\mu$ and $\sigma$ for the performance measure. As suggested by Efron (1987), the 95% confidence intervals are given by

$$\mu \pm t_{95\%}\, \sigma \left(\frac{n}{n-1}\right)^{1/2}, \tag{12}$$

where n is the number of resamples, $t_{95\%}$ is the Student's t value at the 95% confidence level with NP − 1 degrees of freedom, and NP is the number of observation-prediction pairs. Alternatively, the 1000 estimates could be used to directly determine the 95% confidence intervals based on the 5th and 95th percentiles of the distribution. The rule of thumb on the minimum number of observation-prediction pairs is around

20, so that meaningful information on the confidence level can be obtained.

Some practical considerations concerning resampling are given below:

- The evaluation dataset may sometimes appear in blocks. For example, one block of data may be from one experiment trial and another block of data may be from a second experiment trial. Or, one block of data may be from one monitoring arc, and another block of data may be from a second monitoring arc. It may also be appropriate to block data by other independent variables such as wind speed and stability class. In this case, resampling should be restricted to within each block; otherwise, artificial block-to-block variance will be introduced.
- Resampling should be done with replacement. That is, once a sample is drawn, it is allowed to be drawn again.
- Observational and predicted values are sampled concurrently, in order to maintain the relationship between them.

Because the blocking procedure is clearly arbitrary and can have an effect on the resulting confidence limits, some experimentation with one or two options is advised in any operational evaluation.

### *2.5 Taylor's single nomogram method*

Taylor (2001) and Gates et al (1999) recommend a nomogram that can summarize three performance metrics "in a single diagram". The method was first developed for applications to weather forecast and climate models, but also has potential for other types of models, such as air quality models. The three performance measures used by Taylor are the normalized standard deviation (NSD), the normalized root mean square error (NRMSE), and the correlation coefficient (R). NSD and NRMSE are defined below, whereas R has been previously defined by Eq. (5),

$$\mathrm{NSD} = \frac{\sigma_{C_p}}{\sigma_{C_o}}, \tag{13}$$

$$\mathrm{NRMSE} = \frac{\sqrt{\overline{[(C_p - \overline{C_p}) - (C_o - \overline{C_o})]^2}}}{\sigma_{C_o}}, \tag{14}$$

where all variables on the right hand side of Eqs. (14) and (15) are similarly defined as in Eqs. (1)–(6). Note that NSD, NRMSE, and R account for only unsystematic errors, i.e., their values do not depend on mean bias. Therefore, $R = 1.0$, $NSD = 1.0$, and $NRMSE = 0.0$ are only necessary but not sufficient conditions for a perfect model. As a result, if a model systematically overpredicts or underpredicts, but still has the same scatter as the observations, it would yield a perfect NSD of 1.0.

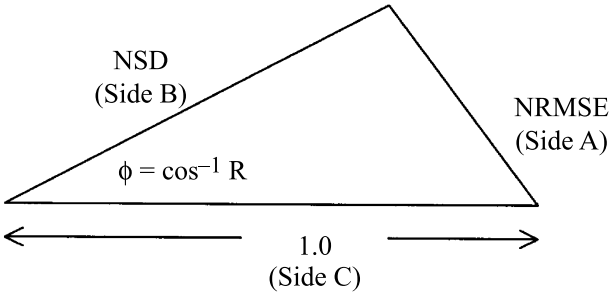
Taylor (2001) shows that a relationship exists among the three performance measures (NSD, NRMSE, and R) based on the law of cosines. It is this relationship that allows the three measures to be plotted in a single diagram. It can be shown that

$$R = \frac{1}{2\,\mathrm{NSD}}(\mathrm{NSD}^2 + 1 - \mathrm{NRMSE}^2). \tag{15}$$

The above equation is in the form of the law of cosines, i.e.,

$$\cos\phi = \frac{1}{2\,BC}(B^2 + C^2 - A^2), \tag{16}$$

where A, B, and C are the three sides of a triangle, and A is the side opposite the angle $\phi$, as depicted by the following diagram:




An example of the nomogram format proposed by Taylor (2001) is given in Fig. 1, where three hypothetical models are used for illustration. A model's performance is indicated by a unique point on the diagram. In Fig. 1, values of constant NSD are indicated by concentric arcs (dashed curves, except that a solid curve is used for $NSD = 1.0$), and values of R are indicated by the polar dial along the edge of the figure. A perfect model ($NSD = R = 1.0$, and $NRMSE = 0.0$) is indicated by the large solid circle. In the figure, Model-A is located closest to the large solid circle, thus the overall best performer.

Taylor's diagram has so far been primarily used for gridded fields. The method should be

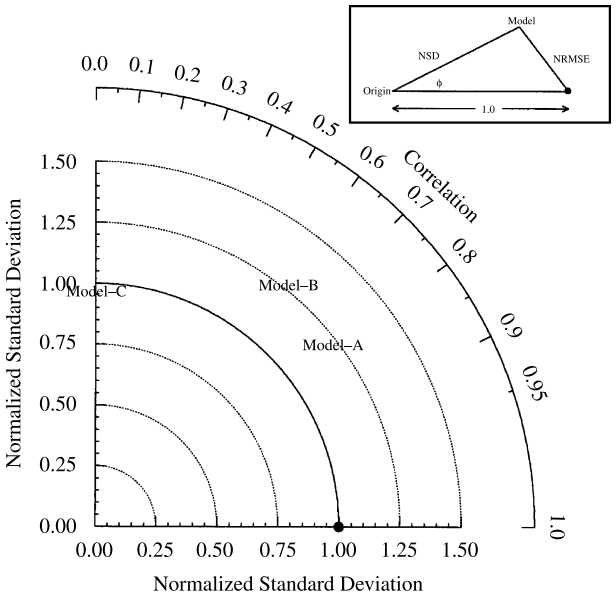



**Fig. 1.** Nomogram proposed by Taylor (2001). The radial distance from the center of the model name to the origin indicates NSD. The distance from the center of the model name to the large solid circle indicates NRMSE. The cosine of the angle formed by the horizontal axis and segment NSD indicates R. The inset further depicts the relationship among NSD, NRMSE, and R explicitly. The large solid circle indicates a perfect model, i.e., NSD = R = 1.0, and NRMSE = 0.0. NSD is measured by concentric arcs (dashed curves, except that a solid curve is used for NSD = 1.0), and R is indicated by the polar dial along the edge of the nomogram. In the above example, Model-A has NSD = 1.2, NRMSE = 0.74, and R = 0.79; Model-B has NSD = 1.25, NRMSE = 1.02, and R = 0.61; and Model-C has NSD = 0.92, NRMSE = 1.36, and R = 0

further reviewed to determine its application to short-range dispersion field experiments where concentrations are typically measured along concentric arcs, and where R is less relevant.

### *2.6 Figure of Merit in Space (FMS) and Measure of Effectiveness (MOE) methods*

Another method that is sometimes used to evaluate model performance is the so-called figure of Merit in Space (FMS), defined as

$$\mathrm{FMS} = \frac{A_p \cap A_o}{A_p \cup A_o}, \tag{17}$$

where $A_p$ is the predicted contour area based on a certain threshold, and $A_o$ is the observed contour area based on the same threshold (Fig. 2). Contour areas can be defined, for example, by a threshold concentration for toxic or flammable materials for hazardous gas models, or by areas of precipitation for weather forecasts. It is clear from the above definition that the FMS does not depend on the detailed distribution within the contour. Therefore, the FMS is more subjective and qualitative than the statistical evaluation procedures described above.

In addition to its routine use to verify precipitation forecasts, the FMS has been used to evaluate the performance of long-range transport and dispersion models for the Chernobyl accident (Klug et al, 1992) and for the European Tracer Experiment (ETEX) (Mosca et al, 1998). It is also often called the threat score (Wilks, 1995; McNally and Tesche, 1993).

The FMS has not often been used for evaluating short-range or mesoscale-range dispersion models. Because of the difficulty in simulating plume directions within 20 to 40 deg, the observed and predicted contours seldom overlap (Weil et al, 1992). Thus, the FMS is usually quite low for plume models applied to point sources, whose concentration contours frequently have a cigar shape.

Emergency response personnel are interested in whether a model is more likely to predict false negatives than false positives. In Fig. 2, the area under $A_o$ but not under $A_p \cap A_o$ can be defined as the area of "false negative" predictions (area $A_{FN}$) for the model, whereas the area under $A_p$ but not under $A_p \cap A_o$ can be defined as the area of "false positive" predictions (area $A_{FP}$) for the model. False negative means that the model did not predict any impact but observations showed otherwise; false positive means that the model predicted impact but observations showed otherwise. Generally, false negatives are more worrisome to emergency responders than false positives. Equation (17) can then be rewritten as

$$\mathrm{FMS} = \frac{A_p \cap A_o}{A_p \cap A_o + A_{FN} + A_{FP}}. \tag{18}$$

One shortcoming with the FMS is that if $A_p$ and $A_o$ are nearly identical in shapes, but do not overlap at all, perhaps due to incorrect wind direction inputs, then the FMS will be zero and no credit will be given to the fact that the model has done a satisfactory job in predicting the shapes.

Warner et al (2001) suggest a more general approach whereby a user can specify the relative importance of the areas of false negative and false positive predictions by inserting weighting factors to $A_{FN}$ and $A_{FP}$ in the above equation.

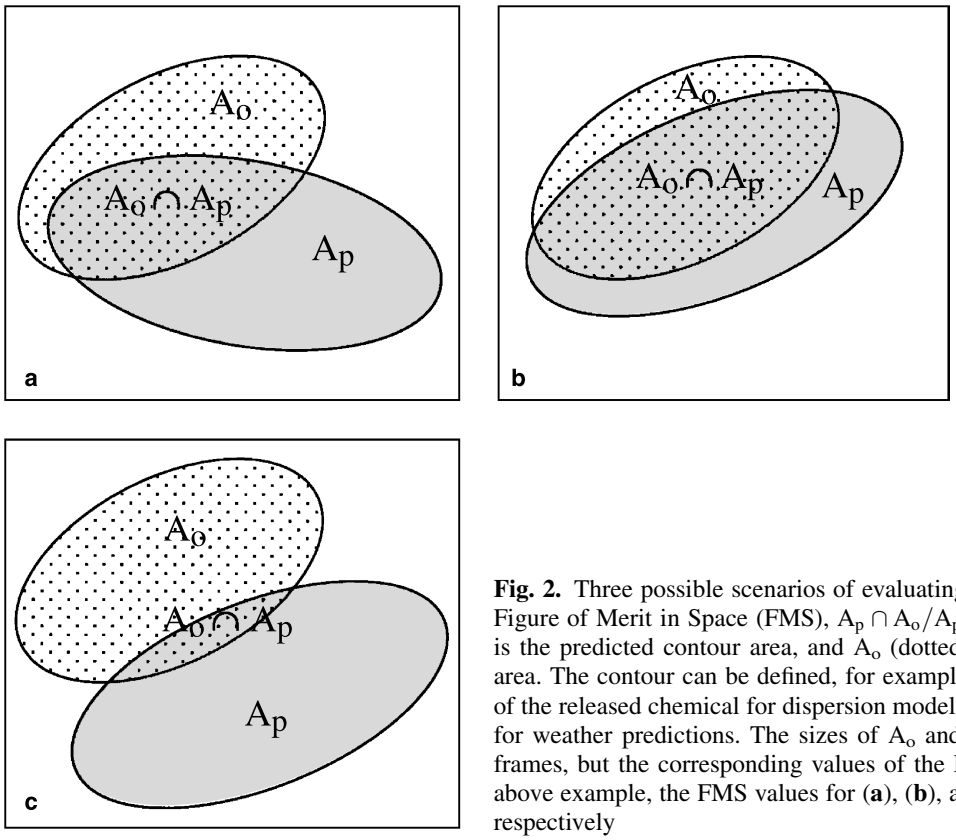



**Fig. 2.** Three possible scenarios of evaluating model performance using the Figure of Merit in Space (FMS), $A_p \cap A_o / A_p \cup A_o$, where $A_p$ (shaded area) is the predicted contour area, and $A_o$ (dotted area) is the observed contour area. The contour can be defined, for example, by a concentration threshold of the released chemical for dispersion modeling, or by areas of precipitation for weather predictions. The sizes of $A_o$ and $A_p$ are the same for all three frames, but the corresponding values of the FMS are quite different. In the above example, the FMS values for (**a**), (**b**), and (**c**) are 0.19, 0.61, and 0.04, respectively

They define the more general metric as the user-oriented one-dimensional (1-D) measure of effectiveness (MOE), i.e.,

$$1-\mathrm{D\ MOE} = \frac{A_p \cap A_o}{A_p \cap A_o + C_{FN}A_{FN} + C_{FP}A_{FP}}, \tag{19}$$

where $C_{FN}$ and $C_{FP}$ are weighting factors or constants to be determined by the user. A family of 1-D MOE estimates can be defined based on different combinations of $C_{FN}$ and $C_{FP}$. The 1-D MOE is the same as the FMS when $C_{FN}$ and $C_{FP}$ both equal 1.0.

Warner et al (2001) also suggest a two-dimensional (2-D) MOE, where two components are used to indicate model performance,

$$\begin{aligned} 2-\mathrm{D\ MOE} &= (\mathrm{MOE_x}, \mathrm{MOE_y}) \\ &= \left(\frac{A_p \cap A_o}{A_o}, \frac{A_p \cap A_o}{A_p}\right) \\ &= \left(\frac{A_p \cap A_o}{A_p \cap A_o + A_{FN}}, \frac{A_p \cap A_o}{A_p \cap A_o + A_{FP}}\right). \end{aligned} \tag{20}$$

Note that the two components are normalized differently, i.e., one by $A_o$ and the other one by $A_p$.

With appropriate configuration, a model can always generate sufficiently-detailed outputs for contouring. However, sufficiently-detailed spatial coverage of observations is needed to permit the accurate drawing of contours to calculate FMS (or MOE). When assessing precipitation forecasts, good coverage is often provided by networks of rain gauges. However, the monitoring instruments for short-range dispersion field experiments are typically arranged in concentric arcs. The data from these arcs are adequate to measure the plume's cross-wind distribution, but are usually insufficient to plot contours. This is especially so if the number of sampling arcs is limited. And in the case of routine air quality monitoring networks, the ten or so monitors are generally scattered around at various angles and distances, and any contours that may be attempted would be highly uncertain.

Warner et al (2001) recommend two alternate methods to compute the MOE under the

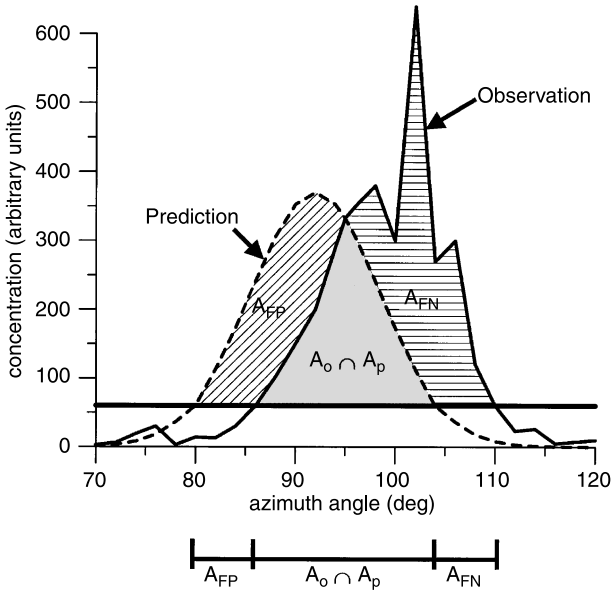



**Fig. 3.** Two methods to estimate the areas required to compute the Figure of Merit in Space (FMS), or the Measure of Effectiveness (MOE), for a typical short-range dispersion field experiment, where concentrations are measured along concentric arcs. Solid curve is observations along a sampling arc, expressed in azimuth angle. Dashed curve is model predictions. Thick solid horizontal line represents a threshold (60 concentration units in this case). Both the FMS and MOE require estimates of the overlap area ($A_p \cap A_o$), the false-positive area ($A_{FP}$), and the false-negative area ($A_{FN}$), which can be given by the following two methods. *Method 1*: The $A_{FP}$, $A_{FN}$, and $A_p \cap A_o$ are determined by the three line segments (or azimuthal distances along the arc) at the bottom of the figure, where $A_{FP}$ indicates the locations where predictions are higher than the threshold and observations are lower than the threshold, $A_{FN}$ is conversely defined, and $A_p \cap A_o$ indicate the locations where both observations and predictions are higher than the threshold. *Method 2*: The $A_{FP}$, $A_{FN}$, and $A_p \cap A_o$ pertain to the cross hatched, horizontal hatched, and shaded areas above the threshold, respectively. In actual implementation, $A_{FP}$, $A_{FN}$, and $A_p \cap A_o$ are simply estimated by the summation of the data

condition where limited data are available (see Fig. 3 for an illustration):

Method 1 applies when sufficient monitoring data are available on a crosswind arc, and these data can be used to define crosswind distance ranges where the concentration (or dosage) exceeds some threshold. In this case, the positions of the predicted and observed distance ranges are used as before to calculate $A_{FP}$, $A_{FN}$, and $A_p \cap A_o$. The difference is that the ''areas'' are now ''line lengths'' in Fig. 3. The MOE that is calculated does not depend on the detailed distributions of observations and predictions. For example, with an assumed concentration threshold of 60, the MOE would remain the same regardless of whether the observed maximum concentration were 650 or 6500 in Fig. 3. It is obvious that a perfect MOE (or FMS) under this method does not necessarily guarantee a perfect agreement between observations and predictions.

Method 2 is applied on a point by point basis, similar to FB or MG, and therefore applies to cases where receptors are not distributed nicely along a sampling arc. It is possible to apply Method 2 to, for example, a set of randomly distributed receptors, as illustrated in Fig. 3. Under Method 2, $A_{FP}$ is given by the sum of the differences between predictions and observations for all those locations where predictions are higher than observations (i.e., false positive), $A_{FN}$ is given by the sum of the differences between predictions and observations for all those locations where predictions are lower than observations (i.e., false negative), and $A_p \cap A_o$ is given by the differences between (1) the smaller of predictions and observations and (2) the threshold at all locations. (Recall that all the above operations are with respect to the threshold value.) For Method 2, a perfect MOE (or FMS) would imply a perfect agreement between observations and predictions.

The BOOT software has incorporated the philosophy of the 2-D MOE determined by Method 2 above by breaking up FB into the false positive ($FB_{fp}$) and false negative ($FB_{fn}$) components (or 2-D FB).

$$FB_{fp} = \frac{0.5[\overline{|C_o - C_p|} + (\overline{C_p} - \overline{C_o})]}{0.5(\overline{C_o} + \overline{C_p})}, \tag{21}$$

where the numerator amounts to considering only those data pairs with $C_p > C_o$ (i.e., overpredicting or false positive).

$$FB_{fn} = \frac{0.5[\overline{|C_o - C_p|} + (\overline{C_o} - \overline{C_p})]}{0.5(\overline{C_o} + \overline{C_p})}, \tag{22}$$

where the numerator amounts to considering only those data pairs with $C_o > C_p$ (i.e., underpredicting or false negative). The difference between $FB_{fn}$ and $FB_{fp}$ yields the original FB,

whereas the sum of $FB_{fn}$ and $FB_{fp}$ is the normalized absolute error (NAE):

$$NAE = \frac{\overline{|C_o - C_p|}}{0.5(\overline{C_o} + \overline{C_p})}. \quad (23)$$

Model evaluation experts such as McNally and Tesche (1993) prefer NAE over NMSE because it is less susceptible to outliers. MG can also be similarly divided into the false positive and false negative components. The following two equations show the relationship between the 2-D FB defined in Eqs. (21) and (22), and the 2-D MOE defined in Eq. (20) and estimated by Method 2 above (Chang, 2002).

$$FB_{fp} = \frac{2 \cdot MOE_x \cdot (1 - MOE_y)}{MOE_x + MOE_y}, \quad (24)$$

$$FB_{fn} = \frac{2 \cdot MOE_y \cdot (1 - MOE_x)}{MOE_x + MOE_y}. \quad (25)$$

### *2.7 ASTM Statistical Model Evaluation method*

The qualitative and quantitative procedures mentioned in the above sections typically involve direct comparisons of model predictions with field observations. However, many studies (e.g., Fox, 1984; Venkatram, 1984 and 1988; Weil et al, 1992) point out that there is a fundamental difficulty in that most dispersion models generate an ensemble-mean prediction (either explicitly or implicitly), whereas an observation corresponds to a single realization of the ensemble. Here, an ensemble is defined as ''a set of experiments corresponding to fixed external conditions'' (Lumley and Panofsky, 1964). Therefore, some researchers have been advocating a framework through which the atmospheric dispersion model predictions can be compared with a grouping of a number of field observations. One such framework was proposed as a standard guide by the American Society for Testing and Materials (ASTM, 2000).

Following the definitions proposed by Venkatram (1984), the basic premise of the ASTM procedure is that a single realization of the observed concentration, $C_o$, can be expressed as

$$C_o = \overline{C_o}(\alpha) + \Delta C_o'(\alpha) + C_o'(\alpha), \quad (26)$$

where $\alpha$ represents the set of model input parameters, $\overline{C_o}(\alpha)$ is the ensemble average of the observations (which dispersion models are ideally supposed to predict), $\Delta C_o'(\alpha)$ represents measurement errors due to calibration or unrepresentative siting, and $C_o'(\alpha)$ represents stochastic fluctuations due to turbulence.

Similarly, the concentration, $C_p$, predicted by most models can be considered to have the following three components:

$$C_p = \overline{C_p}(\alpha) + \Delta C_p'(\alpha) + C_p'(\alpha), \quad (27)$$

where $\overline{C_p}(\alpha)$ is the ensemble average predicted by most models; $\Delta C_p'(\alpha)$ represents the errors due to model input uncertainty; and $C_p'(\alpha)$ represents errors due to factors such as incorrect model physics, unrepresentativeness (such as comparing grid-volume averages with point measurements), and parameters (other than $\alpha$) not accounted for by the model.

The problem of directly comparing observations (which are sets of realizations of ensembles) with model predictions (which are ensemble averages) arises because $C_o$ is compared with $\overline{C_p}(\alpha)$. The ASTM (2000) procedure argues that if the effects of $\Delta C_o'(\alpha)$, $C_o'(\alpha)$, $\Delta C_p'(\alpha)$, and $C_p'(\alpha)$ all somehow average to zero, then it is more appropriate to first separately average observations and modeling results over a number of regimes (or ensembles), which can be defined by independent factors such as downwind distance and stability parameter, and then do the comparison. The goal is to group experiments into regimes of similar conditions. Averaging observations over each of these regimes provides an estimate of what most dispersion models attempt to predict, $\overline{C_o}(\alpha)$. These regime averages of observations and predictions can then be paired to calculate, for example, those performance metrics defined in Sect. 2.3. Like the BOOT software, the ASTM procedure uses the bootstrap resampling technique to calculate the confidence limits of the performance metrics, where resampling is done within each regime, and where the observed and predicted values for the same experiment are sampled concurrently.

Strictly speaking, the calculation of an observed ensemble would require that many experiments be conducted under *identical* conditions. This objective is impossible to achieve in a

field program. Therefore, the regime-average represents a surrogate of a true ensemble average, and is justified because each regime consists of experiments conducted under *similar* conditions.

The ASTM (2000) procedure was initially developed with short-range dispersion field experiments in mind, but can be extended to other types of experiments with appropriate considerations. Traditionally, short-range dispersion experiments have receptors arranged in concentric arcs to maximize plume capture. Previous researchers have often used the centerline concentration to assess model performance. (This was partly motivated by regulatory requirements.) In addition to providing the rationale for the need to combine data within a regime for analysis, the ASTM procedure also suggests that because of wind shifts and concentration fluctuations, the cross-wind concentration distribution along a sampling arc is unlikely to be perfectly Gaussian, a lateral distribution assumed by most air dispersion models. These departures from the ideal Gaussian shape lead to uncertainty in defining the plume centerline position. As a result, the ASTM procedure further recommends treating all ''near-centerline'' observed concentrations as if they were representative of the plume centerline concentration.

One way to define near-centerline concentrations is described below, as suggested by the ASTM (2000) procedure. Figure 4 illustrates

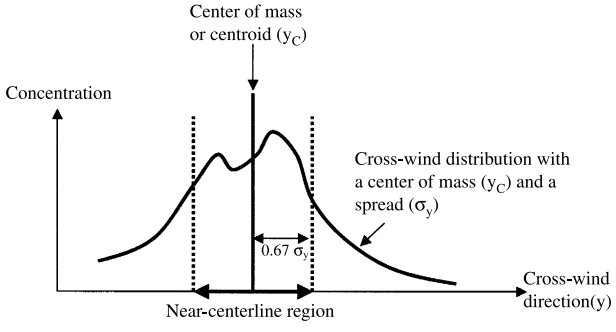



**Fig. 4.** Schematic of the definition of a near-centerline region for the ASTM procedure. A sample cross-wind concentration distribution is shown. The first moment of the distribution defines the center-of-mass (or centroid) location, $y_c$. The second moment defines the spread, $\sigma_y$, of the distribution. The region, marked by dotted lines, that is within 0.67 $\sigma_y$ from the centroid location is the near-centerline region

the method. First, it is important to note that the plume must have been well-captured by the sampling arc before any further analysis can be conducted. This usually involves plotting all observations along the arc and carefully inspecting these plots. Once the data are quality-assured, then consider all those measurements that are within a certain range of the plume center-of-mass (or centroid) location. Due to uncertainty in the plume centerline position, the ASTM procedure suggests that the concentration at a receptor that is within 0.67 $\sigma_y$ from the centroid location is a representative sample of the plume centerline concentration. For a Gaussian distribution, the value at a lateral distance of 0.67 $\sigma_y$ from the centerline would equal 80% of the centerline maximum value.

The BOOT software has been extended to include the ASTM method. There are many similarities between the ASTM (2000) procedure and the BOOT software described earlier, such as

- the calculation of statistical performance measures such as FB,
- the use of the bootstrap resampling technique to estimate the confidence level,
- the paired sampling between observational and predicted values, and
- the grouping of the data in blocks in BOOT and in regimes in ASTM.

However, the ASTM procedure proceeds further by

- calculating the performance measures based on regime-averages (i.e., averaging over all experiments within a regime), rather than the values for individual experiments, and
- if the variable to be evaluated is the centerline concentration, considering the near-centerline observations as representative samples of the centerline value, where the near-centerline region can be defined as within a distance (e.g., 0.67 times the lateral distance scale of the concentration distribution) from the cloud centroid location.

The ASTM (2000) procedure represents a promising approach. However, there are still some issues to be resolved and investigated before the procedure can be fully useful.

– There is a need to study the sensitivity of the evaluation results to the definition of regimes (i.e., how data are stratified).
– There is always only a limited number of regimes (e.g., −20 to 40) that can be defined, regardless of the size of the dataset. As a result, the performance measures are always determined by this limited number of regime averages. It is necessary to carefully examine the implication of accounting for only the variance in regime averages, rather than the full variance in the complete dataset.
– The ASTM procedure has so far only been demonstrated for short-range dispersion experiments with concentric sampling arcs, where multiple observed concentrations near the plume centerline are assumed to represent all possible values of the centerline values. However, many other mesoscale or long range field experiments do not have similar arc-wise configurations.

### *2.8 Cumulative Distribution Function (CDF) method*

Some researchers such as Lewellen et al (1985), Lewellen and Sykes (1989), and Weil et al (1992) suggest that, since observations are sampling a random process and almost all dispersion models predict some sort of ensemble averages, a deterministic evaluation framework is not appropriate. This argument is similar to the premise for the ASTM approach mentioned above. Lewellen et al (1985) devised a method to test the hypothesis that, at any location and time, the observed concentration, $C_o$, is simply a sample taken from the probability density function (PDF) of the predicted concentration, $C_p$. In this method, $C_p$ is treated as a random variable, whereas most air dispersion models give ensemble mean predictions, $\overline{C_p}$. If a model predicts large differences (or residuals) between $C_o$ and $\overline{C_p}$, then it is desirable to check whether the differences are within the expected 95% uncertainty range of $C_p$. One method to generate the required PDF and its 95% range is to use higher-order turbulence closure models such as SCIPUFF (Sykes et al, 1998), which provide predictions of the concentration mean and variance. The shape of the PDF is assumed by Lewellen and Sykes (1986) to follow the clipped normal distribution, which is obtained by replacing all negative values in a general Gaussian distribution with zeros.

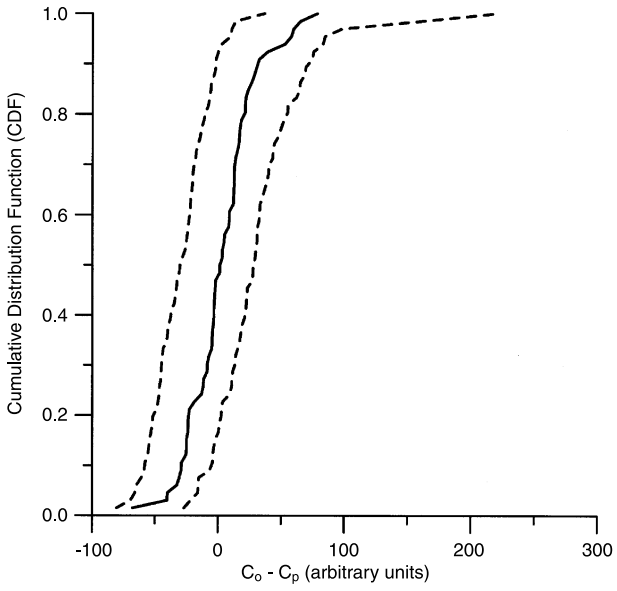



**Fig. 5.** Schematic of the use of the cumulative-distribution-function (CDF) approach for model evaluation. $C_o$ is observation, and $C_p$ is model prediction. In this approach, $C_o$ is treated as a sample taken from the probability density function (PDF) of $C_p$, which is assumed to be a random variable. The PDF can be estimated by higher-order turbulence closure models, or by a Monte Carlo analysis that involves many sensitivity model runs, in addition to the base model run. Solid curve represents the model residuals, $C_o - \overline{C_p}$, for the base run. Each Monte Carlo run creates a new version of the residual CDF. A sufficient number of Monte Carlo runs then allow the estimation of the 95% confidence bounds, as indicated by dashed curves

Besides the use of various analytical distribution functions to define the PDF, a second option is to generate the required PDF by means of a Monte Carlo analysis (e.g., Hanna and Davis, 2002). Figure 5 provides a schematic of the CDF approach based on a Monte Carlo analysis (Hanna and Davis, 2002). The solid curve represents the CDF of a series of model residuals, $C_o - \overline{C_p}$. In this example, the CDF is constructed after rank ordering the differences between the overall maximum observed and predicted hourly ozone concentrations at N sampling stations for a high-ozone episode. Each Monte Carlo sensitivity run creates a new version of the CDF. With a sufficient number (say, $\geq$100) of Monte Carlo runs, the 95% confidence bounds of the CDF can then be estimated, shown by dashed curves in Fig. 5. Since the solid curve is bounded by dashed curves in Fig. 5, it can be concluded that

model performance is consistent with the estimated uncertainty, or that the model is performing as well as expected in light of the stochastic or uncertain values of $C_p$.

The CDF approach mainly checks whether model residuals are consistent with our expectation of model uncertainty. It does not provide quantitative information on model performance, and should be applied in conjunction with other evaluation techniques, such as those mentioned above. There are also other issues that need to be investigated further. For example, the turbulence closure models only address concentration fluctuations due to turbulent wind fields, and cannot adequately address the important spatial and temporal correlation information (Hanna and Davis, 2002; Sykes, 2002). Monte Carlo analysis is more robust in accounting for this correlation information and uncertainties due to other model inputs such as emissions and the chemical reaction rate constants. On the other hand, Monte Carlo analysis cannot easily estimate turbulent contributions from smaller-scale processes.

### *2.9 Summary of Model Evaluation methods*

Various methodologies for evaluating atmospheric dispersion model performance have been presented and discussed. It is recommended that any model evaluation exercise should start with clear definitions of the evaluation goal and the variables to be evaluated, followed by exploratory data analysis, and then statistical performance evaluation. Exploratory data analysis involves the use of many types of plots, including scatter plots, quantile–quantile plots, box-residual plots, and scatter-residual plots, where the residual refers to the ratio of the predicted to observed values. The first two types of plots give an overall assessment of model performance. The last two types of plots are useful in identifying potential flaws in model physics, as indicated by trends of model residuals with independent variables.

The BOOT software package calculates a set of performance measures (or metrics), including the fractional bias (FB), the geometric mean (MG), the normalized mean square error (NMSE), the geometric variance (VG), the correlation coefficient (R), and the fraction of data where predictions are within a factor of two of observations (FAC2). The FB and MG measure systematic bias, whereas NMSE and VG measure systematic bias and random scatter. There is not a single performance measure that is universally applicable to all situations, and a balanced approach is required to look at a number of performance measures. For dispersion modeling where concentrations can easily vary by several orders of magnitude, MG and VG are preferred over FB and NMSE. However, MG and VG may be strongly influenced by very low values, and are undefined for zero values. It is recommended that the instrument threshold, such as the limit of quantitation (LOQ), be used as a lower threshold in calculating MG and VG. The R is probably not a very robust measure because it is sensitive to a few outlier data pairs. Furthermore, because measurements are commonly available in concentric arcs for short-range dispersion field experiments, there is already a pattern in the dataset, i.e., concentration decreasing with downwind distance. Since any reasonable dispersion model would be able to reproduce this pattern, R often mainly reflects this agreement, and is thus not that informative. The FAC2 is probably the most robust performance measure, because it is not overly influenced by either low or high outliers.

Since NMSE and VG measure both systematic bias and random (unsystematic) scatter, it is recommended that they be further partitioned to determine the fractional contributions of the systematic and unsystematic components. It is also recommended that FB, NMSE, MG, and VG be further interpreted by translating them into a quantity (e.g., the equivalent factor-of-N-difference between predictions and observations) that is more easily understood.

Bootstrap resampling can be used to estimate the confidence limits of the performance measures, in order to address questions such as (1) whether the FB for Model-A is significantly different from zero, and (2) whether the FB for Model-A and the FB for Model-B are significantly different.

In addition to the six performance measures included in the BOOT software, there are also other performance measures that can be defined, such as the normalized standard deviation (NSD) and the normalized root mean square error

(NRMSE). It can be shown that NSD, NRMSE, and R satisfy a geometric relationship based on the law of cosines. This relationship allows the plotting of the three performance measures in a two-dimensional nomogram.

The figure of merit in space (FMS) and the measure of effectiveness (MOE) can also be used to measure model performance. The FMS and MOE are naturally more appropriate for fields that can be easily contoured. As previously mentioned, short-range dispersion field experiments typically have the concentration data measured along concentric arcs. This type of data is often not suitable for contouring. In order to implement the MOE, one method suggested by Warner et al (2001) is to replace area estimates with straight data summation. In this case, it can be shown that the MOE and FB are in fact mathematically related. The definitions of the false negative and false positive parts of FB allow their sum to equal the normalized absolute error (NAE).

Many researchers have suggested the inadequacy of a deterministic evaluation framework, because observations are realizations of ensembles and model predictions often represent ensemble averages. The American Society for Testing and Materials (ASTM, 2000) approach suggests (1) grouping experiments under similar conditions (regimes), (2) averaging predictions and observations over each regime, and (3) calculating the performance measures based on these regime averages. The ASTM procedure also recommends the use of bootstrap resampling to estimate the confidence limits. Because there are many similarities between the BOOT and ASTM methodologies, the former can be easily extended to also include the latter.

Another way to evaluate model performance in a stochastic framework is to assume that the observed concentration is simply a random sample taken from the probability density function (PDF) of the predicted concentration. The PDF can be estimated by such techniques as higher-order turbulence closure schemes and Monte Carlo analysis. The cumulative distribution function (CDF) of model residuals (observations minus predictions) can then be plotted to see whether it lies inside the confidence bounds given by the PDF. If so, then large model residuals are said to be consistent with predictions after accounting for uncertainties. The CDF approach is qualitative, and should be used in conjunction with other more quantitative techniques.

### *2.10 Typical magnitudes of performance measures*

There have been many applications of various subsets of the above model evaluation methodologies to a variety of models and scenarios. This experience allows ''typical magnitudes'' of performance measures to be estimated, which could be used to develop model acceptance criteria.

Because a given model's performance usually varies from one field site to the next (typically there may be a 20% mean bias towards overprediction at one site and a 30% mean bias toward underprediction at another site), the most useful model evaluation studies are those that look at a number of models and a number of field sites.

Another important consideration is that model performance will vary with factors such as complexity of the scenario, uncertainty in the source term, and type and amount of meteorological data. Most of the reported model evaluation studies are associated with ''research grade'' field experiments with good instruments and uncomplicated terrain and simple source scenarios. And the experimentalists usually pack up and go home if it starts raining. Only a few of the reported model evaluation studies make use of routine monitoring networks and long time periods (e.g., one year).

One of the more comprehensive recent model evaluation documents is the EPA's evaluation of their new model, AERMOD (American Meteorological Society/Environmental Protection Agency Regulatory Model Improvement Committee Dispersion Model), with numerous field data sets, representing both research grade tracer experiments and routine monitoring networks around industrial stacks (Paine et al, 1998). The evaluations also include the existing model, ISC (Industrial Source Complex Dispersion Model), which AERMOD is intended to replace. Because of the emphasis of the EPA regulations on the highest concentrations, the Paine et al (1998) evaluations are focused more on the high-concentration region of quantile–quantile plots. Concentration averaging times are 1 hr

and 24 hrs. It is seen that AERMOD has better agreement than ISC at higher concentrations, and also improved performance over the middle (near the 50th percentile) region of the concentration distribution. In general, the new model is able to predict the maximum concentrations within about ±20 to 50%, although there is variation from site to site.

Hanna et al (2000) also evaluate AERMOD and ISC, as well as ADMS (Atmospheric Dispersion Modeling System), using data from five field studies (four tracer experiments and one routine monitoring network). This study brought up the issue about how to use multiple model performance measures from multiple sites to arrive at a single conclusion. It was decided to list the performance measures and then rank the models for each experiment and performance measure, and come up with a final ''score''. The types of results that would go into this ranking are listed as an example in Table 1. The data represent the maximum concentration on a monitoring arc and are therefore not paired in space. It is seen that ADMS and AERMOD both overpredict by about 20 or 30% at the high end and at the median, on average, and their scatter is slightly more than a factor of two. These two models' predictions are within a factor of two of the observations about 50% of the time. The ISC model, on the other hand, does not have as good performance across all performance measures except MG (related to mean bias). However, the fairly good mean bias for ISC is actually caused by compensating errors.

Similar conclusions are found in the review paper by Weil et al (1992) on air quality model evaluation. They stress that, because of variations in wind direction, it is almost fruitless to attempt to compare predictions and observations paired in space and time. They point out that the predicted and observed plume shapes and contours may be in very good agreement, but a displacement of 20 degrees may cause the two sets of contours to not overlap at all. For this reason, they stress the use of quantile–quantile plots applied to data unpaired in time or space (perhaps paired by downwind distance arc). For regulatory applications where the highest and second highest concentrations, regardless of their locations, are typically of interest, the unpaired in space or time comparisons are usually sufficient. However, it is also recognized that there are applications such as emergency response, homeland security, and environmental justice, where it is necessary to predict the exposure to a population, in which case not only the magnitude, but also the location and shape of the concentration field are important. This is the reason why despite the challenge due to uncertainty in wind direction, there is nevertheless an interest in the paired in space or time comparisons.

**Table 1.** Median performance measures over five field experiments for three models (from Hanna et al, 2000)

| | ISC3 | ADMS | AERMOD |
|---|---|---|---|
| $\text{Max}\, C_p/\text{Max}\, C_o$ | 6.7 | 0.80 | 0.77 |
| MG | 0.70 | 1.22 | 1.7 |
| VG | 7.7 | 2.4 | 2.9 |
| FAC2 | 0.33 | 0.53 | 0.46 |

Olesen (2001) reviews the numerous applications of the Model Validation Kit, which includes the BOOT software, the ASTM (2000) software, and many field experiment data sets. These studies show that, for the better models, the typical relative mean bias is about ±40% and the amount of random scatter is about a factor of two of the mean. Most of these field data sets involved short-range (up to the order of 10 km or so) dispersion experiments, maximum concentrations along a sampling arc but paired in time, and flat to rolling terrain.

Chang et al (2003) investigated the performance of three models over two mesoscale tracer experiments, where it was shown that wind direction variations over the 30 km domain caused errors in model predictions at specific monitor locations. The general conclusion was that the two better models had mean biases within ±35%, relative random scatter of about a factor of three or four, and 40 to 60% of the predictions (of the maximum concentration on an arc) within a factor of two of the observation.

Hanna (1988; 1993) presented overviews of many model evaluation exercises (about 20 models and about 20 field sites) over short-ranges (less than a few km) to conclude that the typical relative mean bias was about ±20 to 50% and the typical relative scatter was about 60 to 80%. It was suggested that the minimum values of

about ±20% for mean bias and about ±60% for relative scatter represent the best achievable performance of air quality models. Random turbulence (natural or inherent uncertainty) prevents any further improvement in models.

Klug et al (1992) evaluated the performance of many long-range transport models applied to the ETEX tracer release experiment in Europe, where the tracer cloud was tracked for over 1000 km. It was found that the unpaired relative biases and scatter were similar to that found for the short-range experiments and reported above, but that the wind shears, fronts, and rain patterns associated with ''real weather'' at those scales could cause large displacements in position of the predicted and observed clouds.

To conclude, ''good'' performing models appear to have the following typical characteristics based on unpaired in space comparisons:

- The fraction of model predictions within a factor of two of observations is about 50%.
- The mean bias is within ±30% of the mean.
- The random scatter is about a factor of two to three of the mean.

Of course these numbers will be revised as more evidence appears from new model evaluation exercises.

## 3. Test case – Salt Lake City Urban 2000 IOP09

The application of the model performance evaluation methods described in Sect. 2 can be best illustrated using a test case. Two versions of a simple urban baseline dispersion model (Britter and Hanna, 2003; Hanna et al, 2003) are applied to a single Intensive Operating Period (IOP) during the Salt Lake City Urban 2000 field experiment.

### *3.1 Description of field experiment*

The Urban 2000 flow and dispersion experiment in the Salt Lake City area took place in September and October, 2000 (Allwine et al, 2002). There were six nights of experiments, where three $SF_6$ releases were made every two hours and the release duration was one hour from a source near street level in the downtown area.

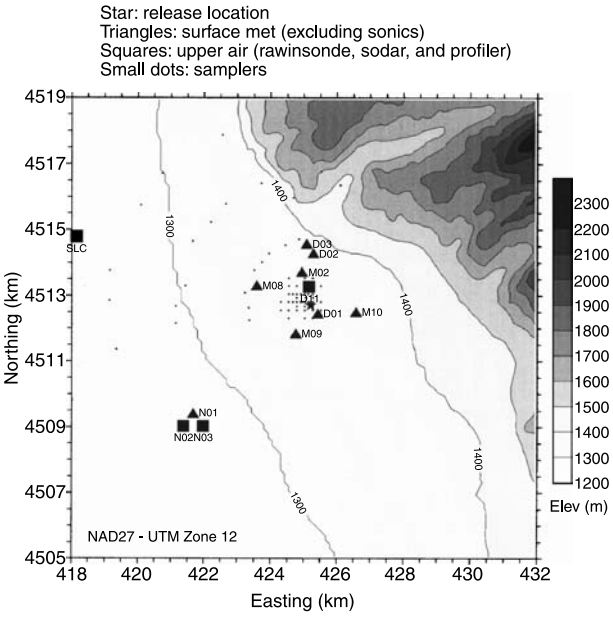



**Fig. 6.** Map of Salt Lake City domain used in the Urban 2000 study, showing the locations of the release, the surface meteorological monitors, the upper air sites, and the $SF_6$ samplers

The current test case involves only one of the six Intensive Operating Periods (IOPs) – IOP09, which took place the night of 20–21 October 2000, with 2 g/s releases at 2100, 2300, and 0100 LST. IOP09 was the experiment with the highest average wind speed – 2.64 m/s, when averaged over six downtown anemometers. All the anemometers were mounted on top of buildings, whose heights are between 4 and 20 m except for one building that is 121 m tall. Figure 6 shows the test domain, which covers a 14 km square area. The release point is marked by a star near the middle of the domain and the $SF_6$ monitors are marked as black dots. Three $SF_6$ sampling arcs are visible at distances of about 2, 4, and 6 km to the northwest of the release point. In addition, in the 1.3 km square area known as the ''Downtown Domain'', there were grids of monitors located on block intersections and midway along the blocks. These monitors were used to define four additional ''arcs'' at distances from about 0.15 km to 1 km. The majority of $SF_6$ concentrations were reported as 30 minute averages over a six-hour period during each night, allowing data from all three releases to be reasonably captured. Some of the meteorological monitors are also shown in Fig. 6. The

Salt Lake City (SLC) National Weather Service (NWS) anemometer is at the airport in the northwest corner of the figure. The N01 surface anemometer, the N02 sodar, and the N03 profiler sites are located in a rural area about 6 km upwind of the urban area. The M02 anemometer is at the top of a 121 m building, and the D11 square marks a sodar at the top of a 36 m building. Average building height, $H_b$, is about 15 m, and the surface roughness length, $z_o$, is estimated to be about 2 m.

The distributions of the observed 30 minute-averaged concentrations on each of the seven monitoring arcs were plotted and the maximum concentration, $C_{max}$, and the lateral standard deviation of the concentration distribution, $\sigma_y$, were calculated if there were sufficient data. In some cases, there were problems because the concentrations were all below the threshold of about 45 ppt, or there was perhaps only a single high observation, or the plume was obviously on the edge of the network. The concentration data were also analyzed for continuity in space and time.

The focus of this test case is on two types of comparisons:

(1) the normalized one-hour averaged arc-maximum concentration, $C_{max}/Q$, anywhere on an arc during the passage of the cloud from each of the three trials;
(2) the normalized one-hour averaged maximum concentration, $C/Q$, at each monitor location during each of the three trials.

Comparisons with data set 1 are called ''unpaired in time and space'', while comparisons with data set 2 are called ''unpaired in time but paired in space''. Actually there is some pairing in time, since the observed one-hour average is taken from a two to three hour sampling period when the $SF_6$ cloud was passing through the network. The observed maximum could have been from 2130 to 2230 LST, while the predicted maximum is assumed to be steady-state in this exercise. Also, in Comparison 1, there is pairing in along-wind distance (i.e., by monitoring arc), even though there is not necessarily pairing in cross-wind distance.

Table 2 contains the observed and predicted $C_{max}/Q$ values, times $10^{-6}\,s/m^3$, for each monitoring arc in the three trials in IOP09. This table would be used for the evaluations made as part of Comparison 1 above. The right side of the table contains the observed and predicted lateral dispersion parameter $\sigma_y$. Concentration data were also available from 66 monitor locations, for use in the evaluations made as part of Comparison 2 above, but these numbers are not listed in this paper.

### *3.2 Model evaluation results*

Table 2 contains the predictions of $C_{max}/Q$ and $\sigma_y$ for two simple urban baseline dispersion models, both of which have a Gaussian structure. Hanna et al (2003) describe the model equations. Both models assume neutral stabilities in urban areas. It is not necessary to know the precise structure of the two models since they are used only for demonstration of the model performance measures. Because the wind speed cancels out of Model A, its predictions of $C_{max}/Q$ and $\sigma_y$ are the same for each trial. Model B does account for variations in wind speed and therefore yields different predictions of $C_{max}/Q$ for each trial.

The results of the evaluation of Models A and B with the IOP09 $SF_6$ tracer data are given in Fig. 7 and Table 3.

Figure 7 contains scatter plots of predicted (for Models A and B) and observed $C/Q$ for paired-in-space $C/Q$ (top) and for arc maximum $C_{max}/Q$ (bottom) comparisons. The scatter plots for paired $C/Q$ on the top consist of 34 points where both observed ($C_o/Q$) and predicted ($C_p/Q$) values exceed $0.12 \times 10^{-6}\,s/m^3$ (or $C > 45$ ppt). The scatter plots of $C_{max}/Q$ on the bottom consist of 18 points (seven monitoring arcs times three trials equals 21, minus three trial-arcs with insufficient data). It can be seen that the plots for the 18 arc-maximum values, $C_{max}/Q$, show less scatter than the plots for the many monitors, because the arc-maximum values do not impose pairing in space and they include only the highest concentrations in the center of the plume. Model A appears to exhibit a slight overprediction and Model B exhibits a slight underprediction. The scatter is seen to be generally within a factor of five to ten for the paired data (top) and within a factor of two for the arc maxima (bottom). However, it must be remembered that the seven monitoring arcs cover a distance range from about 160 m to 6000 m, and so

**Table 2.** Test case observations and predictions of arc-maximum normalized concentration, $C_{max}/Q$, for IOP09 from Salt Lake City Urban 2000. The numbers 1, 2, and 3 refer to the separate release trials in IOP09. Models A and B are from the Hanna et al (2003) study. Model A assumes that the predictions of C/Q are mostly independent of u and employ the all IOP average wind speed of 1.39 m/s. Model B assumes the actual wind speed for each release trial. For IOP09, the wind speeds are 2.69 m/s, 2.47 m/s, and 3.23 m/s for trials 1, 2, and 3, respectively. Observed and predicted $\sigma_y$ are listed

| Arc | | $C_{max}/Q$ ($10^{-6}$ s/m$^3$) | | | | | | | | $\sigma_y$ (m) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | x | Trial 1 | Trial 2 | Trial 3 | All trial avg | Model A All trial avg | Model B Trial 1 | Model B Trial 2 | Model B Trial 3 | Trial 1 | Trial 2 | Trial 3 | All trial avg | Model A All trial avg | Model B All trial avg |
| | (m) | Obs | Obs | Obs | Obs | Pred | Pred | Pred | Pred | Obs | Obs | Obs | Obs | Pred | Pred |
| 1 | 156 | 129 | 244 | 115 | 163 | 229 | 141.7 | 154.4 | 118 | 76 | 56 | 49 | 60 | 34.7 | 29 |
| 2 | 394 | 49.7 | 56.8 | 37.3 | 47.9 | 52.3 | 30 | 32.6 | 25 | 137 | 123 | 120 | 127 | 73.4 | 66.2 |
| 3 | 675 | 44.9 | 32.5 | 11 | 29.5 | 21.2 | 12.2 | 13.3 | 10.2 | 153 | 170 | 150 | 158 | 115 | 103 |
| 4 | 928 | 4.77 | n/a | 7.56 | 6.17 | 12.5 | 7.17 | 7.81 | 5.98 | 224 | 189 | 228 | 214 | 150 | 134 |
| 5 | 1974 | n/a | n/a | 2.63 | 2.63 | 3.71 | 2.14 | 2.33 | 1.78 | n/a | 256 | n/a | 256 | 273 | 244 |
| 6 | 3907 | 0.63 | 1.06 | 1.5 | 1.06 | 1.36 | 0.784 | 0.854 | 0.653 | 503 | 494 | n/a | 499 | 447 | 398 |
| 7 | 5998 | 0.49 | 1.01 | n/a | 0.75 | 0.76 | 0.439 | 0.478 | 0.365 | 820 | 1075 | n/a | 948 | 593 | 529 |
| u (m/s) | | 2.69 | 2.47 | 3.23 | 2.64 | 1.39 | | | | | | | | | |

**Table 3.** Results of model performance evaluation for Models A and B for Comparison 1 (Arc Max, last row) and for four alternate methods of addressing the data in Comparison 2

| | | FB | $FB_{fn}$ | $FB_{fp}$ | NMSE | MG | VG | FAC2 | R | High | 2nd high | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| All $C_o$ and $C_p$, paired in space | Model A | −0.043 | 0.210 | 0.253 | 3.47 | 3.23 | 233 | 0.57 | 0.93 | 229 | 229 | 147 |
| | Model B | 0.518 | 0.544 | 0.026 | 3.47 | 4.28 | 561 | 0.55 | 0.95 | 154 | 142 | |
| $C_o > 45$ ppt and $C_p > 45$ ppt, paired in space | Model A | −0.134 | 0.131 | 0.265 | 0.86 | 1.19 | 4.1 | 0.56 | 0.92 | 229 | 229 | 34 |
| | Model B | 0.431 | 0.459 | 0.028 | 0.85 | 2.70 | 21.4 | 0.47 | 0.94 | 154 | 142 | |
| $C_o > 45$ ppt and all $C_p$, paired in space | Model A | −0.041 | 0.210 | 0.253 | 1.68 | 8.92 | 48974 | 0.28 | 0.93 | 229 | 229 | 71 |
| | Model B | 0.518 | 0.544 | 0.026 | 1.68 | 15.52 | 306520 | 0.24 | 0.95 | 154 | 142 | |
| All $C_o$ and $C_p > 45$ ppt, paired in space | Model A | −0.135 | 0.130 | 0.265 | 0.91 | 1.08 | 4.3 | 0.53 | 0.92 | 229 | 229 | 36 |
| | Model B | 0.431 | 0.459 | 0.028 | 0.90 | 2.47 | 18.4 | 0.47 | 0.94 | 154 | 142 | |
| Arc Max | Model A | −0.230 | 0.066 | 0.296 | 0.60 | 0.84 | 1.2 | 0.89 | 0.91 | 229 | 229 | 21 |
| | Model B | 0.294 | 0.318 | 0.024 | 0.46 | 1.48 | 1.4 | 0.78 | 0.95 | 154 | 142 | |

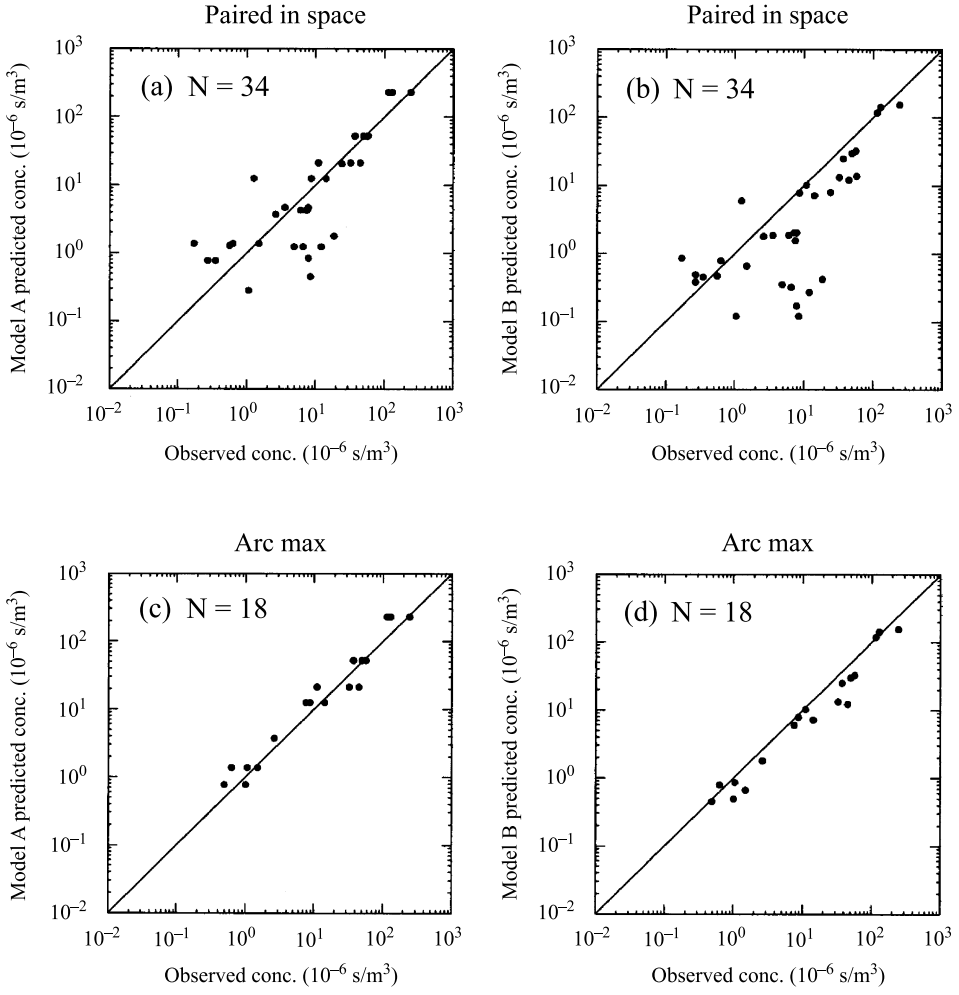



**Fig. 7.** Scatter plots of observed ($C_o/Q$) and predicted ($C_p/Q$) concentrations ($10^{-6}\,s/m^3$) for IOP 9 of Urban 2000; **a** Model A paired in space, and **b** Model B paired in space; when $C_o/Q$ and $C_p/Q$ are both greater than $0.12 \times 10^{-6}\,s/m^3$ (or 45 ppt when not normalized by the emission rate); **c** Model A arc maximum, and **d** Model B arc maximum. The number of points (N) is also indicated in each frame

the predictions and observations both show a three-order-of-magnitude decrease from close distances (with high concentrations) to far distances (with low concentrations).

The plots with all 147 points, including many cases with $C_o$ and/or $C_p$ less than 45 ppt, are not shown here. They exhibit much more scatter, especially at low concentrations, since there are many monitors on the edges of the observed or predicted plumes. The model performance measures are influenced by assumptions regarding the minimum or threshold concentration. For this experiment, concentrations below 45 ppt ($0.12 \times 10^{-6}\,s/m^3$ when normalized by the emission rate) were considered to be rather uncertain. Also, because the global background concentration of $SF_6$ is 3 ppt ($\sim 0.01 \times 10^{-6}\,s/m^3$ when not normalized by the emission rate), and that value was added to all concentrations predicted by the models, there are 63 points at $0.01 \times 10^{-6}\,s/m^3$. These 63 points will be seen to affect some of the statistical performance measures. It can be shown that the model performance measures are strongly affected by whether we add the global background of 3 ppt to the model predictions or subtract 3 ppt from the observations.

Many of the statistical model performance measures discussed in Sect. 2 are calculated for the data in the scatter plots. Table 3 contains the results for Models A and B for Comparison 1

(Arc Max) and for four alternate methods of addressing the data in Comparison 2:

- Method 1 (top row of the table) includes "all $C_o$ and $C_p$, paired in space", no matter how low the concentrations. There are 147 points.
- Method 2 (second row of the table) includes only monitors where both $C_o$ and $C_p$ exceed the threshold of 45 ppt (or $C/Q = 0.12 \times 10^{-6}\,s/m^3$). There are 34 points.
- Method 3 (third row of the table) includes only monitors where $C_o > 45$ ppt, but $C_p$ can have any value. There are 71 points.
- Method 4 (fourth row of the table) includes only monitors where $C_p > 45$ ppt but $C_o$ can have any value. There are 36 points.

These four methods are applied in order to show the range of the results depending on the assumptions concerning low concentrations. Note that there are 71 points with $C_o > 45$ ppt but only 36 points with $C_p > 45$ ppt, reflecting the fact that the predicted plume width is too small. Probably the one-sided comparisons are less valid in rows three and four of the table (where one of $C_o$ or $C_p$ is not allowed to drop below 45 ppt, while the other one is allowed to have any value).

Despite the range in some of the performance measures in Table 3, due to the various methods for considering low concentrations, the overall results in the table appear to be fairly conclusive: Overall, the values of FB suggest that Model A overpredicts by a slight amount (about 10 to 20%), while Model B underpredicts by about 20 to 40%. The fraction of predictions within a factor of two, or FAC2, is slightly less for Model B than for Model A, and is about 0.8 for the arc-maximum values and about 0.5 for all monitors. The NMSE less than unity indicates that the magnitude of the scatter is less than the mean concentration. The scatter, as measured by NMSE or VG, appears to be more strongly influenced by the assumption of which monitors to use and which low concentration threshold to select. This result has been found in other studies, and is caused by the fact that outliers have a stronger influence when they are squared in the calculation of the performance measure. In the case of VG, where the logarithm of C is taken, the assumption regarding the threshold can dominate the calculated VG, causing variations in calculated VG of several orders of magnitude. The correlation, R, is always between 0.90 and 0.95, due to the dominant influence of the decrease in concentrations with distance from the source, and is therefore not very useful as a method to discern model performance. As mentioned previously, the sum of the false negative and false positive components of FB (i.e., $FB_{fn}$ and $FB_{fp}$) yields the normalized absolute error (NAE). The values of NAE (not shown in Table 3) indicate that both Models A and B have a comparable absolute error that is about 30 to 40% of the mean, which means that Model B's NAEs are more due to persistent underpredictions, whereas Model A's NAEs are due to more overpredictions and less underpredictions.

Table 3 also contains the predicted "high" and "2nd high" C/Q values, to show whether the models can match the worst case observed concentrations. The observed high and 2nd high C/Q are 244 and $129 \times 10^{-6}\,s/m^3$ (see Table 2), while the Model A predictions are 229 and 229, and the Model B predictions are 154 and 142 (see Table 3). These predictions are relatively close to the observations, when compared to previous model evaluations. The relative difference in Models A and B is consistent with that found for the mean relative bias calculations.

The relative magnitudes of the false negatives (underpredictions) and false positives (overpredictions) can be seen in Table 3 by looking at the $FB_{fn}$ and $FB_{fp}$. As shown in Sect. 2, $FB = FB_{fn} - FB_{fp}$. On the other hand, the sum $FB_{fn} + FB_{fp}$ is the normalized absolute error, which is a more robust measure of scatter than NMSE or VG.

Table 4 addresses the questions concerning whether the performance measures are significantly different from zero. This table focuses on the fractional bias, FB, although any performance measure could be used, and although the numerical value in question may be different for different performance measures. The models (A and B), data set and other assumptions are the same in Table 4 as in Table 3. Note that there are five rows given in the table, related to the four ways of defining minimum C for the paired-in-space data set, and the arc-maximum data in the last row, where the minimum C is never reached so it is not of concern. Because there is not a "×" in any box for Model A, it follows that the FB for Model A is not significantly different

**Table 4.** Results of significance testing to determine (1) whether the FB for Models A and B ($FB_{Model\ A}$ and $FB_{Model\ B}$) are significantly different from zero, and (2) whether the difference in FB between Models A and B ($\Delta FB$) is significantly different from zero, for Comparison 1 (Arc Max, last row) and for four alternate methods of addressing the data in Comparison 2. A "×" indicates that the results are significant at 95% confidence intervals

| | $FB_{Model\ A}$ | $FB_{Model\ B}$ | $\Delta FB$ |
|---|---|---|---|
| All $C_o$ and $C_p$, paired in space | | × | × |
| $C_o > 45$ ppt and $C_p > 45$ ppt, paired in space | | × | × |
| $C_o > 45$ ppt and all $C_p$, paired in space | | × | × |
| All $C_o$ and $C_p > 45$ ppt, paired in space | | × | × |
| Arc Max | | × | × |

from zero (at the 95% confidence level). However, because there is a "×" in all the boxes for Model B, the FB for that model is significantly different from zero. And of interest is the model-to-model comparison in the last column, where it is concluded that the FB for Model A is significantly different from the FB for Model B. This information is routinely output from BOOT.

Three types of plots commonly used in statistical model evaluation exercises are given in Figs. 8–10.

Figure 8 contains quantile–quantile plots for Model A (left) and Model B (right), where only monitors with both $C_o$ and $C_p$ exceeding 45 ppt (or C/Q of $0.12 \times 10^{-6}$ s/m$^3$) are considered. To create these plots, all observed data are rank-ordered, all predicted data are rank-ordered, and each point on the plot represents a particular rank number (e.g., the maximum $C_o$ and the maximum $C_p$). As previously mentioned, the goal of the quantile–quantile plot is to see whether the distributions of all the observed and predicted values as a whole are comparable, where it is not necessary to require that, for example, the maximum $C_o$ and the maximum $C_p$ take place under the same conditions. The general tendency for underprediction of Model B at all parts of the distribution can be seen, although Model B does well with the three highest concentrations. The distribution suggests little bias for Model A over most of the range.

Figure 9 contains residual plots of $C_p/C_o$ as a function of arc distance from the source. This is for the same data set as used in Fig. 8. Both figures suggest little overall trend with distance, although some arc distances (e.g., 400 m) show an underprediction and other arc distances (e.g., 6 km) show an overprediction. The underprediction at

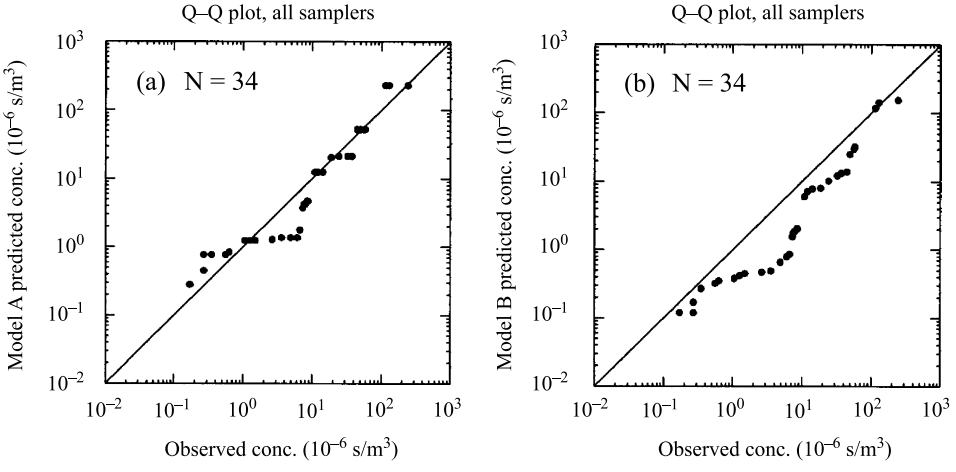



**Fig. 8.** Quantile–quantile (Q–Q) plots of observed ($C_o/Q$) and predicted ($C_p/Q$) concentrations ($10^{-6}$ s/m$^3$) when $C_o/Q$ and $C_p/Q$ are separately ranked; (**a**) Model A for all sampler locations, and (**b**) Model B for all sampler locations; when $C_o/Q$ and $C_p/Q$ are both greater than $0.12 \times 10^{-6}$ s/m$^3$ (or 45 ppt when not normalized by the emission rate). The number of points (N) is also indicated in each frame

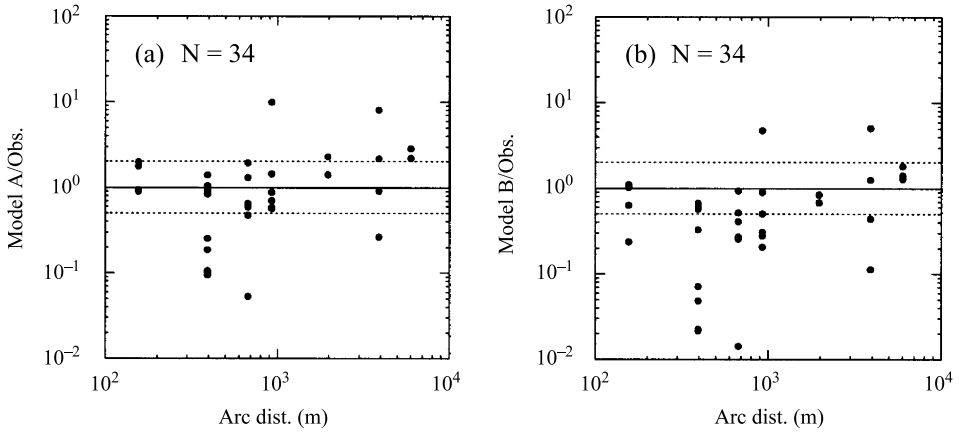



**Fig. 9.** Residual (ratio of model prediction to observation) plots for IOP 9 of Urban 2000 as a function of arc distance (m); (**a**) Model A for all sampler locations, and (**b**) Model B for all sampler locations; when $C_o/Q$ and $C_p/Q$ are both greater than $0.12 \times 10^{-6}\,s/m^3$ (or 45 ppt when not normalized by the emission rate). Dashed lines indicate factor-of-two scatter

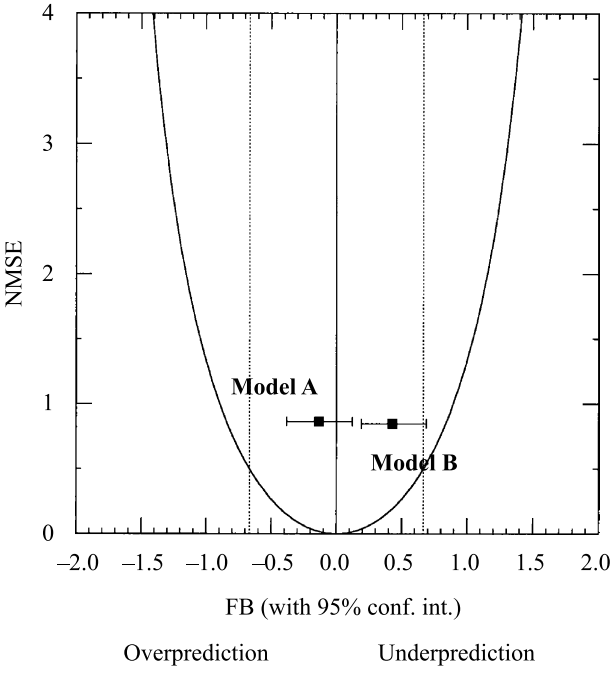



**Fig. 10.** FB, together with the 95% confidence intervals, and NMSE for Models A and B, when $C_o/Q$ and $C_p/Q$ are both greater than $0.12 \times 10^{-6}\,s/m^3$ (or 45 ppt when not normalized by the emission rate). The parabola indicates the minimum NMSE for a given FB

the 400-m arc might be due to the fact that the arc is in the downtown area with a number of tall buildings. Nevertheless, these plots also clearly show the points where the predictions are within a factor of two of the observations (i.e., the points within the range defined by the dashed lines).

Figure 10 is an "FB versus NMSE" plot which is often used as a single plot to indicate overall relative model performance. The perfect model would have FB = NMSE = 0.0. Models A and B are the same ones used before, but this time the emphasis is only on data monitors where both $C_o$ and $C_p$ exceed 45 ppt (the same as in the previous two plots). The figure suggests that Model A and Model B have similar scatter (NMSE) but Model A has a lower relative mean bias (FB).

## 4. Conclusions

This paper investigates methodologies for evaluating the performance of dispersion (air quality) models. Dispersion models are used to predict the fate of gases and aerosols after they are released into the atmosphere. For example, these pollutants can be from regularly emitted industrial sources, accidental releases, and intentional releases of harmful chemical and biological agents. The dispersion model results have been used in applications such as granting permits and preparing risk management programs for industrial facilities, designing emissions control strategies for regions that are not in compliance with existing air quality standards, and conducting forensic studies of historical events (e.g., dosage reconstruction for areas surrounding a nuclear fuel processing site, and possible exposures due to the derailment of rail cars containing toxic chemicals). Because of the large economic and

public-health impacts, these atmospheric dispersion models should be properly evaluated before they can be used with confidence.

As is the case for all other kinds of computer models, there are also uncertainties in the dispersion model results. For example, random turbulence in the atmosphere causes the concentration field to fluctuate. As a result, the concentration field can only be described through gross statistical properties (e.g. mean and variance). In addition, the input data used to run dispersion models are subject to errors due to such factors as instrument accuracy and representativeness. Physical parameterizations included in the models can also contain errors, approximations, and uncertainties. Therefore, it is critical to understand, and quantify if possible, the uncertainty in the dispersion model results. The current paper mainly focuses on the issue of model evaluation, and does not go into detailed analyses of model uncertainty.

Section 2 provided a review of various methodologies for evaluating the performance of atmospheric dispersion models. Any model evaluation exercise should start with clear definitions of the evaluation goal (i.e., the statistical null hypothesis to test) and the types of model outputs to be evaluated. Before conducting any statistical performance evaluation, it is useful to first conduct exploratory data analysis by means of different graphical displays, including scatter plots, quantile–quantile plots, box-residual plots, and scatter-residual plots. (Here the residual refers to the ratio of the predicted to observed values.) Scatter and quantile–quantile plots provide a general assessment of model performance. Residual plots are useful in identifying potential problems in the model physics.

One commonly-used evaluation methodology is the BOOT software package previously co-developed by the author (Hanna et al, 1993). It calculates a set of quantitative performance measures, including the fractional bias (FB), the geometric mean (MG), the normalized mean square error (NMSE), the geometric variance (VG), the correlation coefficient (R), and the fraction of data where predictions are within a factor of two of observations (FAC2). Most previous studies where the BOOT software had been used simply quoted the values of these measures, which are not that informative to readers in general. It was suggested that measures such as FB, MG, NMSE, and VG be further expressed in terms of a quantity, such as the equivalent ratio of the predicted to observed values, that is easier to interpret.

Depending on the situation, some performance measures might be more appropriate than others. Hence, there is not a single best measure, and it is necessary to use a combination of the performance measures. The FAC2 is more robust, because it is not sensitive to the distribution of the variables to be evaluated. For dispersion modeling where concentrations or dosages often vary by several orders of magnitude, MG and VG are better than FB and NMSE. However, a lower threshold was suggested when calculating MG and VG, because these two measures are strongly influenced by very low values and are undefined for zero values. The R is not a very robust measure, because it is sensitive to a few aberrant data pairs, and often mainly reflects the obvious pattern (i.e., concentration decreasing with downwind distance) that already exists in the dataset.

The confidence limits of the performance measures can be estimated with bootstrap resampling (Efron, 1987) to answer questions such as (1) whether the FB for one model is significantly different from zero, and (2) whether the FBs for two models are significantly different at the 95% confidence level.

In addition to the six performance measures described above, other measures can also be defined, including the normalized standard deviation (NSD), and the normalized root mean square error (NRMSE). It has been shown that NSD, NRMSE, and R can be concisely plotted on a nomogram, based on the law of cosines (Taylor, 2001). Note that all these three measures only account for random scatter, but not systematic bias.

The figure of merit in space (FMS) and the measure of effectiveness (MOE) are similar measures that are essentially, in their original definitions, based on the ratio of the intersection to the union of the predicted and observed contour areas. The two-dimensional (2-D) version of the MOE separates the false-negative (or underpredicting) and false-positive (or overpredicting) components of predictions. For situations where contours are not easily defined (because, for example, receptors are arranged in concentric

arcs or the number of receptors is simply too few), Warner et al (2001) recommend using straight data summation as one of the options to provide the area estimates necessary for calculating the MOE. It can be shown that the 2-D MOE is closely related to a 2-D FB that is adapted from the traditional one-dimensional FB (Chang, 2002). One advantage of the 2-D FB is that, like the 2-D MOE, it separates the underpredicting and overpredicting components, so that cases of compensating errors will be easily detected.

Nevertheless, it is emphasized that the MOE has been traditionally calculated for each sampling arc of each trial, where the data at all sampler locations are paired in space. The average of these individual MOEs then provides a summary of model performance for all the experiments as a whole. On the other hand, the FB has been traditionally based on the arc-wise maximum concentrations (or dosages) for all experiments, so it directly gives an assessment of the overall model performance. Therefore, even though the MOE and FB are mathematically related, it is important to discern how they are actually implemented in order to interpret the overall model performance for all the experiments as a whole.

Because observations are realizations of infinite ensembles, and model predictions mostly represent some sort of ensemble averages, it has been recommended that observations and predictions under similar regimes (e.g., as defined by similar atmospheric stability and downwind distance) should be first averaged before conducting any performance evaluation (ASTM, 2000).

The ASTM procedure demonstrated the inadequacy of performing model evaluation in a deterministic manner. Lewellen et al (1985) recommend an evaluation methodology based on the assumption that the observed concentration is simply a random sample taken from the probability density function (PDF) of the predicted concentration. As a result, the cumulative distribution function (CDF) of model residuals (e.g., observations minus predictions) should be inspected to see if it is within the confidence bounds given by the PDFs of predicted concentrations. The PDF can be estimated by such techniques as higher-order turbulence closure schemes (e.g., Sykes, 1984) or Monte Carlo analysis (e.g., Hanna and Davis, 2002).

The performance measures reported in many model evaluation exercises with many sets of field data were reviewed. It was concluded that a ''good'' model would be expected to have about 50% of the predictions within a factor of two of the observations, a relative mean bias within $\pm 30\%$, and a relative scatter of about a factor of two or three.

To demonstrate the use of many of the performance measures, a model evaluation exercise was carried out using one six-hour period (IOP09) of the Salt Lake City Urban 2000 field experiment (Allwine et al, 2002). Two alternate baseline urban dispersion models (Britter and Hanna, 2003; Hanna et al, 2003) were run for this test case, and standard plots were presented and discussed such as scatter plots, quantile–quantile plots, residual plots, and ''FB versus NMSE'' plots. Tables are presented containing the performance measures, as well as significance tests on whether the fractional bias, FB, is significantly different from zero or the difference in FB between the two models is significantly different from zero. These performance measures were calculated using several alternate interpretations of the data (e.g., maximum C/Q on each downwind distance arc, and C/Q ''paired in time and space'' at each monitor but using various assumptions concerning the minimum C that is allowed. It is found that the assumption regarding minimum C can have a large effect on some of the performance measures.

**Acknowledgments**

This research was supported by the Defense Threat Reduction Agency (DTRA) and the Department of Energy (DOE). Major Brian Beitler and Mr. John Pace have been DTRA's technical contract representatives, and Dr. Michael Brown of Los Alamos National Laboratory has managed the DOE portion of the research. We thank Dr. K. Jerry Allwine of Battelle Pacific Northwest Laboratory for providing the Salt Lake City Urban 2000 field data.

**References**

Allwine KJ, Shinn JH, Streit GE, Clawson KL, Brown M (2002) Overview of Urban 2000. Bull Am Meteor Soc 83(4): 521–536

Anthes RA, Kuo Y-H, Hsie E-Y, Low-Nam S, Bettge TW (1989) Estimation of skill and uncertainty in regional

numerical models. Quart J Royal Meteor Soc 115: 763–806

ASTM (2000) Standard guide for statistical evaluation of atmospheric dispersion model performance. American Society for Testing and Materials, Designation D 6589-00. ASTM, 100 Barr Harbor Drive, West Conshohocken, PA 19428-2959

Beck MB, Ravetz JR, Mulkey LA, Barnwell TO (1997) On the problem of model validation for predictive exposure assessments. Stoch Hydrol Hydraulics 11: 229–254

Britter RE, Hanna SR (2003) Flow and dispersion in urban areas. Ann Rev Fluid Mech 35: 469–496

Chang JC (2002) Methodologies for evaluating performance and assessing uncertainty of atmospheric dispersion models. Ph.D. thesis, George Mason University, Fairfax, 277 pp. Available from: *www.lib.umi.com/dissertations/search*. The abstract can be seen by selecting ''Pub Number (PN)'' from the first pull-down menu and entering 3068631. The full document can be ordered on-line at $34 per unbound copy

Chang JC, Franzese P, Chayantrakom K, Hanna SR (2003) Evaluations of CALPUFF, HPAC, and VLSTRACK with two mesoscale field data sets. J Appl Meteor 42: 453–466

Cullen AC, Frey HC (1998) Probabilistic techniques in exposure assessment: A handbook for addressing variability and uncertainty in models and inputs. Plenum, 352 pp

Efron B (1987) Better bootstrap confidence intervals. J Am Stat Assoc 82: 171–185

Efron B, Tibshirani RJ (1993) An introduction to bootstrap. Monographs on Statistics and Applied Probability, vol. 57. New York: Chapman & Hall, 436 pp

EPA (1999) User Manual for the EPA Third-Generation Air Quality Modeling System (Models-3 Version 3.0). Office of Research and Development, U.S. Environmental Protection Agency, Washington, D.C. 20460. EPA-600/R-99/055. Report available from: *www.epa.gov/asmdnerl/models3/doc/user/user.html*

EPA (2002) Example application of modeling toxic air pollutants in an urban area. EPA-454/R-02-003, OAQPS/EPA, RTP, NC 27711, *www.epa.gov/scram001/tt25htm#toxics* (91 pages + appendix)

Fox DG (1984) Uncertainty in air quality modeling. B Amer Meteor Soc 65: 27–36

Gates WL, Boyle JS, Covey C, Dease CG, Doutriaux CM, Drach RS, Fiorino M, Gleckler PJ, Hnilo JJ, Marlais SM, Phillips TJ, Potter GL, Santer BD, Sperber KR, Taylor KE, Williams DN (1999) An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). Bull Amer Meteor Soc 80: 29–55

Grell GA, Dudhia J, Stauffer DR (1994) A description of the Fifth-Generation Penn State/NCAR mesoscale model (MM5). NCAR Tech. Note NCAR/TN 398 + STR, 138 pp. Available from NCAR, P.O. Box 3000, Boulder, CO 80307

Hamill TM, Mullen SL, Snyder C, Toth Z, Baumhefner DP (2000) Ensemble forecasting in the short to medium range: Report from a workshop. Bull Amer Meteor Soc 81: 2653–2664

Hanna SR (1988) Air quality model evaluation and uncertainty. J Air Poll Control Assoc 38: 406–412

Hanna SR (1989) Confidence limits for air quality model evaluations as estimated by bootstrap and jackknife resampling methods. Atmos Environ 23: 1385–1398

Hanna SR (1993) Uncertainties in air quality model predictions. Bound Layer Meteorol 62: 3–20

Hanna SR, Britter RE, Franzese P (2003) A baseline urban dispersion model evaluated with Salt Lake City and Los Angeles Tracer data. Atmos Environ (submitted)

Hanna SR, Chang JC, Strimaitis DG (1993) Hazardous gas model evaluation with field observations. Atmos Environ 27A: 2265–2285

Hanna SR, Egan BA, Purdum J, Wagler J (2000) Evaluation of the ADMS, AERMOD, and ISC3 Dispersion Models with the Kincaid, Indianapolis, Lovett, Sweeny, and Duke Forest Field Data Sets. Int J Environ Poll

Hanna SR, Strimaitis DG, Chang JC (1991) Hazard response modeling uncertainty (A quantitative method), vol. I: User's guide for software for evaluating hazardous gas dispersion models; vol. II: Evaluation of commonly-used hazardous gas dispersion models; vol. III: Components of uncertainty in hazardous gas dispersion models. Report no. A119/A120, prepared by Earth Tech, Inc., 196 Baker Avenue, Concord, MA 01742, for Engineering and Services Laboratory, Air Force Engineering and Services Center, Tyndall Air Force Base, FL 32403; and for American Petroleum Institute, 1220 L Street, N.W., Washington, D.C., 20005

Hanna SR, Yang R (2001) Evaluations of mesoscale model predictions of near-surface winds, temperature gradients, and mixing depths. J Appl Meteor 40: 1095–1104

Hanna SR, Davis JM (2002) Evaluation of a photochemical grid model using estimates of concentration probability density functions. Atmos Environ 36: 1793–1798

Helton JC (1997) Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. J Stat Comp Simul 1997: 3–76

Hodur RM (1997) The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). Mon Wea Rev 125: 1414–1430

Ichikawa Y, Sada K (2002) An atmospheric dispersion model for the environmental impact assessment of thermal power plants in Japan – A method for evaluating topographical effects. J Air Waste Manage Assoc 52: 313–323

Klug W, Graziani G, Grippa G, Pierce D, Tassone C (1992) Evaluation of long range atmospheric transport models using environmental radioactivity data from the chernobyl accident. The ATMES Report. Elsevier Science Publishing

Lewellen WS, Sykes RI, Parker SF (1985) An evaluation technique which uses the prediction of both concentration mean and variance. Proc. DOE/AMS Air Pollution Model Evaluation Workshop, Savannah River Lab. Report No. DP-1701-1, Section 2, 24 pp

Lewellen WS, Sykes RI (1986) Analysis of concentration fluctuations from lidar observations of atmospheric plumes. J Clim Appl Meteor 25: 1145–1154

Lewellen WS, Sykes RI (1989) Meteorological data needs for modeling air quality uncertainties. J Ocean Atmos Tech 6: 759–768

Lorenz E (1963) Deterministic nonperiodic flow. J Atmos Sci 20: 130–141

Lumley JL, Panofsky HA (1964) The structure of atmospheric turbulence. New York: Wiley Interscience, 239 pp

McNally D, Tesche TW (1993) MAPS sample products. Alpine Geophysics, 16225 W. 74th Dr., Golden, CO 80403

Morgan MG, Henrion M (1990) Uncertainty. A guide to dealing with uncertainty in quantitative risk and policy analysis (with a chapter from M. Small). Cambridge University Press, 332 pp

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

Nappo CJ, Essa KSM (2001) Modeling dispersion from near-surface tracer releases at Cape Canaveral, FL. Atmos Environ 35: 3999–4010

Nasstrom JS, Sugiyama G, Ermak D, Leone JM Jr (2000) A real-time atmospheric dispersion modeling system. Proc. 11th Joint Conf. on the Application of Air Pollution Meteorology with the A&WMA, Amer Meteor Soc, Boston, MA

Olesen HR (2001) Ten years of harmonization activities: past, present, and future. 7th Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Belgirate, Italy. National Environmental Research Institute, Roskilde, Denmark *(www.harmo.org)*

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

Paine RJ, Lee RF, Brode R, Wilson RB, Cimorelli AJ, Perry SG, Weil JC, Venkatram A, Peters WD (1998) Model evaluation results for AERMOD. USEPA, RTP. NC 27711

Pielke RA (2002) Mesoscale meteorological modeling, 2nd ed. Academic Press, 676 pp

Pielke RA, Pearce RP (1994) Mesoscale modeling of the atmosphere. Meteorological Monographs No. 47. Amer Meteor Soc, Boston, MA, 167 pp

Saltelli A, Chan K, Scott EM (2000) Sensitivity analysis. Wiley, 475 pp

Seaman NL (2000) Meteorological modeling for air quality assessments. Atmos Environ 34: 2231–2259

Seigneur C, Pun B, Pai P, Louis J-F, Solomon P, Emery C, Morris R, Zahniser M, Worsnop D, Koutrakis P, White W, Tombach I (2000) Guidance for the performance evaluation of three-dimensional air quality modeling systems for particulate matter and visibility. J Air Waste Manage Assoc 50: 588–599

Sykes RI (1984) The variance in time-averaged samples from an intermittent plume. Atmos Environ 18: 121–123

Sykes RI (2002) The use of concentration fluctuation variance predictions in model evaluation. Proc. 12th Joint Conf. on the Applications of Air Pollution Meteorology with A&WMA, 20–24 May 2002, Norfolk, VA, Amer Meteor Soc, Boston, MA

Sykes RI, Parker SF, Henn DS, Cerasoli CP, Santos LP (1998) PC-SCIPUFF Version 1.2PD, Technical Documentation. Titan Corporation, Titan Research and Technology Division, ARAP Group, P.O. Box 2229, Princeton, NJ 08543-2229, 172 pp

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 106(D7): 7183–7192

Venkatram A (1979) The expected deviation of observed concentrations from predicted ensemble means. Atmos Environ 13: 1547–1549

Venkatram A (1984) The uncertainty in estimating dispersion in the convective boundary layer. Atmos Environ 18: 307–310

Venkatram A (1988) Topics in applied dispersion modeling. Lectures on Air Pollution Modeling, chap. 6. Amer Meteor Soc, Boston, MA, 390 pp

Warner S, Platt N, Heagy JF, Bradley S, Bieberbach G, Sugiyama G, Nasstrom JS, Foster KT, Larson D (2001) User-oriented measures of effectiveness for the evaluation of transport and dispersion models. Institute for Defense Analyses, IDA Paper P-3554, 815 pp. IDA, 1801 N. Beauregard Street, Alexandria, VA 22311-1772

Weil JC, Sykes RI, Venkatram A (1992) Evaluating air-quality models: review and outlook. J Appl Meteor 31: 1121–1145

Wilks DS (1995) Statistical methods in the atmospheric sciences. New York: Academic Press, 467 pp

Willmott CJ (1982) Some comments on the evaluation of model performance. Bull Amer Meteor Soc 63: 1309–1313

Wilson DJ (1995) Concentration fluctuations and averaging times in vapor clouds. American Institute of Chemical Engineers, 345 East 47th Street, New York, NY 10017, 208 pp

Corresponding author's address: J. C. Chang, School of Computational Sciences, George Mason University, MS 5B2, Fairfax, VA 22030-4444, USA (E-mail: jchang4@scs.gmu.edu)

Verleger: Springer-Verlag GmbH, Sachsenplatz 4–6, 1201 Wien, Austria. – Herausgeber: Prof. Dr. Reinhold Steinacker, Institut für Meteorologie und Geophysik, Universität Wien, Althanstraße 14, 1090 Wien, Austria. – Redaktion: Innrain 52, 6020 Innsbruck, Austria. – Satz und Umbruch: Thomson Press (India) Ltd., Chennai. – Druck und Bindung: Grasl Druck&Neue Medien, 2540 Bad Vöslau, Austria. – Verlagsort: Wien. – Herstellungsort: Bad Vöslau. – Printed in Austria.

Offenlegung gem. § 25 Abs. 1 bis 3 Mediengesetz: Unternehmensgegenstand: Verlag von wissenschaftlichen Büchern und Zeitschriften
An der Springer-Verlag GmbH ist beteiligt: Springer-Verlag GmbH, Heidelberger Platz 3, 14197 Berlin, Germany, zu 100%, Geschäftsführer: Rudolf Siegle, Sachsenplatz 4–6, 1201 Wien, Austria