Technical Descriptions and User's Guide for the BOOT Statistical Model Evaluation Software Package, Version 2.0

by

Joseph C. Chang[1,2] and Steven R. Hanna[3]

[1]Comprehensive Atmospheric Modeling Program
School of Computational Sciences
George Mason University
4400 University Drive, MS 5B2
Fairfax, VA 22030-4444

[2]also affiliated with
Homeland Security Institute
2900 South Quincy Street, Suite 800
Arlington, VA 22206-2231

[3]Harvard School of Public Health
Landmark Center, Room 404J
401 Park Drive
Boston, MA 02215-0013

July 10, 2005

Table of Contents

1. Introduction

Computer models are powerful tools that are often used to study scientific or social phenomena. For example, various sophisticated models are applied to analyze the economic development of a country, to study the spread of a disease, to study the interaction among galaxies, to forecast tomorrow's weather and next decade's climate, and to predict the potential health effects due to the emission of toxic pollutants into the atmosphere. It is important that these models be properly evaluated in order to demonstrate their fidelity in simulating the phenomena of interest.

This report mainly describes the technical basis and user's instructions for the BOOT statistical model evaluation package. Although BOOT has been primarily used to evaluate the performance of air dispersion models, the same procedures and approaches implemented in BOOT also apply to other types of models.

A model can be evaluated in at least three areas: statistical (*e.g.*, Hanna *et al.* 1993), scientific (*e.g.*, Nappo *et al.* 1998), and operational (*e.g.*, Chang *et al.* 1998). Statistical evaluation mainly involves comparing model predictions with observations. It provides concise information on model performance, but caution must be exercised to avoid a possible situation where the model produces the right answers but for the wrong reasons (*i.e.*, compensating errors). Scientific evaluation examines model algorithms, physics, and assumptions in detail for their accuracy, efficiency, and sensitivity; and requires in-depth knowledge of the model's scientific basis. Operational evaluation mainly considers issues related to the user-friendliness of the model, such as the user's guide, the user interface, error checking of input data, internal model diagnostics, output display, and consistency of application by multiple users. BOOT is mainly for statistical evaluation. However, as described later, some graphical techniques also allow for preliminary scientific evaluation.

Statistical evaluation calls for the comparison of model predictions with certain reference states, which in most cases are simply "observations." Observations can be directly measured by instruments, or can be 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 clearly 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 pairing options may result in different conclusions.

The BOOT statistical model evaluation software package is originally based on recommendations by Hanna (1989). An earlier version (V1.0) of BOOT is described in Hanna *et al.* (1991). The software has been extensively used by scientists in assessing model performance (*e.g.*, Ichikawa and Sada 2002; Nappo and Essa 2001; Mosca *et al.* 1998). In the European Initiative on "Harmonisation within Atmospheric Dispersion

Modelling for Regulatory Purposes," the software was also used as a common framework upon which the performance of many dispersion models was intercompared (Olesen 2001). The current version (V2.0) of BOOT incorporates some recent upgrades described in Chang (2002), and Chang and Hanna (2004). These upgrades mainly include the consideration of additional performance measures, and the implementation of the ASTM (2000) model evaluation procedure.

This report is more than just a "software user's guide." It describes the life cycle of a comprehensive model evaluation exercise, including the definition of the evaluation objective (Section 2), exploratory data analysis (Section 3), and the methodology of statistical performance evaluation (Sections 4 and 5). Section 6 gives the actual software user's guide for BOOT. Section 7 provides a demonstration of BOOT. A summary is given in Section 8.

2. Evaluation Objective

An evaluation objective must first be clearly defined for any model evaluation study. Similarly, when conducting a statistical significance test, a null hypothesis should also be clearly defined. Depending on the study's objective and emphasis, there are a number of potential outputs of dispersion modeling that could be evaluated, such as

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 and shape 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
For dosage (concentration integrated with time):
• The maximum dosage along a sampling line • The cross-wind integrated dosage along a sampling line
For cloud timing:
• The cloud arrival and departure times, and the effective cloud speed

For example, for regulatory applications with routine emissions the primary objective might be how well a model simulates the maximum hourly-averaged concentration anywhere on the sampling network, whereas for accidental releases of

flammable substances the instantaneous maximum is more important than the average concentration. Selection of an appropriate averaging time is thus quite important. While the location of the maximum impact may be of less importance for regulatory applications, it is important for environmental justice applications to evaluate model predictions at specific locations such as densely-populated neighborhoods. For military applications, the location and shape of the footprint of a chemical warfare agent cloud are important, since a military commander can use that information to decide whether to order troops to put on protective gears, or to order responsive troop maneuvers. For a forensic study concerning whether individuals were impacted by a hazardous gas cloud, it might be of interest to correctly estimate the cloud arrival and departure times. Moreover, for any field experiment, there are usually practical constraints resulting in only a limited number of the above evaluation objectives that can actually be considered. In other words, a certain evaluation objective might be desirable but cannot be met due to lack of data. All of these issues should be carefully considered so that a meaningful model evaluation objective can be defined.

3. Exploratory Data Analysis

Before calculating various statistical performance measures (or metrics), it is recommended that exploratory data analysis be first performed by simply plotting the data in different ways. Human eyes can often glean much more inherent information from these plots than pure statistics. These plots can also provide clues as to why a model performed in a certain way. Some of the commonly-used plots are

- Scatter plots (Fig. 1)
- Quantile-quantile plots (Fig. 2)
- Residual (box) plots (Figs. 3 and 4)
- Conditional scatter plots (Fig. 5)

Depending on such factors as the range and amount of data, and the information to be conveyed, a combination of plots is usually necessary.

These plots are demonstrated below with a sample, artificial database listed in Table 1. The sample database contains the maximum hourly observed concentrations anywhere on the monitoring network, and the predictions from three models. The database also contains the corresponding values for time of day (hour), representative wind speed (m s^{-1}), mixing height (m), and the Pasquill-Gifford stability class.

3.1. Scatter Plot

Figure 1 shows the scatter plots of observed versus predicted concentrations for the three models. The scatter plots provide a first look of the overall model performance. Direct visual inspection shows that, as will also be verified by quantitative calculations in Section 7, Model-A appears to have the best performance. The highest predictions given

by Model-A and Model-B also correspond to high observed values. This merit may be of primary importance because of the obvious impacts on the public health due to high pollutant concentrations. On the other hand, the ability of a model to correctly predict low concentrations may sometimes also be important for highly toxic chemicals such as chemical and biological warfare agents. Model-C is seen to show almost no correlation between observations and predictions.

3.2. Quantile-Quantile Plot

Dispersion models are often used for air pollution regulatory purposes, which in the U.S. typically involve multiple-year simulations on an hourly basis for the entire time period. The highest short-term (*e.g.*, 1-hr and 8-hr) concentrations and the average long-term (*e.g.*, 1-yr) concentrations or dosages are then investigated for possible violation. Thus, it is also of interest to find out whether a model can generate a concentration distribution that is similar to the observed, especially at the range of high concentrations. To generate the quantile-quantile plots shown in Fig. 2, predicted and observed concentrations are separately ranked using the dataset in Fig. 1. It can be seen that although Fig. 1 shows that Model-C has a poor performance in terms of correlation, the model's ability to simulate the observed concentration distribution appears better in Fig. 2. For example, the highest observed concentrations and Model-C's predicted concentrations have similar magnitude. However, the model clearly overpredicts overall, and may be correctly predicting the values of the highest few observed concentrations but for the wrong reasons.

3.3. Residual (Box) Plots

Scatter plots and quantile-quantile plots often do not provide an adequate understanding of why a model performed in a certain way. This question could be addressed by reviewing the model algorithms in great detail. The question can also be addressed using residual analyses, which often employ box plots. Figures 3 and 4 show the box plots of model residuals, defined as the ratio of predicted (C_p) to observed (C_o) concentrations, for Model-A and Model-C, respectively, as functions of four of the model independent variables: time of day, ambient wind speed, mixing height, and atmospheric stability. Again, these are the same sets of model predictions and observations presented in Figs. 1 and 2. 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 symbol. The significant points for each box indicate the 2^{nd}, 16^{th}, 50^{th}, 84^{th}, and 98^{th} percentiles of the cumulative distribution of the n points considered in the bin of data used in the box. A good performing model should not show any trend with independent variables. This ideal behavior is evident for Model-A by visual inspection of Fig. 3. However, visual inspection of Fig. 4 shows slight trends for Model-C with time of day and atmospheric stability. Model-C generally overpredicts during nighttime hours when the atmosphere is stable. As a consequence, the dispersion algorithms in Model-C under nighttime, stable conditions should be carefully reviewed to identify potential flaws.

3.4. Residual Scatter Plots

Box plots are useful in summarizing the distribution of model residuals when the number of data points in each bin is large. However, when the number is small, it makes more sense to simply plot all the data points without the use of box symbols. Figure 5 is essentially the same as Fig. 3, except that all data points (model residuals) for Model-A are now represented on a scatter plot without any binning. In order to convey additional information, different symbols are used to further represent different ranges of a second independent variable on each panel of Fig. 5. Therefore, Fig. 5 can be considered a *conditional* residual plot, and is similar to a two-way contingency table. The figure shows that most of the data points fall within the factor-of-two dashed lines. The two hours where Model-A greatly underpredicts (by two orders of magnitude) can be easily identified to have the following characteristics: near midnight, under stable conditions, wind speed ~4 m s^{-1}, and mixing height = 2000 and 0 m. A review of the raw data in Table 1 further indicates that the two hours are consecutive.

Model-A's poor performance for these two hours can be attributed to inadequate model physics or to incorrect or unrepresentative input data. Before one decides to "blame" the model physics for poor performance, model inputs should be carefully reviewed. First of all, the mixing heights for the two consecutive hours are 2000 and 0 m, indicating a sudden collapse of the atmospheric mixed layer. This phenomenon should be confirmed by re-examining the original temperature profiles from which mixing heights were estimated. Moreover, the two hours have a moderate wind speed of ~4 m s^{-1} but are characterized as "very stable" (stability class = 6, see Table 1). This is unusual, since very stable conditions are usually associated with lower wind speeds. Thus, it is likely that the model inputs for these two hours are incorrect, and that validated model inputs might lead to an improved model performance.

In conclusion, it can be seen that exploratory data analysis, although qualitative in nature, can reveal much valuable information.

4. Quantitative Performance Measures Implemented in BOOT Software

Hanna (1989) and Hanna *et al.* (1991 and 1993) recommend a set of quantitative statistical performance measures for evaluating air dispersion models, and implement the procedures in a software package called BOOT. These 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).

The quantitative performance measures considered in the BOOT software are described in this section, together with some recent enhancements mentioned in Chang

(2002), and Chang and Hanna (2004). User's instructions for the software are given in Section 6.

4.1 Definitions of Performance Measures

4.1.1 Original Performance Measures

The original BOOT software (Hanna *et al.* 1991) incorporated some basic statistical performance measures, which at the time were being suggested by the U.S. Environmental Protection Agency (EPA; Cox and Tikvart 1990) as an basis for air quality model evaluation. These performance measures 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):

$$\text{FB} = \frac{\left(\overline{C_o} - \overline{C_p}\right)}{0.5\ \left(\overline{C_o} + \overline{C_p}\right)} \tag{1}$$

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

$$\text{NMSE} = \frac{\overline{\left(C_o - C_p\right)^2}}{\overline{C_o}\ \overline{C_p}} \tag{3}$$

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

$$\text{R} = \frac{\overline{\left(C_o - \overline{C_o}\right)\left(C_p - \overline{C_p}\right)}}{\sigma_{C_p}\sigma_{C_o}} \tag{5}$$

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

where C_p denotes model predictions, C_o denotes observations, overbar ($\overline{C}$) denotes the average over the dataset, and σ_c denotes the standard deviation over the dataset. $\overline{C_o} - \overline{C_p}$ is used to define mean bias, rather than $\overline{C_p} - \overline{C_o}$, because that was the definition used by the U.S. EPA.

A perfect model would have MG, VG, R, and FAC2 = 1.0; and FB and NMSE = 0.0. The properties of these measures are further described in Section 4.2. The above six performance measures are by no means exhaustive. Other measures can also be used if necessary for a specific application or concern.

In the above, we have simply assumed that the evaluation dataset contains pairs of C_p and C_o, and that they represent averages over an averaging time, T_a. The pairing is completely generic, and can be:

- 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 location);
- 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); or
- in both time and space.

Pairing in both time and space is clearly most stringent. Decisions concerning these pairing options are obviously part of the evaluation objective (see Section 2) that should be clearly defined for any evaluation exercise.

4.1.2 New Performance Measures

In addition to the six basic performance measures defined above, the new BOOT software also considers the following measures, which are all closely related to FB. Firstly, the new BOOT software considers the overpredicting (or false-positive) and underpredicting (or false-negative) components of FB. To show this, Eq. (1) can be written as

$$FB = \frac{\frac{1}{N}\cdot\sum_i\left(C_{oi}-C_{pi}\right)}{\frac{1}{2N}\cdot\sum_i\left(C_{oi}+C_{pi}\right)} = \frac{\sum_i\left(C_{oi}-C_{pi}\right)}{\frac{1}{2}\cdot\sum_i\left(C_{oi}+C_{pi}\right)} \tag{7}$$

where C_{oi} is the i^{th} observed value, C_{pi} is the i^{th} predicted value, and N is the total number of observation-prediction pairs. The above equation can be rearranged as follows.

$$\begin{aligned} FB &= \frac{\frac{1}{2}\sum_i\left[\left|C_{oi}-C_{pi}\right|+\left(C_{oi}-C_{pi}\right)\right]}{\frac{1}{2}\cdot\sum_i\left(C_{oi}+C_{pi}\right)} - \frac{\frac{1}{2}\sum_i\left[\left|C_{oi}-C_{pi}\right|+\left(C_{pi}-C_{oi}\right)\right]}{\frac{1}{2}\cdot\sum_i\left(C_{oi}+C_{pi}\right)} \\ &= FB_{FN} - FB_{FP} \end{aligned} \tag{8}$$

where

$$FB_{FN} = \frac{\frac{1}{2}\sum_i \left[\left| C_{oi} - C_{pi} \right| + \left(C_{oi} - C_{pi} \right) \right]}{\frac{1}{2} \cdot \sum_i \left(C_{oi} + C_{pi} \right)} \tag{9}$$

$$FB_{FP} = \frac{\frac{1}{2}\sum_i \left[\left| C_{oi} - C_{pi} \right| + \left(C_{pi} - C_{oi} \right) \right]}{\frac{1}{2} \cdot \sum_i \left(C_{oi} + C_{pi} \right)} \tag{10}$$

FB_{FN} can be considered as the underpredicting (false-negative) component of FB, *i.e.*, only those (C_o, C_p) pairs with $C_p < C_o$ are considered in the calculation. Similarly, FB_{FP} can be considered as the overpredicting (false-positive) component of FB, *i.e.*, only those (C_o, C_p) pairs with $C_p > C_o$ are considered in the calculation. Equations (9) and (10) show that FB_{FN} and FB_{FP} are always non-negative. Properties of FB_{FN} and FB_{FP} are further described in the next section.

As shown in Eq. (8), FB equals the difference between FB_{FN} and FB_{FP}. The sum of FB_{FN} and FB_{FP}, on the other hand, is often called the absolute fractional bias, AFB, another commonly-used performance measure (*e.g.*, ASTM 2000), where

$$AFB = FB_{FN} + FB_{FP} = \frac{\left| \overline{C_o} - \overline{C_p} \right|}{0.5 \left(\overline{C_o} + \overline{C_p} \right)} \tag{11}$$

The numerator of Eq. (11), $\left| \overline{C_o} - \overline{C_p} \right|$, is often called the mean absolute error, MAE (*e.g.*, Wilks 1995).

The underpredicting (false-negative) and overpredicting (false-positive) components of MG (*i.e.*, MG_{FN} and MG_{FP}) can be similarly defined by considering only those ($\ln C_o$, $\ln C_p$) pairs with $\ln C_p < \ln C_o$ and $\ln C_p > \ln C_o$, respectively. In other words,

$$MG_{FN} = \exp\left[\frac{1}{2N} \sum_i \left[\left| \ln C_{oi} - \ln C_{pi} \right| + \left(\ln C_{oi} - \ln C_{pi} \right) \right] \right] \tag{12}$$

$$MG_{FP} = \exp\left[\frac{1}{2N} \sum_i \left[\left| \ln C_{oi} - \ln C_{pi} \right| + \left(\ln C_{pi} - \ln C_{oi} \right) \right] \right] \tag{13}$$

$$MG = MG_{FN} / MG_{FP} \tag{14}$$

Another way of evaluating model performance is to consider the so-called Figure of Merit in Space (FMS; *e.g.*, Wilks 1995; Mesinger 1996; Mosca *et al.* 1998; Ebert 2001), defined as

$$\mathrm{FMS} = \frac{\mathrm{A_p} \cap \mathrm{A_o}}{\mathrm{A_p} \cup \mathrm{A_o}} \tag{15}$$

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. 6). FMS is also often called the threat score (McNally and Tesche 1993). The portion of A_p that is outside the intersection ($A_p \cap A_o$) can be considered as the false-positive (or overpredicting) area, A_{FP}, *i.e.*, a hazard area predicted by the model but not observed. The portion of A_o that is outside the intersection ($A_p \cap A_o$) can be considered as the false-negative (or underpredicting) area, A_{FN}, *i.e.*, a hazard area observed but not predicted by the model.

Warner *et al.* (2001) suggest a more general expression of FMS, or a two-dimensional (2-D) Measure of Effectiveness (MOE), where two components are used to indicate model performance,

$$\begin{aligned} 2-\mathrm{D\ MOE} &= \left(\mathrm{MOE_{FN}}, \mathrm{MOE_{FP}}\right) \\ &= \left(\frac{\mathrm{A_p} \cap \mathrm{A_o}}{\mathrm{A_o}}, \frac{\mathrm{A_p} \cap \mathrm{A_o}}{\mathrm{A_p}} \right) \\ &= \left(\frac{\mathrm{A_p} \cap \mathrm{A_o}}{\mathrm{A_p} \cap \mathrm{A_o} + \mathrm{A_{FN}}}, \frac{\mathrm{A_p} \cap \mathrm{A_o}}{\mathrm{A_p} \cap \mathrm{A_o} + \mathrm{A_{FP}}} \right) \end{aligned} \tag{16}$$

MOE_{FN} measures a degree of underprediction (false-negative), whereas MOE_{FP} measures a degree of overprediction (false-positive). The two components, however, are normalized differently, *i.e.*, MOE_{FN} by A_o, and MOE_{FP} by A_p. On the other hand, FB_{FN} and FB_{FP} are normalized by the same quantity (Eqs. (9) and (10)).

It is not always straightforward to define a contour area for a field experiment due to reasons such as limited samplers and limited measuring arcs. One possible surrogate for the area estimate is data summation (Warner *et al.* 2001, Chang 2002). In this case, A_{FN} is given by the sum of the difference between C_o and C_p for those (C_o, C_p) pairs with $C_p < C_o$ (*i.e.*, false negative),

$$\mathrm{A_{FN}} = \frac{1}{2} \sum_{\mathrm{i}} \left[\left| \mathrm{C_{oi}} - \mathrm{C_{pi}} \right| + \left(\mathrm{C_{oi}} - \mathrm{C_{pi}} \right) \right] \tag{17}$$

A_{FP} is given by the sum of the difference between C_o and C_p for those (C_o, C_p) pairs with $C_p > C_o$ (*i.e.*, false positive),

$$A_{FP} = \frac{1}{2}\sum_i \left[\left| C_{oi} - C_{pi} \right| + \left(C_{pi} - C_{oi} \right) \right] \tag{18}$$

$A_p \cap A_o$ is given by the difference between the smaller of (C_o, C_p) and a threshold, C_t, for all data pairs,

$$\begin{aligned} A_p \cap A_o &= \frac{1}{2}\sum_i \left[\min\left(C_{oi}, C_{pi} \right) - C_t \right] \\ &= \frac{1}{2}\sum_i \left[\left(C_{oi} + C_{pi} \right) - \left| C_{oi} - C_{pi} \right| - C_t \right] \end{aligned} \tag{19}$$

Substituting Eqs. (17) through (19) into Eqs. (8) through (10), while assuming $C_t = 0$, would yield alternative expressions of FB, FB_{FN}, and FB_{FP} in terms of area estimates, *i.e.*,

$$FB = FB_{FN} - FB_{FP} = \frac{A_{FN} - A_{FP}}{A_p \cap A_o + \frac{1}{2}\cdot\left(A_{FN} + A_{FP} \right)} \tag{20}$$

$$FB_{FN} = \frac{A_{FN}}{A_p \cap A_o + \frac{1}{2}\cdot\left(A_{FN} + A_{FP} \right)} \tag{21}$$

$$FB_{FP} = \frac{A_{FP}}{A_p \cap A_o + \frac{1}{2}\cdot\left(A_{FN} + A_{FP} \right)} \tag{22}$$

Comparison of Eq. (16) with Eqs. (21) and (22) suggests similarity between (MOE_{FN}, MOE_{FP}) and (FB_{FN} and FB_{FP}). Chang (2002) shows that the two pairs of performance measures are indeed related, as given by the following formulas where FB_{FN} and FB_{FP} are expressed as a function of MOE_{FN} and MOE_{FP}:

$$FB_{FN} = \frac{2\cdot MOE_{FP}\cdot\left(1 - MOE_{FN}\right)}{MOE_{FN} + MOE_{FP}} \tag{23}$$

$$FB_{FP} = \frac{2\cdot MOE_{FN}\cdot\left(1 - MOE_{FP}\right)}{MOE_{FN} + MOE_{FP}} \tag{24}$$

$$FB = FB_{FN} - FB_{FP} = \frac{-2\cdot\left(MOE_{FN} - MOE_{FP} \right)}{MOE_{FN} + MOE_{FP}} \tag{25}$$

Equations (23) and (24) can also be inverted to express MOE_{FN} and MOE_{FP} as a function of FB_{FN} and FB_{FP}:

$$MOE_{FN} = \frac{1 - \frac{FB_{FN}}{2} - \frac{FB_{FP}}{2}}{1 + \frac{FB_{FN}}{2} - \frac{FB_{FP}}{2}} = \frac{2 - FB_{FN} - FB_{FP}}{2 + FB} \quad (26)$$

$$MOE_{FP} = \frac{1 - \frac{FB_{FN}}{2} - \frac{FB_{FP}}{2}}{1 - \frac{FB_{FN}}{2} + \frac{FB_{FP}}{2}} = \frac{2 - FB_{FN} - FB_{FP}}{2 - FB} \quad (27)$$

In summary, the new BOOT software calculates the following performance measures: FB, MG, NMSE, VG, R, FAC2, FB_{FN}, FB_{FP}, MG_{FN}, MG_{FP}, MOE_{FN}, and MOE_{FP} (Eqs. (1) through (6), (9), (10), (12), (13), and (16), respectively). It has also been shown that FB_{FN}, FB_{FP}, MOE_{FN}, and MOE_{FP} are closely related to one another (Eqs. (23), (24), (26), and (27)) and to FB. AFB is not directly calculated by BOOT since it is readily given by the sum of FB_{FN} and FB_{FP} (Eq. (11)).

4.2. Properties of Performance Measures

It is necessary to consider multiple performance measures, 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 distribution of the variable of interest. The distribution resembles a log-normal distribution for atmospheric pollutant concentrations. In this case, linear measures FB and NMSE are strongly influenced by infrequently occurring high observed and predicted concentrations, whereas logarithmic measures MG and VG provide a more balanced treatment of extremely high and low values. Therefore, for a dataset where both predicted and observed concentrations vary by many orders of magnitude, MG and VG would probably be more appropriate. FAC2, on the other hand, is the most robust measure, because it is not overly influenced by high and low outliers.

However, MG and VG are also known to be strongly influenced by extremely low values (*e.g.*, Hanna and Chang 2001, Chang *et al.* 2001), and are undefined for zero values. These low and zero values are not uncommon in dispersion modeling, where a low concentration value might be at a receptor that the plume has missed. Therefore, when calculating MG and VG, it is useful to impose a minimum threshold for data values. It is recommended that an instrument threshold, such as the limit of detection (LOD), be used as the lower bound for both C_p and C_o. In this case, whenever C_p is lower than the threshold, it is set to the LOD; and whenever C_o is lower than the threshold, it is also set to the LOD.

FB and MG are measures of mean bias and indicate only systematic errors, whereas NMSE and VG are measures of 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. 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 an FB = 0. A solution to the problem is to consider a modified version of FB where the two error components (*i.e.*, overprediction and underprediction) are separately considered (see Eq. (8)).

The correlation coefficient, R, reflects the linear relationship between two variables and is thus insensitive to either an additive or a multiplicative factor. That is, if $C_p = \alpha + \beta C_o$, where α and β (>0) are arbitrary constants, R will always equal 1.0 between C_p and C_o. Therefore, a perfect correlation coefficient is only a necessary, but not sufficient, condition for a perfect model. Also, R is sensitive to a few aberrant data pairs (*e.g.*, Isaaks and Srivastava 1989). 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.

Moreover, it is typical for short-range dispersion field experiments to have concentration data measured along concentric arcs, 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 is capable of simulating. Therefore, R is less useful in a typical evaluation exercise for dispersion models when data are arranged in arcs. On the other hand, it might be more useful when gridded fields are involved (*e.g.*, McNally and Tesche 1993).

It is sometimes 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 (Conover 1980; EPA 1997). R_{rank} is a more robust measure than R. Large differences between R and R_{rank} are usually due to the locations of extreme data pairs on the scatter plot. A high value of R_{rank} and a low value of R may indicate that there are a few erratic pairs in an otherwise good correlation. A low value of R_{rank} and a high value of R may be a result of a few extreme pairs close to the diagonal line on a scatter plot.

It is also sometimes useful to consider the "real" correlation between two variables after the influence of a third variable is removed. In this case, the partial correlation coefficient between variables X_1 and X_2, while holding the third variable X_3 constant, is (Panofsky and Brier 1958):

$$R_{12,3} = \frac{R_{12} - R_{13}R_{23}}{\sqrt{1-R_{13}^2}\sqrt{1-R_{23}^2}} \tag{28}$$

where R_{12}, R_{13}, and R_{23} are the correlation coefficients between X_1 and X_2, between X_1 and X_3, and between X_2 and X_3, respectively.

As mentioned above, both NMSE and VG account for systematic and random errors. It can be shown that the minimum NMSE, *i.e.*, without any random errors, for a certain value of FB is given by the following expression (Hanna *et al.* 1991, Chang 2002):

$$NMSE_{min} = \frac{4FB^2}{4 - FB^2} \tag{29}$$

Similarly, the minimum possible VG, *i.e.*, without any random errors, for a certain value of MG is given by the following expression (Hanna *et al.* 1991, Chang 2002):

$$VG_{min} = \exp\left(\left(\ln MG\right)^2\right) \tag{30}$$

Equation (8) suggests that the traditional FB can be partitioned into two orthogonal components, FB_{FN} and FB_{FP}, where the difference between the two gives the original FB. Thus, a model's performance can be indicated in a two-dimensional diagram with the coordinates given by FB_{FN} and FB_{FP} (Fig. 7), where the x-axis (FB_{FN}) denotes the degree of false negative, and the y-axis (FB_{FP}) denotes the degree of false positive. All points along a line whose slope is 1.0 will have the same value of the one-dimensional FB *(i.e.*, $FB_{FN} - FB_{FP} = FB$ = constant). Figure 7 shows three lines (dashed) that correspond to FB = 0, FB = 2/3, and FB = –2/3. Therefore, the diagonal shaded band can be considered as the "factor-of-two" region in the 2-D FB diagram, where any model that is located inside the band will have a one-dimensional FB that is between ±2/3, *i.e.*, with a mean bias within a factor of two of the observed (see also Eq. (31) below).

Chang (2002) describes other features of the 2-D FB diagram:

- The diagram covers a triangular area, because $FB_{FN} + FB_{FP}$ must be ≤ 2 (see Eqs. (21) and (22)).
- A perfect model would be located at the origin of the diagram, *i.e.*, (FB_{FN}, FB_{FP}) = (0, 0).
- Compensating errors can be easily identified in the diagram. Even though a model whose predictions are completely out of phase with observations (*i.e.*, $A_{FN} = A_{FP}$, and $A_p \cap A_o = 0$) will lead to FB = 0, the model will be located at (FB_{FN},

FB_{FP}) = (1, 1), clearly distinguishable from the perfect-agreement location, *i.e.*, (FB_{FN}, FB_{FP}) = (0, 0).

- The hypotenuse of the triangle indicates that there is no overlap between predictions and observations, *i.e.*, predictions are zero whenever observations are finite, and vice versa.
- The x-axis of the diagram means that predictions are systematically lower than observations (*i.e.*, systematic underprediction, or no false positive).
- The y-axis of the diagram means that predictions are systematically higher than observations (*i.e.*, systematic overprediction, or no false negative).
- The point of (FB_{FN}, FB_{FP}) = (2, 0) means that predictions are zero everywhere but all observations are finite.
- The point of (FB_{FN}, FB_{FP}) = (0, 2) means that observations are zero everywhere but all predictions are finite.

Likewise, a similar diagram in the 2-D MOE space can also be used to indicate model performance (Warner *et al.* 2001).

Equations (26) and (27) describe the relationship between (FB_{FN}, FB_{FP}) and (MOE_{FN}, MOE_{FP}). Figure 8 depicts this relationship, which graphically shows that once FB_{FN} and FB_{FP} are known, MOE_{FN} and MOE_{FP} are also known. The figure also shows that $MOE_{FN} = 1$ is asymptotic to $FP_{FN} = 0$, and that $MOE_{FP} = 1$ is asymptotic to $FP_{FP} = 0$.

4.3. Interpretations of FB, MG, NMSE, and VG

The FB, MG, NMSE, and VG defined in Eqs. (1) through (4) quantitatively define model performance. However, direct quotation of their values are often not that informative. For example, it will be difficult for a user to discern what NMSE = 9 and VG = 13 mean. As a result, it is recommended that the values of FB, NMSE, MG, and VG be further interpreted in terms of a measure that is more easily comprehended, such as the equivalent ratio of C_p to C_o. These interpretations are briefly described below, where Chang (2002) provides more details.

First of all, Eq. (1) can be easily rearranged and becomes

$$\frac{\overline{C_p}}{\overline{C_o}} = \frac{1-\frac{1}{2}FB}{1+\frac{1}{2}FB} \tag{31}$$

Therefore, for example, FB = 2/3 would correspond to a factor of two underprediction, and FB = –2/3 would correspond to a factor of two overprediction.

To interpret NMSE, *assume C_p and C_o are constant*, then $\overline{C_p} = C_p$, $\overline{C_o} = C_o$, and $\overline{\left(C_o - C_p\right)^2}$ in Eq. (3) equals $\left(\overline{C_o} - \overline{C_p}\right)^2$. Equation (3) can then be expressed as

$$\frac{\overline{C_p}}{\overline{C_o}} = \frac{2 + \text{NMSE} \pm \sqrt{\left(2 + \text{NMSE}\right)^2 - 4}}{2} \tag{32}$$

In this case, for example, NMSE = 0.5 would correspond to an equivalent factor of two mean bias. Since NMSE involves the square of the difference between C_p and C_o, it does not differentiate whether the factor of two mean bias is underprediction or overprediction.

Equation (2) can be shown to be equivalent to

$$\frac{\left\langle C_p \right\rangle}{\left\langle C_o \right\rangle} = \frac{1}{\text{MG}} \tag{33}$$

where $\left\langle C_o \right\rangle$ and $\left\langle C_p \right\rangle$ are geometric means of C_o and C_p. As a result, for example, a factor of two mean bias would mean MG = 0.5 or 2.0, and MG = 3.0 would mean a factor of three underprediction.

One way to relate the value of VG to a more easily understood quantity is to *assume that the ratio of C_p/C_o equals a constant*, A, which amounts to ignoring the random scatter between C_p and C_o. In this case, $C_p/C_o = A = \left\langle C_p \right\rangle / \left\langle C_o \right\rangle$, where $\langle\ \rangle$ again represents the geometric mean. Then it can be shown that

$$\frac{\left\langle C_p \right\rangle}{\left\langle C_o \right\rangle} = \exp\left[\pm\sqrt{\ln \text{VG}}\right] \tag{34}$$

In other words, for example, a factor of two mean bias (*i.e.*, $\left\langle C_p \right\rangle / \left\langle C_o \right\rangle = 2.0$ or 0.5) would mean VG = 1.6, and VG = 12 would indicate a random scatter that is equivalent to roughly a factor of five mean bias (*i.e.*, $\left\langle C_p \right\rangle / \left\langle C_o \right\rangle = 4.84$, or $\left\langle C_o \right\rangle / \left\langle C_p \right\rangle = 4.84$).

Figure 9 depicts the relationships between the equivalent ratio of C_p to C_o and FB, NMSE, MG, and VG, as described in Eqs. (31) through (34). Note that Eqs. (32) and (34) are not exact since some assumptions were involved. These assumptions mainly provide a means to more easily interpret NMSE and VG. Equations (31) and (33), on the other hand, are exact relationships. In a BOOT demonstration to be described in Section

7, Eqs. (31) through (34) will be used to help interpret the model evaluation results, rather than simply quoting the values of FB, NMSE, MG, and VG.

In summary, FB = ±2/3, MG = 0.5 or 2.0, NMSE = 0.5, and VG = 1.6 would correspond to an equivalent factor-of-two mean bias. FB = ±4/3, MG = 0.2 or 5.0, NMSE = 3.2, and VG = 13.33 would correspond to an equivalent factor-of-five mean bias. Finally, due to the nature of VG, it tends to have a large value when compared to NMSE for a large discrepancy between C_p and C_o. For example, Eqs. (32) and (34) and Fig. 9 show that NMSE = 8 and VG = 200 would both indicate an equivalent factor-of-ten mean bias.

4.4. Model Acceptance Criteria

One inevitable question for model performance evaluation is "how good is good enough?" Typical magnitudes of the above performance measures and estimates of model acceptance criteria have been summarized by Chang and Hanna (2004) based on extensive experience with evaluating many models with many field data sets. It was concluded that, for comparisons of maximum concentrations on arcs (*i.e.*, unpaired in space) and for research-grade field experiments, "acceptable" performing models have the following typical performance measures:

- The fraction of predictions within a factor of two of observations is about 50% or greater (*i.e.*, FAC2 > 0.5).
- The mean bias is within ± 30% of the mean (*i.e.*, roughly $|FB| < 0.3$ or $0.7 < MG < 1.3$).
- The random scatter is about a factor of two to three of the mean (*i.e.*, roughly NMSE < 1.5 or VG < 4).

However, these are not firm guidelines and it is necessary to consider all performance measures in making a decision concerning model acceptance. Since most of these criteria are based on research grade field experiments, model performance would be expected to deteriorate as the quality of the inputs decreases, or as more stringent data pairing options (*e.g.*, paired in space and time) are used.

The criterion of $|FB| < 0.3$ is readily shown in Fig. 8 as the shaded area. This shaded area is essentially $|FB_{FN} - FB_{FP}| < 0.3$. The figure also shows a cross hatched area that corresponds to AFB < 0.3, or $FB_{FN} + FB_{FP} < 0.3$. It is clear that this is a much more stringent performance criterion for a model to satisfy.

4.5. Confidence Limits Estimated by Bootstrap Resampling

A model might appear to have skills based on, for example, a small normalized mean square error (NMSE). A model might also appear to have a better performance

than other models based on, for example, a smaller fractional bias (FB). Hence, there are two hypotheses that could be tested:

- When compared to observations, are a model's performance measures significantly different from zero at the 95% confidence level? (For geometric measures VG and MG, it is necessary to consider their logarithmic values instead, since by definition VG and MG will always be positive.)
- When comparing the performance of two models, are the differences in performance measures for the two models (*e.g.*, FB for Model-A minus FB for Model-B) significantly different from zero at the 95% confidence level?

The bootstrap resampling (Efron 1987, Efron and Tibshirani 1993) is used to estimate the confidence limits of a performance measure. Two types of confidence limits, Student's t and percentile, can be given by the bootstrap resampling (Efron and Tibshirani 1993).

Student's t confidence limits

With, say, 1000, bootstrap resamples, there will be 1000 estimates for a performance measure. These 1000 estimates are used to estimate the mean, μ, and the standard deviation, σ, for the performance measure. The 95% Student's t confidence limits are then given by the following standard formula:

$$\mu \pm t_{95\%}\,\sigma\left(\frac{N}{N-1}\right)^{1/2} \tag{35}$$

where N is the number of observation-prediction pairs, $t_{95\%}$ is the Student's t value at the 95% confidence level with N – 1 degrees of freedom.

Percentile confidence limits

Alternatively, the 2.5^{th} and 97.5^{th} percentiles of the cumulative distribution function of the 1000 estimates also provide another estimate of the 95% confidence limits. According to Efron and Tibshirani (1993), the percentile confidence limits are more robust than the Student's t confidence limits. The BOOT software calculates both types of confidence limits, but mainly uses the percentile confidence limits for the significance tests.

For geometric measures VG and MG, the Student's t confidence limits are first calculated with the μ and σ that are based on the logarithms of VG and MG. These confidence limits are used in significance tests. The exponentiation of these confidence limits is then used for reporting. Such transformation is unnecessary for the percentile confidence limits.

Below are some practical considerations concerning the bootstrap resampling:

- The dataset may appear in blocks, with each block corresponding to, for example the data from one field program or the data for one value of an independent variable such as wind speed, stability class, or downwind distance. In this case, resampling should be restricted to within each block to avoid introducing block-to-block variance.
- Resampling should also be done with replacement. That is, once a sample is drawn, it is allowed to be drawn again.
- Observed and predicted values should be sampled concurrently in order to maintain the relationship between them.

It is helpful to illustrate the above considerations. Suppose there are seven experiments (numbered 1 through 7), with seven corresponding pairs of observations (C_o) and model predictions (C_p),

Original Version of Data

C_{o1}	C_{p1}	Block 1
C_{o2}	C_{p2}	
C_{o3}	C_{p3}	
C_{o4}	C_{p4}	Block 2
C_{o5}	C_{p5}	
C_{o6}	C_{p6}	
C_{o7}	C_{p7}	

Experiments 1 through 3 belong to Block 1, and Experiments 4 through 7 belong to Block 2. Then one possible bootstrap sample of the original series is

One Possible Resampled Version of Data

C_{o2}	C_{p2}	Block 1
C_{o2}	C_{p2}	
C_{o1}	C_{p1}	
C_{o7}	C_{p7}	Block 2
C_{o5}	C_{p5}	
C_{o4}	C_{p4}	
C_{o4}	C_{p4}	

The above resampled series shows that:

- Resampling was restricted to within each block. In other words, the first three data pairs (Block 1) of the new series were drawn from Experiments 1 through 3, and the last four data pairs (Block 2) of the new series were drawn from Experiments 4 through 7.
- Some experiments (2 and 4 in this case) were drawn more than once (twice in this case), while some experiments (3 and 6 in this case) were not drawn at all.
- Observed and predicted values were drawn concurrently. For example, C_{o4} and C_{p4} were always drawn together.

On the other hand, the following series would *not* be created based on the resampling procedure prescribed above.

A Version of Data That Will Not be Resampled

C_{o7}	C_{p7}	Block 1
C_{o2}	C_{p2}	
C_{o1}	C_{p1}	
C_{o7}	C_{p7}	Block 2
C_{o5}	C_{p5}	
C_{o4}	C_{p1}	
C_{o4}	C_{p4}	

This is because:

- Resampling must be restricted to within each block, but this procedure was not followed in the above example. For instance, the first three data pairs (Block 1) include Experiment 7, which in fact belongs to Block 2.
- The observed and predicted values must be drawn concurrently. For example, the sixth data pair consists of C_{o4} and C_{p1}, *i.e.*, Experiment 4's observation was matched with Experiment 1's prediction, thus violating the concurrent resampling rule.

5. ASTM Procedure

5.1. Framework

The qualitative and quantitative procedures mentioned in Sections 3 and 4, respectively, 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) suggest that there is a fundamental question in that most dispersion models generate ensemble-mean predictions (either explicitly or implicitly), whereas observations correspond to realizations of ensembles. 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 new frameworks upon which atmospheric dispersion model predictions and field observations could be properly compared, including some of the effects of uncertainties. One such framework was proposed as a standard guide by the American Society for Testing and Materials (ASTM 2000), whose primary author was John Irwin of the U.S. EPA.

The basic assumption of the ASTM procedure is that a realization of the observed concentration, C_o, can be expressed as:

$$C_o = \overline{C_o} + \Delta C_o' + C_o' \tag{36}$$

where $\overline{C_o}$ is the ensemble average that a dispersion model is supposed to predict ideally, $\Delta C_o'$ represents measurement errors due to calibration or unrepresentative instrument siting, and C_o' represents stochastic fluctuations due to turbulence.

The predicted concentration, C_p, can be considered to have the following three components:

$$C_p = \overline{C_p}(\alpha) + \Delta C_p'(\alpha) + C_p'(\alpha) \tag{37}$$

where α represents the set of model input parameters, $\overline{C_p}(\alpha)$ is the ensemble average given by the model, $\Delta C_p'(\alpha)$ represents the effects 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 α) not accounted for by the model.

Direct comparison of observations (which are realizations of ensembles) with model predictions (which are ensemble averages) amounts to comparing C_o with $\overline{C_p}(\alpha)$. The ASTM (2000) procedure suggests that if the effects of $\Delta C_o'$, C_o', $\Delta C_p'(\alpha)$, and $C_p'(\alpha)$ all somehow average to zero, then it is more appropriate to first separately average observations and model predictions over a number of regimes (or ensembles of similar conditions), which can be defined by independent variables such as downwind distance and stability parameter, and then do the comparison. Averaging observed values over each of these regimes may provide an estimate that is closer to what most dispersion models attempt to predict, $\overline{C_o}$. These regime averages of observations and predictions can then be paired to calculate, for example, the performance measures defined in Section 4. Like the BOOT software, the ASTM procedure uses the bootstrap resampling technique to calculate the confidence limits of performance measures, where resampling is done within each regime, and where the observed and predicted values for the same experiment are sampled concurrently.

Strictly speaking, an ensemble would require many experiments conducted under *identical* conditions. This objective, however, is impossible to achieve in a field program. Therefore, regime average represents a surrogate for a true ensemble average, because each regime consists of experiments conducted under *similar* conditions.

5.2. A Sample Implementation of the ASTM Procedure for Short-Range Dispersion Experiments

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 the possibility of plume capture. Moreover, 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 recognizes that because of wind shifts and concentration fluctuations, the cross-wind concentration distribution along an sampling arc is unlikely to be perfectly Gaussian, a lateral distribution assumed by most air dispersion models. These departures from an 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 to be representative of the plume centerline concentration for these short-range dispersion experiments.

ASTM (2000) suggests one way to define near-centerline concentrations. First of all, it is required that the plume must have been well captured by the sampling arc. 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. In practice, this can be done by first calculating the first moment of the cross-wind distribution, which gives the centroid location, y_c,

$$y_c = \frac{\int Cydy}{\int Cdy} \tag{38}$$

where C is the concentration along the arc, and y is the cross-wind coordinate measured in, for example, distance or azimuth angle. The spread, or the second moment, of the cross-wind distribution, σ_y, is then given by

$$\sigma_y = \sqrt{\frac{\int C(y - y_c)^2 dy}{\int Cdy}} \tag{39}$$

To account for uncertainty in the plume centerline position, the ASTM procedure suggests that the concentration at any receptor that is located within 0.67 σ_y from y_c is a representative sample of the plume centerline concentration (Fig. 10). For a Gaussian distribution, the concentration at a lateral distance of 0.67 σ_y from the centerline would equal 80% of the centerline, maximum value.

With the consideration of a finite region ($\pm$ 0.67 σ_y from the centroid), it is likely that an experiment will have multiple near-centerline observed values, even though there is just one predicted centerline value. This is illustrated by an example in Fig. 11, where it is assumed that there are ten experiments that can be grouped into three regimes. Black solid circles indicate the ten centerline predictions for the ten experiments. Red open and red solid circles indicate near-centerline observations that are considered to be representative of the centerline concentrations for the ten experiments. Red solid circles further indicate the maximum among these near-centerline values. Traditional model evaluation compares black solid with red solid circles. The ASTM procedure compares black solid with red solid *and* red open circles.

5.3. Extension of BOOT Software to Include ASTM Procedure

It is apparent that there are many similarities between the ASTM (2000) procedure and the BOOT software described in Section 4. These similarities include

- the calculation of various statistical performance measures,
- the use of the bootstrap resampling to estimate the confidence level,
- the paired sampling between observed and predicted values, and
- the grouping of the data in blocks in BOOT versus in regimes in ASTM.

However, the ASTM procedure proceeds further by

- calculating performance measures based on regime averages (*i.e.*, averaging over all experiments within a regime), rather than based on the values of individual experiments, and
- if the variable to be evaluated is the centerline concentration, considering near-centerline observations to be representative samples of the centerline value.

This section mainly describes the necessary steps to extend the BOOT software in order to incorporate the new ASTM procedure.

Figure 11 can be used to further demonstrate the difference between the BOOT and ASTM methodologies. There are ten experiments that are grouped into three regimes or blocks in Fig. 11. Regimes 1, 2, and 3 have four, three, and three experiments, respectively. There are ten predicted centerline values (black solid circles) for the ten

experiments. However, because of the consideration of near-centerline values for observations, there are a total of 33 observed values (red solid and open circles) for the ten experiments, where red solid circles indicate the maximum observed values for each experiment. The BOOT methodology calculates statistical performance measures, such as the fractional bias (FB) and the normalized mean square error (NMSE), using the ten pairs of predicted and maximum observed values (black solid and red solid circles, respectively). The ASTM procedure first calculates the average of the four predicted values in Regime 1, $\overline{C_{p,R1}}$, and the average of the 13 near-centerline observed values in Regime 1, $\overline{C_{o,R1}}$. Averages for other two regimes are similarly calculated. The ASTM procedure then uses the *three* pairs of regime averages, ($\overline{C_{p,R1}}$, $\overline{C_{o,R1}}$), ($\overline{C_{p,R2}}$, $\overline{C_{o,R2}}$), and ($\overline{C_{p,R3}}$, $\overline{C_{o,R3}}$), to calculate performance measures.

As mentioned before, both BOOT and ASTM involve the bootstrap resampling where resampling is done within each regime or block. The resampling in BOOT is straightforward, because the numbers of observed and predicted values are the same. The ASTM procedure, however, requires special considerations, because the numbers of observed and predicted values can be different (33 vs. 10 in Fig. 11). The table below lists all the data points for Regime 1 in Fig. 11, where there are four experiments (numbered 1 through 4) with four predicted values (right column) and 13 observed values (left column).

Original Version of Data for Regime 1

Observed	Predicted
$C_{o1,1}$, $C_{o1,2}$, $C_{o1,3}$, $C_{o1,4}$	C_{p1}
$C_{o2,1}$, $C_{o2,2}$, $C_{o2,3}$	C_{p2}
$C_{o3,1}$, $C_{o3,2}$	C_{p3}
$C_{o4,1}$, $C_{o4,2}$, $C_{o4,3}$, $C_{o4,4}$	C_{p4}

Since all of the near-centerline measurements for each experiment are adjacent, one approach would be to sample *a pair of neighboring observed values* each time in order to preserve the correlation of these observed values. This approach is found to be more robust than sampling one observed value each time (Irwin and Rosu 1998). As a result, assuming there are N observed values in a regime, the ASTM procedure suggests that *INT*(N/2) pairs of adjacent samples be drawn, where *INT*(N/2) is the truncated integer result of dividing N by 2. For the example above, since N = 13, *INT*(13/2) = 6. Hence, for each bootstrap sample of Regime 1, six experiments will be randomly selected from the four experiments in that regime, where for each experiment a pair of adjacent observed values will be drawn. Since there is only a single predicted value for each experiment, that value will be drawn twice in order to maintain concurrent sampling of observed and predicted values. Assume the six randomly selected experiments are 3, 3, 2, 4, 1, 2, then a possible sample of Regime 1 following the above procedures would be

One Possible Resampled Version of Data for Regime 1

Observed	Predicted
$C_{o3,1}$, $C_{o3,2}$	$C_{p3,}$ C_{p3}
$C_{o3,1}$, $C_{o3,2}$	$C_{p3,}$ C_{p3}
$C_{o2,2}$, $C_{o2,3}$	C_{p2}, C_{p2}
$C_{o4,3}$, $C_{o4,4}$	C_{p4}, C_{p4}
$C_{o1,1}$, $C_{o1,2}$	C_{p1}, C_{p1}
$C_{o2,1}$, $C_{o2,2}$	C_{p2}, C_{p2}

It can be seen that the sampled observed values are always adjacent, that each predicted value is always sampled twice, and that the resampling is done with replacement. There are some noticeable differences between this resampled version and the original version of the data:

- The original version of the data has four experiments in Regime 1, but the resampled version has six (= *INT*(13/2)) experiments instead.
- The original version has four predicted values and 13 observed values in Regime 1. The average of the four predicted values gives the nominal (median) value of $\overline{C_{p,R1}}$, and the average of the 13 predicted values gives the nominal (median) value of $\overline{C_{o,R1}}$. The resampled version has 12 predicted and 12 observed values. These values are then averaged to give a new estimate of regime averages $\overline{C_{p,R1}}$ and $\overline{C_{o,R1}}$.

As previously mentioned, one way for the BOOT procedure to estimate confidence limits is to use Eq. (35), where $t_{95\%}$ is the Student's t value at the 95% confidence level with N – 1 degrees of freedom, and N is the number of observation-prediction pairs (or experiments). In implementing the ASTM procedure, the degrees of freedom should be N – NR – 1, where NR is the number of regimes. The reason why the degrees of freedom are reduced by NR is because NR regime averages will first have to be calculated. No special treatment is necessary for the percentile confidence limits.

Based on the above discussions, the following are additional optional procedures that have been implemented in the new BOOT software in order to incorporate the ASTM methodology:

- Allow multiple observed values for each experiment.
- Calculate the average for each regime. Then, calculate all performance measures based on these averages, rather than based on original individual values.
- Sample two adjacent observed values each time during resampling.

- Reduce the degrees of freedom by the number of regimes when calculating the Student's t confidence limits.

Therefore, it can be concluded that the main extensions associated with the ASTM procedure are the treatment of regime averages, and the generation of bootstrap resamples for regime averages. The procedures by which statistical performance measures are calculated essentially remain the same.

6. User's Instructions for the BOOT Software

Section 4 described the quantitative performance measures (FB, MG, NMSE, VG, R, FAC2, FB_{FN}, FB_{FP}, MG_{FN}, MG_{FP}, MOE_{FN}, and MOE_{FP}) that the new BOOT software calculates. Section 5 describes the additional procedures that have been added in BOOT in order to implement the ASTM procedure. This section provides the user's instructions for the BOOT software.

6.1. Run-Time Environment

The BOOT software has been primarily applied in a Microsoft Windows environment. However, as described later, the program can be easily ported to other computer platforms because of the standard programming language used. The only installation requirement is to place the BOOT program and all the associated input and output files in the same directory folder.

In order to run BOOT, the user will need to respond to a number of prompts at the command line (see Section 6.2), and prepare a mandatory input file (see Section 6.3). The program in turn generates one mandatory output file and two optional output files (see Section 6.3). It is assumed that the BOOT program and all the associated data files reside in the same folder.

Follow these steps to launch BOOT:

- Click on `Start / Run…`
- Hit the `Browse` button to navigate to the folder where BOOT and the associated data files reside.
- Hit the `OK` button to launch the program.

Alternatively, BOOT can also be launched by these steps:

- Open a `Command Prompt` window.
- Change to the folder where BOOT and the associated data files reside using the `CD` command.
- Type `BOOT` to launch the program.

Because of its simple I/O requirements, the BOOT program can be easily run in batch mode by redirecting command-line inputs to an external file.

6.2. Command-Line Prompts

Once BOOT is launched, the user will be prompted with the following questions (each enclosed by a rectangle) at the command line. Some questions are conditional.

```
Name of input file:
```

The name of the mandatory input data file. There is no default for this prompt. See Section 6.3 for file format.

```
Name of output file:
```

The name of the mandatory output data file. There is no default for this prompt. See Section 6.3 for file format.

```
Select one from the following options:
(1) straight Co and Cp comparison
(4) consider ln(Co) and ln(Cp)
```

Select the option (`IMENU`) of comparing C_o and C_p directly, or comparing the logarithms of C_o and C_p. `IMENU` must equal 1 or 4, and there is no default value assumed. When `IMENU` = 1, NMSE and all measures related to FB (*i.e.*, FB_{FN}, FB_{FP}, MOE_{FN}, and MOE_{FP}) will be calculated, but not VG and all measures related to MG (*i.e.*, MG_{FN} and MG_{FP}). When `IMENU` = 4, VG and all measures related to MG will be calculated, but not NMSE and all measures related to FB. R is based on C_o and C_p when `IMENU = 1`, and is based on $\ln(C_o)$ and $\ln(C_p)$ when `IMENU = 4`. FAC2 is always based on C_o and C_p regardless of the value of `IMENU`. Note that C_o and C_p must be positive when `IMENU` = 4. (C_o and C_p are always non-negative by definition for physical plume quantities such as concentration, dosage, and cloud width.)

```
Use ASTM procedure? (y/<N>)
```

Decide whether (option `IASTM`) to consider the ASTM procedure (see Section 5). Enter `Y` or `N` (case-insensitive), or simply hit the `Enter` key. Hitting the `Enter` key amounts to selecting the default option, `N`, *i.e.*, not using the ASTM procedure.

```
Print out original data? (y/<N>)
```

Decide whether to echo the original data in the mandatory output file. Enter Y or N (case-insensitive), or simply hit the Enter key. Hitting the Enter key amounts to selecting the default option, N, *i.e.*, not printing the original data.

```
Use E- or F-format for mean, sigma, and bias? (<F>/e)
```

Decide whether to use the F (floating point without exponentiation) or E (explicit exponential notation) format for the mean, standard deviation, and bias (difference in the means). Enter F or E (case-insensitive), or simply hit the Enter key. Hitting the Enter key amounts to selecting the default option, F, *i.e.*, using the floating point format without exponentiation.

```
Calculate partial correlation? (y/<N>)
That is, the influence from a certain model is removed.
```

Decide whether (option IPART) to calculate the partial correlation coefficient (Panofsky and Brier 1958; see Eq. (28)) between observations and a model's predictions after the influence of another model's predictions is removed. Enter Y or N (case-insensitive), or simply hit the Enter key. Hitting the Enter key amounts to selecting the default option, N, *i.e.*, not considering partial correlation.

```
The influence of which model, Cp(i), do you want to remove?
Note that i corresponds to the i+1 th model in the input file.
Enter i now:
```

The prompt will appear only if IPART equals Y.

Decide which model whose influence is to be removed. Since the mandatory input file (Section 6.3) consider the "first model" as observations, the predictions for model i actually correspond to the $i+1^{th}$ "model" in the file.

```
Do the bootstrap resampling? (<Y>/n)
```

Decide whether (option IBOOT) to do the bootstrap resampling. Enter Y or N (case-insensitive), or simply hit the Enter key. Hitting the Enter key amounts to selecting the default option, Y, *i.e.*, conducting the bootstrap resampling.

```
Print out detailed information on confidence limits? (y/<N>)
```

The prompt will appear only if IBOOT equals Y.

Decide whether to print out detailed quantitative information on confidence limits resulting from the bootstrap resampling. Enter Y or N (case-insensitive), or simply hit the

Enter key. Hitting the Enter key amounts to selecting the default option, N, *i.e.*, not printing out the detailed quantitative information on confidence limits.

```
Create files containing FB (with its 95% confidence limits) and
NMSE that can later be plotted? (<Y>/n)
```

The prompt will appear only if IBOOT equals Y.

Decide whether (option INN) to create two optional files containing the information on the confidence limits for FB and NMSE if IMENU = 1, or for MG and VG if IMENU = 4, that can later be plotted (see Fig. 12 for an example). Enter Y or N (case-insensitive), or simply hit the Enter key. Hitting the Enter key amounts to selecting the default option, Y, *i.e.*, creating optional files. See Section 6.3 for file format.

The information contained in these two optional files in fact will also be included in the mandatory output file if the user chooses to print out detailed quantitative information on confidence limits, but is separately reproduced here to facilitate plotting.

```
Additional two files will be opened:
If IMENU = 1, then
     Enter name of file 1 that contains FB and NMSE info.
     Enter name of file 2 that contains d(FB) and d(NMSE) info.
If IMENU = 4, then
     Enter name of file 1 that contains MG and VG info.
     Enter name of file 2 that contains d(MG) and d(VG) info.
```

The prompt will appear only if both IBOOT and INN equal Y.

If IMENU = 1, the first file contains the information on FB, together with confidence limits, and NMSE for each model. The second file contains the information on the difference in FB, together with confidence limits, and the difference in NMSE for all model pairs. There are no default file names assumed.

If IMENU = 4, the first file contains the information on MG, together with confidence limits, and VG for each model. The second file contains the information on the difference in MG, together with confidence limits, and the difference in VG for all model pairs. There are no default file names assumed.

```
Make another run? (y/<N>)
```

Decide whether to make another run. Enter Y or N (case-insensitive), or simply hit the Enter key. Hitting the Enter key amounts to selecting the default option, N, *i.e.*, not making another run. If Y is entered, the whole command-prompt sequence will repeat without exiting the program.

6.3. Input and Output File Formats

All input and output files of BOOT are in ASCII (*i.e.*, plain text). Formats of these files are described below.

The mandatory input data file consists of four header records, followed by numerical data. All records are free format with space delimiters. Table 2 shows the input file structure in detail, and a sample mandatory input data file is shown in Table 3. The data in Table 3 are also consistent with the sample database in Table 1.

Table 4 shows a sample BOOT mandatory output file, which is based on use of the sample input file in Table 3. The mandatory output file consists of a number of segments, with some segments appearing only if certain options have been selected at the command line (see Section 6.2). Table 4's caption provides a detailed description of each segment. Note that no information on MG and VG is listed in Table 4 because `IMENU` = `1`. Set `IMENU` = `4` to obtain the information on MG and VG.

Table 5 shows a sample BOOT optional output file that corresponds to the sample input file in Table 3. The file displays *for each model* the FB (together with the 95% confidence limits) and NMSE if `IMENU` = 1, or the MG (together with the 95% confidence limits) and VG if `IMENU` = 4. The information displayed in this file can also be found in the mandatory output file with appropriate command-line options. The percentile confidence limits are used in the table.

Table 6 shows an additional sample BOOT optional output file that corresponds to the sample input file in Table 3. The file displays *for each model pair* the difference in FB (together with the 95% confidence limits) and the difference in NMSE if `IMENU` = 1, or the difference in MG (together with the 95% confidence limits) and the difference in VG if `IMENU` = 4. The information displayed in this file can also be found in the mandatory output file with appropriate command-line options. The percentile confidence limits are used in the table.

6.4. Programming Notes

The BOOT software was developed using the FORTRAN 95 programming language. Since the program allocates memory dynamically, there is no need to recompile the code for a problem with more data points, more models, or more data blocks (regimes). Moreover, because of the standard programming language used, BOOT can be easily ported to other computer platforms such as UNIX and LINUX.

BOOT is currently set to take 1,000 bootstrap resamples (`MAXSS = 1000`). Recompilation is necessary if a different number of bootstrap resamples is desired.

The current version of BOOT uses the `RAN3` random number generator routine described in Press *et al.* (1992) to generate a sequence of random numbers for the purpose resampling. (The previous version of BOOT originally programmed by Hanna (1989) used a fixed set of random numbers from an external data file.)

7. A Demonstration

This section contains a further demonstration of the use of the BOOT software using the sample database listed in Table 3 (or Table 1). The database includes 79 hours of observed pollutant concentrations and predicted concentrations from three dispersion models, Model-A, Model-B, and Model-C. These are the same data used in earlier figures and tables (Figs. 1 through 5 and Tables 1 through 6). Some corresponding independent variables used by the models to carry out calculations are also listed in Table 1, including time of day, ambient wind speed, mixing height, and atmospheric stability class. The first 40 records of the database are from one field program, and the remaining 39 records are from a second field program. Performance measures for the three models are summarized in Table 7, where the information is retrieved from the sample BOOT mandatory output file in Table 4.

It can be seen that all three models correctly predict the highest and second highest observed values within about 10%. The values of FB suggest (see Eq. (31)) that the mean bias on a linear scale is ~0, 6% underprediction, and 40% overprediction for Model-A, B, and C, respectively. The relatively small bias for Model-A and Model-B is also evident from that fact that FB_{FN} nearly cancels out FB_{FP} for the two models. For Model-A, even though the mean bias is almost zero, the underpredicting and overpredicting errors are both about 20% of the mean.

The values of MG suggest (see Eq. (33)) that the mean bias on a logarithmic scale is about 20% underprediction, 25% underprediction, and 50% overprediction for Model-A, B, and C, respectively. The values of NMSE suggest (see Eq. (32)) that the random scatter on a linear scale for the three models is about a factor of 1.5 to 2 of the mean. The values of VG suggest (see Eq. (34)) that the random scatter on a logarithmic scale is about a factor of 2.5 to 3.5 of the mean. The R for Model-C is almost zero, also evident from the scatter plot in Fig. 1. Model-A has about 80% of predictions within a factor of 2 of observations, and Model-B and C have about 60% of predictions within a factor of 2 of observations.

Based on the above, it appears that Model-A has the best performance, and Model-C is the worst performer.

Table 8 summarizes the results of significance tests based on the bootstrap resampling, where a mark "×" indicates significantly different from zero at the 95% confidence level, and Δ means the difference between two models. The information in Table 8 was presented in Tables 5 and 6, which are listings of two optional BOOT output

files. Significance tests were not conducted for each model's NMSE and ln(VG), because they are always greater than zero by definition. It can be seen that although Model-A has the lowest FB (= 0.001) and an MG (= 1.22) that is closest to 1.0 in Table 7, these values are not significantly different from the FB (= 0.057) and MG (= 1.34) for Model-B. This shows the importance of conducting significance tests, or incorrect conclusions might be reached.

Figure 12 shows the MG and VG for the three models. The 95% confidence limits for MG based on the bootstrap resampling (the percentile confidence limits) are also shown as horizontal bars. The figure is a useful way of summarizing the MG and VG statistics in a single diagram, where the parabola indicates the minimum VG defined by Eq. (30). In other words, no points should be located below the parabola. Dashed lines represent a factor-of-two mean bias. A perfect model would have VG = MG = 1.0, and is located at the bottom center of the diagram. It can be seen that Model-A has an MG that is closest to 1.0. Model-C has the smallest VG (= 2.28). Perhaps in the future Fig. 12 should be improved to include the 95% confidence limits for ln(VG) as well, so that the performance information will not be misinterpreted.

8. Summary

A general methodology for evaluating atmospheric dispersion model performance has been 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. In addition to statistical performance evaluation, a model can also be evaluated scientifically and operationally. Exploratory data analysis involves the use of various types of plots, including scatter plots (Fig. 1), quantile-quantile plots (Fig. 2), box-residual plots (Figs. 3 and 4), and scatter-residual plots (Fig. 5). Here 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 any trends of model residuals with independent variables.

The methodology for statistical performance evaluation has been implemented in the BOOT software package. Hanna (1989) and Hanna *et al.* (1991) describe the first version of BOOT. This report describes Version 2.0 of the software. The original BOOT 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) (Eqs. (1) through (6), respectively). 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 usually required to consider a number of performance measures. For dispersion modeling where concentrations can easily vary by several orders of magnitude, MG and VG are probably preferred over FB and NMSE. However, MG and VG can 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 detection (LOD), be used as a lower threshold in calculating MG and VG. R is generally not a very robust measure because it is sensitive to a few aberrant data pairs. Furthermore, measurements are commonly available in concentric arcs for short-range dispersion field experiments. As a result, 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. FAC2 is probably the most robust performance measure, because it is not overly influenced by either low or high outliers.

It is also shown how FB, NMSE, MG, and VG can 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. See Fig. 9 and Eqs. (31) through (34) for examples of this interpretation.

The bootstrap resampling can be used to estimate the confidence limits of 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. Figure 12 shows a way of presenting both MG and VG in a single diagram, where MG's confidence limits and the minimum VG for a given MG (Eq. (30)) are also plotted. Similar diagram for FB and NMSE can also be constructed.

To address the fact that FB and MG measure only systematic bias, the new BOOT software further separates FB and MG into underpredicting (false-negative) and overpredicting (false-positive) components. In addition, the new BOOT software also calculates the two-dimensional measure of effectiveness (MOE) recommended by Warner *et al.* (2001). It is shown that MOE is closely related to FB, and in fact one can be expressed as a function of the other.

Many researchers have pointed out the inadequacy of a deterministic model 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 performance measures based on these regime averages. The new BOOT software has been extended to include the ASTM procedure as an option.

Finally, this report provides detailed user's instructions for BOOT, together with a demonstration of the use of the software package.

References

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.

Chang, J.C., 2002: *Methodologies for Evaluating Performance and Assessing Uncertainty of Atmospheric Dispersion Models.* Ph.D. thesis, George Mason University, Fairfax, VA, 277 pp. Available from http://wwwlib.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, J.C., K. Chayantrakom, and S.R. Hanna, 2001: Evaluation of CALPUFF, HPAC, and VLSTRACK with the Over-Land Alongwind Dispersion (OLAD) Field Data. George Mason University, Fairfax, VA. Available from http://camp.gmu.edu.

Chang, J.C., M.E. Fernau, J.S. Scire, and D.G., Strimaitis, 1998: A Critical Review of Four Types of Air Quality Models Pertinent to MMS Regulatory and Environmental Assessment Missions. Prepared for U.S. Department of the Interior, Minerals Management Service, Gulf of Mexico OCS Region, 1201 Elmwood Park Blvd., New Orleans, LA 70123, by Earth Tech, Inc., 196 Baker Avenue, Concord, MA 01742.

Chang, J.C., and S.R. Hanna, 2004: Air quality model performance evaluation. *Meteorol. and Atmos. Phys*., **87**, 167-196.

Conover, W.J., 1980: *Practical Nonparametric Statistics, Second Edition.* John Wiley & Sons, 493 pp.

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

Ebert, E.E., 2001: Ability of a pool man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461-2480.

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

Efron, B., 1987: Better bootstrap confidence intervals. *J. Am. Stat. Asso.*, **82**, 171-185.

EPA, 1997: *Guiding Principles for Monte Carlo Analysis*. EPA/630/R-97/001, Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, D.C. 20460.

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

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

Hanna, S.R., and J.C. Chang, 2001: Examples of evaluations of new models with field data. Conference on Guideline on Air Quality Modeling, A New Beginning, Newport, Rhode Island; Air and Waste Management Association, Pittsburgh, Pennsylvania.

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

Hanna, S.R., D.G. Strimaitis, and J.C. Chang, 1991: *Hazard Response Modeling Uncertainty (A Quantitative Method), Volume I: User's Guide for Software for Evaluating Hazardous Gas Dispersion Models; Volume II: Evaluation of Commonly-Used Hazardous Gas Dispersion Models; Volume 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.

Ichikawa, Y., and K. Sada, 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. Asso.*, **52**, 313-323.

Irwin, J.S., M.-R., Rosu, 1998: Comments on a draft practice for statistical evaluation of atmospheric dispersion models. *Proc. 10th Joint Conference on the Application of Air Pollution Meteorology with A&WMA*, 11-16 January 1998, Phoenix, Arizona. American Meteorological Society, Boston, MA.

Isaaks, E.H., and R.M. Srivastava, 1989: *An Introduction to Applied Geostatistics*. Oxford University Press, 561 pp.

Lumley, J.L., and H.A. Panofsky, 1964: *The Structure of Atmospheric Turbulence*. Wiley Interscience, New York, 239 pp.

McNally D., and T.W. Tesche, 1993: *MAPS Sample Products*. Alpine Geophysics, 16225 W. 74th Dr., Golden, CO 80403.

Mesinger, F., 1996: Improvements in quantitative precipitation forecast with the Eta regional model and the National Centers for Environmental Prediction: the 48-km upgrade. *Bull. Amer. Meteor. Soc.*, **77**, 2637-2649. Corrigendum: **78**, 506.

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

Nappo, C.J., and K.S.M. Essa, 2001: Modeling dispersion from near-surface tracer releases at Cape Canaveral, Florida. *Atmos. Environ.*, **35**, 3999-4010.

Nappo, C.J., R.M. Eckman, K.S. Rao, J.A. Herwehe, and R.L. Gunter, 1998: *Second Order Closure Integrated Puff (SCIPUFF) Model Verification and Evaluation Study*. NOAA Technical Memorandum ERL ARL-227, Air Resources Laboratory, National Oceanic and Atmospheric Administration, Silver Spring, MD.

Olesen, H.R., 2001: Ten years of harmonization activities: past, present, and future. *7th International conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes*, Belgirate, Italy. National Environmental Research Institute, Roskilde, Denmark (http://www.dmu.dk/AtmosphericEnvironment/HARMONI.htm).

Panofsky, H.A., and G.W. Brier, 1958: *Some Applications of Statistics to Meteorology*. Pennsylvania State University Press, 224 pp.

Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, 1992: *Numerical Recipes in Fortran, the Art of Scientific Computing, Second Edition*. Cambridge University Press, 993 pp.

Venkatram, A., 1988: Topics in applied dispersion modeling, Chapter 6 of *Lectures on Air Pollution Modeling*. American Meteorological Society, Boston, MA, 390 pp.

Venkatram, A., 1984: The uncertainty in estimating dispersion in the convective boundary layer. *Atmos. Environ.*, **18**, 307-310.

Warner, S., N. Platt, J.F. Heagy, S. Bradley, G. Bieberbach, G. Sugiyama, J.S. Nasstrom, K.T. Foster, and D. Larson, 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, J.C., R.I. Sykes, and A. Venkatram, 1992: Evaluating air-quality models: review and outlook. *J. Appl. Meteor.*, **31**, 1121-1145.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, New York, 467 pp.

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

Tables and Figures

Table 1. A sample model evaluation database that contains observed concentrations (ppm); predictions (ppm) by Model-A, Model-B, and Model-C; and the corresponding values for time of day (hour, local standard time or LST), wind speed (m s^{-1}), mixing height (m), and Pasquill-Gifford stability class (1 = very unstable, 2 = unstable, 3 = moderately unstable, 4 = neutral, 5 = moderately stable, 6 = very stable). (It is also common to use A through F, instead of 1 through 6, to indicate stability class.) There are 79 records in the database. The first 40 records correspond to one field program, and the remaining 39 records (shaded) correspond to another field program.

Observed (ppm)	Model-A (ppm)	Model-B (ppm)	Model-C (ppm)	Hour (LST)	Wind speed (m s^{-1})	Mixing height (m)	PG stability class
616.0	708.7	594.7	516.5	11	3.0	800	2
604.1	689.2	585.8	496.7	12	3.4	1000	2
868.0	674.8	580.3	516.8	13	3.5	1100	2
498.6	668.8	652.1	548.3	14	3.8	1200	2
393.1	560.2	704.7	581.9	15	4.7	1300	2
409.0	740.9	570.1	621.4	16	5.2	1000	3
640.2	249.6	510.1	553.5	17	5.4	1100	3
265.3	259.6	463.4	446.0	18	4.9	1100	4
192.7	91.6	131.0	485.0	19	4.2	1100	5
1149.1	1217.5	1116.1	520.6	10	2.6	1600	2
972.8	1275.8	1175.1	536.9	11	3.2	1900	2
1137.5	1225.7	1081.7	617.4	12	3.8	1600	2
669.5	1052.8	905.1	637.3	13	4.5	1600	2
595.5	862.0	862.0	664.1	14	5.0	1500	2
741.2	589.5	767.0	665.3	15	5.1	1500	2
612.6	602.4	728.2	672.4	16	5.0	1500	3
312.0	398.9	657.5	659.5	17	5.2	1500	3
400.2	340.2	412.3	586.0	18	5.1	1500	4
264.7	612.1	774.2	705.9	16	5.7	1400	3
290.0	428.4	757.3	708.8	17	5.1	1800	3
459.5	355.0	512.3	602.4	18	5.1	2000	4
444.0	216.0	441.4	681.1	19	4.4	2000	5
175.1	216.6	456.1	825.4	20	4.6	2000	6
102.3	126.1	255.6	522.9	21	4.9	2000	6
128.8	16.5	0.5	834.9	22	4.6	0	6
200.2	301.9	208.9	728.0	23	5.4	0	6
358.3	481.8	354.0	742.4	24	5.4	0	6
611.1	1010.2	987.1	679.0	14	4.4	1500	2
499.3	752.5	921.6	725.7	15	5.0	1500	2
537.8	724.0	826.8	675.9	16	4.7	1500	3
220.0	523.3	908.2	640.8	17	3.9	1800	3
479.2	357.5	788.6	544.7	18	4.2	2000	4
133.2	195.3	383.1	738.5	19	3.1	1800	5
98.2	167.3	213.5	1064.9	20	3.2	1500	6

92.5	104.6	142.2	741.2	21	3.1	1200	6
21.0	127.4	176.3	805.2	22	3.3	1200	6
353.0	307.8	167.1	576.9	20	3.8	2000	5
358.0	280.9	188.4	225.3	21	2.3	2000	4
233.3	355.3	234.9	719.1	22	2.4	2000	5
198.3	12.7	184.0	745.2	23	3.6	2000	6
507.2	3.0	126.3	664.9	24	3.5	2000	6
313.7	0.2	30.0	667.1	1	4.2	0	6
165.1	16.5	4.0	703.9	2	3.6	0	6
295.6	329.9	454.6	695.3	4	5.2	0	6
527.7	308.0	295.9	775.0	5	4.7	0	6
454.1	301.0	1.0	995.6	6	2.9	0	6
240.3	417.5	361.1	933.8	7	3.4	0	6
590.8	579.3	144.2	666.5	8	3.1	1500	5
638.3	756.6	608.9	400.1	9	3.4	1500	4
949.8	1004.2	805.4	528.9	10	3.4	1500	3
886.8	855.6	706.2	517.4	11	3.0	1300	2
635.5	761.0	670.9	596.6	12	4.5	1200	2
359.3	412.6	232.5	937.6	1	2.3	1200	6
484.7	360.7	226.8	979.0	2	2.5	1200	6
529.7	332.0	202.5	980.0	3	2.4	1200	6
585.8	291.4	186.1	1100.1	4	2.1	1200	6
367.7	368.0	260.2	1005.6	5	2.1	1200	6
324.7	270.9	72.7	1058.6	6	2.0	1200	6
489.0	274.6	208.5	942.2	7	2.6	1200	6
570.8	337.1	218.0	646.5	8	2.8	1200	5
419.7	254.4	206.1	344.0	9	4.3	1200	4
532.8	414.2	197.9	477.0	9	4.8	1800	3
425.2	365.7	198.7	469.5	10	7.1	1700	4
467.5	411.5	228.5	455.3	11	7.5	2000	4
362.2	306.4	147.6	405.2	12	5.1	2000	4
429.2	287.4	139.2	450.6	13	5.4	2000	4
446.0	338.1	169.5	461.2	14	5.7	2000	4
192.9	253.8	145.6	460.7	15	5.8	2400	4
630.3	322.5	257.2	460.5	16	7.3	2700	4
364.9	326.7	251.1	510.6	17	7.8	3000	4
111.4	196.4	248.5	314.4	23	1.9	250	4
89.8	146.5	254.9	123.2	24	1.9	250	4
82.5	248.0	160.9	80.9	1	2.9	250	4
296.5	253.2	193.2	230.4	2	2.8	250	4
215.4	299.7	165.0	339.5	3	2.9	250	4
454.5	274.2	154.0	120.4	4	3.5	250	4
384.7	324.6	163.2	251.7	5	3.5	250	4
253.2	488.3	175.6	122.4	6	3.5	250	4
289.5	304.1	193.1	153.8	7	3.1	250	4

Table 2. Format of the mandatory input data file for BOOT. The file contains four header records, followed by nn (see below) numerical records. All records are free format with space delimiters. See Table 3 for a sample input file.

Record No.	Descriptions
1	Contains three variables: `nn, mm, kk` All three variables are integers, where `nn` is the number of experiments in total, `mm` is the number of "models" where the first "model" is assumed to be observations, and `kk` is the number of blocks (or regimes). Since observed values are counted as one "model," the actual number of models is in fact `mm – 1`.
2	Contains `kk` variables: `nkk(i), i = 1, kk` All variables are integers, where `nkk` is the umber of experiments in each block (regime). The sum of `nkk` over all blocks must equal `nn`.
3	Contains `mm` variables: `modnam(i), i = 1, mm` All variables are characters (maximum length 8 bytes) enclosed by single quotation marks, where `modnam` is the name of each model. Note that quotation marks are not part of `modnam`, and thus not counted as part of the 8-byte limit.
4	Contains `kk` variables: `blknam(i), i = 1, kk` All variables are characters (maximum length 30 bytes) enclosed by single quotation marks, where `blknam` is the name of each block (regime). Note that quotation marks are not part of `blknam`, and thus not counted as part of the 30-byte limit.
Next `nkk(1)` records	Each record contains `nnobs + mm` variables: `nnobs, (co(i), i = 1, nnobs), (cp(i), i = 1, mm - 1)`, corresponding to one of the nkk(1) experiments in block 1. The first variable is integer and the rest of variables are real numbers, where `nnobs` is the number of observed values (`co`) for the experiment, and `cp` is the predictions for the mm – 1 models. Note that `nnobs = 1` if the ASTM procedure is not used, and that `nnobs` ≥ `1` if the ASTM procedure is used. Since `mm` includes observations as one "model," there

	are in fact mm – 1 models. It can be seen that in order to implement the ASTM algorithm, the current input file structure allows multiple observed values for each experiment, but only one predicted value for each model.
Next `nkk(2)` records	Each record contains `nnobs + mm` variables: `nnobs, (co(i), i = 1, nnobs), (cp(i), i = 1, mm - 1),` corresponding to one of the `nkk(2)` experiments in block 2.
…	…
Next `nkk(kk)` records	Each record contains `nnobs + mm` variables: `nnobs, (co(i), i = 1, nnobs), (cp(i), i = 1, mm - 1),` corresponding to one of the `nkk(kk)` experiments in block `kk`.

Table 3. Sample mandatory input file for BOOT. See Table 2 for file format. This sample input file is *not* for the ASTM procedure because there is always a single observed value for each experiment, as indicated by the "1" leading all numerical records (*i.e.*, nnobs in Table 2 always equals 1). The numerical data (*i.e.*, observed and predicted concentrations) are the same as those shown in Table 1.

```
   79    4  2
   39 40
'OBS.' 'MODEL-A' 'MODEL-B' 'MODEL-C'
'Urban data set' 'Rural data set'
1  616.0  708.7  594.7  516.5
1  604.1  689.2  585.8  496.7
1  868.0  674.8  580.3  516.8
1  498.6  668.8  652.1  548.3
1  393.1  560.2  704.7  581.9
1  409.0  740.9  570.1  621.4
1  640.2  249.6  510.1  553.5
1  265.3  259.6  463.4  446.0
1  192.7   91.6  131.0  485.0
1 1149.1 1217.5 1116.1  520.6
1  972.8 1275.8 1175.1  536.9
1 1137.5 1225.7 1081.7  617.4
1  669.5 1052.8  905.1  637.3
1  595.5  862.0  862.0  664.1
1  741.2  589.5  767.0  665.3
1  612.6  602.4  728.2  672.4
1  312.0  398.9  657.5  659.5
1  400.2  340.2  412.3  586.0
1  264.7  612.1  774.2  705.9
1  290.0  428.4  757.3  708.8
1  459.5  355.0  512.3  602.4
1  444.0  216.0  441.4  681.1
1  175.1  216.6  456.1  825.4
1  102.3  126.1  255.6  522.9
1  128.8   16.5    0.5  834.9
1  200.2  301.9  208.9  728.0
1  358.3  481.8  354.0  742.4
1  611.1 1010.2  987.1  679.0
1  499.3  752.5  921.6  725.7
1  537.8  724.0  826.8  675.9
1  220.0  523.3  908.2  640.8
1  479.2  357.5  788.6  544.7
1  133.2  195.3  383.1  738.5
1   98.2  167.3  213.5 1064.9
1   92.5  104.6  142.2  741.2
1   21.0  127.4  176.3  805.2
1  353.0  307.8  167.1  576.9
1  358.0  280.9  188.4  225.3
1  233.3  355.3  234.9  719.1
1  198.3   12.7  184.0  745.2
1  507.2    3.0  126.3  664.9
1  313.7    0.2   30.0  667.1
1  165.1   16.5    4.0  703.9
1  295.6  329.9  454.6  695.3
1  527.7  308.0  295.9  775.0
1  454.1  301.0    1.0  995.6
1  240.3  417.5  361.1  933.8
1  590.8  579.3  144.2  666.5
1  638.3  756.6  608.9  400.1
1  949.8 1004.2  805.4  528.9
1  886.8  855.6  706.2  517.4
1  635.5  761.0  670.9  596.6
1  359.3  412.6  232.5  937.6
1  484.7  360.7  226.8  979.0
1  529.7  332.0  202.5  980.0
1  585.8  291.4  186.1 1100.1
1  367.7  368.0  260.2 1005.6
1  324.7  270.9   72.7 1058.6
1  489.0  274.6  208.5  942.2
1  570.8  337.1  218.0  646.5
1  419.7  254.4  206.1  344.0
1  532.8  414.2  197.9  477.0
1  425.2  365.7  198.7  469.5
1  467.5  411.5  228.5  455.3
1  362.2  306.4  147.6  405.2
1  429.2  287.4  139.2  450.6
1  446.0  338.1  169.5  461.2
1  192.9  253.8  145.6  460.7
1  630.3  322.5  257.2  460.5
1  364.9  326.7  251.1  510.6
1  111.4  196.4  248.5  314.4
1   89.8  146.5  254.9  123.2
1   82.5  248.0  160.9   80.9
1  296.5  253.2  193.2  230.4
1  215.4  299.7  165.0  339.5
1  454.5  274.2  154.0  120.4
1  384.7  324.6  163.2  251.7
1  253.2  488.3  175.6  122.4
1  289.5  304.1  193.1  153.8
```

Table 4. Sample BOOT mandatory output file based on the sample input file shown in Table 3. The output is divided into the following segments, delineated by artificial blue lines for clarity. Segment 1: overall summary of the input data and the processing option chosen by the user. Segment 2 (optional): Echo of the original dataset, where the block (regime) number has been added to the first column. Segment 3: Median (nominal) values of performance measures for the complete dataset as a whole, and for each block. In addition to the performance measures defined in Section 4, the highest and second highest values of the dataset are also listed. Segment 4 (optional): Detailed (quantitative) information on confidence limits. Note that the Student's t confidence limits are given by Eq. (35), whereas the percentile confidence limits are given by the 2.5^{th} and 97.5^{th} percentiles of the cumulative distribution function for the resamples. Segment 5: Summary of the results of significance tests; *e.g.*, whether the difference in NMSE between Model-A and Model-C is significantly different from zero, and whether the FB for Model-B is significantly different from zero at the 95% confidence level. The results are based on the percentile confidence limits. There is no information on MG and VG listed because the option of using straight C_o and C_p for evaluation has been selected for this run (see Segment 1). Choose the option of using $\ln(C_o)$ and $\ln(C_p)$ for evaluation to obtain the information on MG and VG.

Segment 1

```
OUTPUT OF THE BOOT PROGRAM, LEVEL 10/15/2004

No. of experiments                    =   79
No. of models                         =    4
(with the observed data counted as one)
No. of observations                   =   79
(there might be multiple observations in each experiment, if the ASTM option is chosen)
(there is only one prediction in each experiment)
No. of observations available for
paried sampling                       =   78
(there might be odd number of observations in each block)
No. of blocks (regimes)               =    2
No. of experiments in each block (regime)
  39  40


Out of the following options:
(1) straight Co and Cp comparison
(4) consider ln(Co) and ln(Cp)
 1 was selected
```

Segment 2

```
Input data:  Co, Cp1, Cp2 ...

 1   1  616.0      708.7      594.7      516.5
 1   1  604.1      689.2      585.8      496.7
 1   1  868.0      674.8      580.3      516.8
 1   1  498.6      668.8      652.1      548.3
 1   1  393.1      560.2      704.7      581.9
 1   1  409.0      740.9      570.1      621.4
 1   1  640.2      249.6      510.1      553.5
 1   1  265.3      259.6      463.4      446.0
 1   1  192.7      91.60      131.0      485.0
 1   1  1149.      1218.      1116.      520.6
 1   1  972.8      1276.      1175.      536.9
 1   1  1138.      1226.      1082.      617.4
 1   1  669.5      1053.      905.1      637.3
 1   1  595.5      862.0      862.0      664.1
 1   1  741.2      589.5      767.0      665.3
 1   1  612.6      602.4      728.2      672.4
 1   1  312.0      398.9      657.5      659.5
 1   1  400.2      340.2      412.3      586.0
 1   1  264.7      612.1      774.2      705.9
 1   1  290.0      428.4      757.3      708.8
 1   1  459.5      355.0      512.3      602.4
 1   1  444.0      216.0      441.4      681.1
 1   1  175.1      216.6      456.1      825.4
 1   1  102.3      126.1      255.6      522.9
 1   1  128.8      16.50     0.5000      834.9
 1   1  200.2      301.9      208.9      728.0
 1   1  358.3      481.8      354.0      742.4
 1   1  611.1      1010.      987.1      679.0
 1   1  499.3      752.5      921.6      725.7
 1   1  537.8      724.0      826.8      675.9
 1   1  220.0      523.3      908.2      640.8
 1   1  479.2      357.5      788.6      544.7
```

```
  1   1  133.2      195.3      383.1      738.5
  1   1  98.20      167.3      213.5      1065.
  1   1  92.50      104.6      142.2      741.2
  1   1  21.00      127.4      176.3      805.2
  1   1  353.0      307.8      167.1      576.9
  1   1  358.0      280.9      188.4      225.3
  1   1  233.3      355.3      234.9      719.1
  2   1  198.3      12.70      184.0      745.2
  2   1  507.2      3.000      126.3      664.9
  2   1  313.7     0.2000      30.00      667.1
  2   1  165.1      16.50      4.000      703.9
  2   1  295.6      329.9      454.6      695.3
  2   1  527.7      308.0      295.9      775.0
  2   1  454.1      301.0      1.000      995.6
  2   1  240.3      417.5      361.1      933.8
  2   1  590.8      579.3      144.2      666.5
  2   1  638.3      756.6      608.9      400.1
  2   1  949.8      1004.      805.4      528.9
  2   1  886.8      855.6      706.2      517.4
  2   1  635.5      761.0      670.9      596.6
  2   1  359.3      412.6      232.5      937.6
  2   1  484.7      360.7      226.8      979.0
  2   1  529.7      332.0      202.5      980.0
  2   1  585.8      291.4      186.1      1100.
  2   1  367.7      368.0      260.2      1006.
  2   1  324.7      270.9      72.70      1059.
  2   1  489.0      274.6      208.5      942.2
  2   1  570.8      337.1      218.0      646.5
  2   1  419.7      254.4      206.1      344.0
  2   1  532.8      414.2      197.9      477.0
  2   1  425.2      365.7      198.7      469.5
  2   1  467.5      411.5      228.5      455.3
  2   1  362.2      306.4      147.6      405.2
  2   1  429.2      287.4      139.2      450.6
  2   1  446.0      338.1      169.5      461.2
  2   1  192.9      253.8      145.6      460.7
  2   1  630.3      322.5      257.2      460.5
  2   1  364.9      326.7      251.1      510.6
  2   1  111.4      196.4      248.5      314.4
  2   1  89.80      146.5      254.9      123.2
  2   1  82.50      248.0      160.9      80.90
  2   1  296.5      253.2      193.2      230.4
  2   1  215.4      299.7      165.0      339.5
  2   1  454.5      274.2      154.0      120.4
  2   1  384.7      324.6      163.2      251.7
  2   1  253.2      488.3      175.6      122.4
  2   1  289.5      304.1      193.1      153.8

 Regime averaged data:  Co, Cp1, Cp2 ...

 439.4      509.5      569.1      636.3
 414.1      345.2      241.2      569.3
```

Segment 3

```
 Nominal (median) results          (No. of regimes =    2)
 MODEL      MEAN      SIGMA       BIAS       NMSE   CORR   FA2     FB      HIGH   2nd HIGH   PCOR
 OBS.       427.     235.39       0.00       0.00  1.000  1.000  0.000     1149.      1138.    n/a
                        (FBfn=  0.000, FBfp=  0.000, MOEfn=  1.000, MOEfp=  1.000, FB=FBfn-FBfp)

 MODEL-A    426.     286.37       0.29       0.17  0.784  0.835  0.001     1276.      1226.    n/a
                        (FBfn=  0.167, FBfp=  0.166, MOEfn=  0.833, MOEfp=  0.834, FB=FBfn-FBfp)

 MODEL-B    403.     296.46      23.48       0.34  0.612  0.570  0.057     1175.      1116.    n/a
                        (FBfn=  0.266, FBfp=  0.209, MOEfn=  0.742, MOEfp=  0.785, FB=FBfn-FBfp)

 MODEL-C    602.     228.27    -175.77       0.54  0.001  0.620 -0.342     1100.      1065.    n/a
                        (FBfn=  0.114, FBfp=  0.456, MOEfn=  0.862, MOEfp=  0.610, FB=FBfn-FBfp)


 Block   1: Urban data set                (N=   39)
 MODEL      MEAN      SIGMA       BIAS       NMSE   CORR   FA2     FB      HIGH   2nd HIGH   PCOR
 OBS.       439.     273.79       0.00       0.00  1.000  1.000  0.000     1149.      1138.    n/a
                        (FBfn=  0.000, FBfp=  0.000, MOEfn=  1.000, MOEfp=  1.000, FB=FBfn-FBfp)

 MODEL-A    509.     329.36     -70.05       0.16  0.847  0.821 -0.148     1276.      1226.    n/a
                        (FBfn=  0.087, FBfp=  0.234, MOEfn=  0.907, MOEfp=  0.782, FB=FBfn-FBfp)

 MODEL-B    569.     304.22    -129.70       0.24  0.747  0.718 -0.257     1175.      1116.    n/a
                        (FBfn=  0.056, FBfp=  0.313, MOEfn=  0.936, MOEfp=  0.723, FB=FBfn-FBfp)

 MODEL-C    636.     134.77    -196.86       0.57 -0.384  0.590 -0.366     1065.       835.    n/a
                        (FBfn=  0.118, FBfp=  0.484, MOEfn=  0.856, MOEfp=  0.591, FB=FBfn-FBfp)


 Block   2: Rural data set                (N=   40)
 MODEL      MEAN      SIGMA       BIAS       NMSE   CORR   FA2     FB      HIGH   2nd HIGH   PCOR
 OBS.       414.     189.82       0.00       0.00  1.000  1.000  0.000      950.       887.    n/a
                        (FBfn=  0.000, FBfp=  0.000, MOEfn=  1.000, MOEfp=  1.000, FB=FBfn-FBfp)

 MODEL-A    345.     207.08      68.87       0.20  0.709  0.850  0.181     1004.       856.    n/a
                        (FBfn=  0.265, FBfp=  0.083, MOEfn=  0.757, MOEfp=  0.908, FB=FBfn-FBfp)

 MODEL-B    241.     173.99     172.84       0.57  0.593  0.425  0.527      805.       706.    n/a
                        (FBfn=  0.581, FBfp=  0.053, MOEfn=  0.541, MOEfp=  0.928, FB=FBfn-FBfp)

 MODEL-C    569.     288.07    -155.20       0.50  0.239  0.650 -0.316     1100.      1059.    n/a
                        (FBfn=  0.111, FBfp=  0.427, MOEfn=  0.868, MOEfp=  0.632, FB=FBfn-FBfp)
```

Note: The Percentile 95% Confidence Limits are based on the 2.5th and 97.5th percentiles
of the cumulative distribution function.
The Student's t 95% Confidence Limits are based on calculated mean and standard deviation.

Segment 4

Model(s)		Student's t 95% Conf. limits		Student t	Mean	S.D.	Percentile 95% Conf. limits	
OBS.	MEAN	373.007	476.322	16.366	424.665	25.949	371.310	473.776
MODEL-A	NMSE	0.109	0.243	5.222	0.176	0.034	0.120	0.252
	FB	-0.084	0.085	0.007	0.000	0.043	-0.082	0.084
	FBfn	0.109	0.225	5.701	0.167	0.029	0.113	0.231
	FBfp	0.119	0.214	6.991	0.167	0.024	0.122	0.215
	CORR	0.670	0.886	14.360	0.778	0.054	0.653	0.864
MODEL-B	NMSE	0.224	0.466	5.688	0.345	0.061	0.240	0.470
	FB	-0.046	0.160	1.098	0.057	0.052	-0.047	0.161
	FBfn	0.200	0.334	7.951	0.267	0.034	0.206	0.332
	FBfp	0.140	0.281	5.941	0.210	0.035	0.145	0.283
	CORR	0.442	0.764	7.449	0.603	0.081	0.422	0.736
MODEL-C	NMSE	0.380	0.709	6.581	0.544	0.083	0.396	0.730
	FB	-0.484	-0.207	-4.971	-0.345	0.069	-0.481	-0.207
	FBfn	0.060	0.166	4.242	0.113	0.027	0.065	0.166
	FBfp	0.353	0.565	8.626	0.459	0.053	0.354	0.566
	CORR	-0.189	0.184	-0.029	-0.003	0.094	-0.200	0.176

Model(s)		Student's t 95% Conf. limits		Student t	Mean	S.D.	Percentile 95% Conf. limits	
MODEL-A - MODEL-B	NMSE	-0.269	-0.070	-3.396	-0.169	0.050	-0.279	-0.081
	FB	-0.132	0.019	-1.493	-0.057	0.038	-0.135	0.018
	FBfn	-0.155	-0.046	-3.647	-0.100	0.027	-0.156	-0.048
	FBfp	-0.096	0.009	-1.637	-0.043	0.027	-0.097	0.006
	CORR	0.072	0.278	3.380	0.175	0.052	0.082	0.289
MODEL-A - MODEL-C	NMSE	-0.536	-0.202	-4.392	-0.369	0.084	-0.549	-0.224
	FB	0.183	0.509	4.224	0.346	0.082	0.186	0.509
	FBfn	-0.037	0.144	1.182	0.054	0.045	-0.037	0.143
	FBfp	-0.397	-0.187	-5.549	-0.292	0.053	-0.396	-0.195
	CORR	0.551	1.011	6.761	0.781	0.115	0.543	1.000
MODEL-B - MODEL-C	NMSE	-0.369	-0.029	-2.336	-0.199	0.085	-0.379	-0.044
	FB	0.243	0.562	5.033	0.403	0.080	0.254	0.554
	FBfn	0.059	0.249	3.221	0.154	0.048	0.060	0.248
	FBfp	-0.350	-0.147	-4.877	-0.249	0.051	-0.355	-0.153
	CORR	0.350	0.861	4.712	0.606	0.129	0.357	0.843

Segment 5

SUMMARY OF CONFIDENCE LIMITS ANALYSES BASED ON PERCENTILE CONFIDENCE LIMITS

D(NMSE) among models: an 'X' indicates significantly different from zero at 95% confidence limits

	MODEL-A	MODEL-B	MODEL-C
MODEL-A		X	X
MODEL-B			X

D(FB) among models: an 'X' indicates significantly different from zero at 95% confidence limits

	MODEL-A	MODEL-B	MODEL-C
MODEL-A			X
MODEL-B			X

D(FBfn) among models: an 'X' indicates significantly different from zero at 95% confidence limits

	MODEL-A	MODEL-B	MODEL-C
MODEL-A		X	
MODEL-B			X

D(FBfp) among models: an 'X' indicates significantly different from zero at 95% confidence limits

```
             M   M   M
             O   O   O
             D   D   D
             E   E   E
             L   L   L
             -   -   -
             A   B   C
            -----------
MODEL-A |           X
MODEL-B |           X

 D(CORR) among models: an 'X' indicates significantly different from zero at 95% confidence limits

             M   M   M
             O   O   O
             D   D   D
             E   E   E
             L   L   L
             -   -   -
             A   B   C
            -----------
MODEL-A |       X   X
MODEL-B |           X

 FB for each model: an 'X' indicates significantly different from zero at 95% confidence limits

             M   M   M
             O   O   O
             D   D   D
             E   E   E
             L   L   L
             -   -   -
             A   B   C
            -----------
                    X

 FBfn for each model: an 'X' indicates significantly different from zero at 95% confidence limits

             M   M   M
             O   O   O
             D   D   D
             E   E   E
             L   L   L
             -   -   -
             A   B   C
            -----------
             X   X   X

 FBfp for each model: an 'X' indicates significantly different from zero at 95% confidence limits

             M   M   M
             O   O   O
             D   D   D
             E   E   E
             L   L   L
             -   -   -
             A   B   C
            -----------
             X   X   X

 CORR for each model: an 'X' indicates significantly different from zero at 95% confidence limits

             M   M   M
             O   O   O
             D   D   D
             E   E   E
             L   L   L
             -   -   -
             A   B   C
            -----------
             X   X
```

Table 5. Sample optional output file generated by BOOT, based on the sample input file shown in Table 3, that shows the FB (together with the 95% confidence limits) and NMSE for each model. The first four lines are specific to an in-house plotting package, and are mainly for the x and y labels of the graph, and the number (3 in this case) of data objects to display. The remaining lines have five values each, including the median (nominal) value of NMSE, the lower confidence limit of FB, the median (nominal) value of FB, the upper confidence limit of FB, and the model name. The same information is also listed in the mandatory output file given appropriate command-line options. (See Segment 3 of Table 4 for the information on median values, and Segment 4 of Table 4 for the information on confidence limits.) For example, the sample below (the fifth line) shows that for Model-A the median NMSE is 0.174, the median FB is 0.000677, and the 95% confidence interval for FB is (-0.0822, 0.0838).

```
0
FB (with 95% conf. int.)
NMSE
3  1
0.174467146 -8.22452828E-02 6.76780124E-04 8.38098750E-02  'MODEL-A'
0.339648187 -4.73854281E-02 5.66054508E-02 0.161254287  'MODEL-B'
0.538153350 -0.481164068 -0.341653824 -0.206860647  'MODEL-C'
```

Table 6. Sample optional output file generated by BOOT, based on the sample input file shown in Table 3, that shows the difference in FB (d(FB), together with the 95% confidence limits) and the difference in NMSE (d(NMSE)) for each model pair. The first four lines are specific to an in-house plotting package, and are mainly for the x and y labels of the graph, and the number (3 in this case) of data objects to display. The remaining lines have five values each, including the median (nominal) value of d(NMSE), the lower confidence limit of d(FB), the median (nominal) value of d(FB), the upper confidence limit of d(FB), and the name of the model pair. The same information is also listed in the mandatory output file given appropriate command-line options. (See Segment 3 of Table 4 for the information on median values, and Segment 4 of Table 4 for the information on confidence limits.) For example, the sample below (the fifth line) shows that for the pair of Model-A and Model-B, the median d(NMSE) is –0.165, the median d(FB) is –0.0559, and the 95% confidence interval for d(FB) is (-0.135, 0.0183).

```
0
d(FB) (with 95% conf. int.)
d(NMSE)
3  1
-0.165181041 -0.135485172 -5.59286699E-02 1.82778761E-02  'MODEL-A-MODEL-B'
-0.363686204 0.186301097 0.342330605 0.508606791  'MODEL-A-MODEL-C'
-0.198505163 0.254026473 0.398259282 0.554421723  'MODEL-B-MODEL-C'
```

Table 7. Summary of performance measures, including FB, MG, NMSE, VG, R, FAC2, the average, the standard deviation (σ), the highest, the second highest, FB_{FN}, FB_{FP}, MOE_{FN}, and MOE_{FP} for the sample database shown in Table 3 (or Table 1). The information is retrieved from the sample BOOT mandatory output file in Table 4.

	Observed	Model-A	Model-B	Model-C
FB (Eq. (1))	n/a	0.001	0.057	-0.342
MG (Eq. (2))	n/a	1.22	1.34	0.65
NMSE (Eq. (3))	n/a	0.17	0.34	0.54
VG (Eq. (4))	n/a	4.20	4.99	2.28
R (Eq. (5))	n/a	0.784	0.612	0.001
FAC2 (Eq. (6))	n/a	0.84	0.57	0.62
Average	427	426	403	602
σ	235	286	296	228
Highest	1149	1276	1175	1100
2^{nd} Highest	1138	1226	1116	1065
FB_{FN} (Eq. (9))	n/a	0.167	0.266	0.114
FB_{FP} (Eq. (10))	n/a	0.166	0.209	0.456
MOE_{FN} (Eq. (12))	n/a	0.833	0.742	0.862
MOE_{FP} (Eq. (13))	n/a	0.834	0.785	0.610

Table 8. Summary of significance tests for the FB, R, and ln(MG) *for each model* (rows 2 through 4); and for the differences (Δ) in NMSE, FB, R, ln(VG), and ln(MG) *for each model pair* (rows 5 through 7) for the sample database shown in Table 3 (or Table 1). Tables 5 and 6 listed optional BOOT output files for these same parameters. A mark "×" means that the parameter is significantly different from zero at the 95% confidence level based on the percentile confidence limits. For example, the table shows that the FB for Model-A is not significantly different from zero, and that the difference in the FB's for Model-A and Model-C is significantly different from zero. Note that, by definition, NMSE and ln(VG) for each model is always greater than zero, thus "n/a" in the corresponding cells.

	NMSE	FB	R	ln(VG)	ln(MG)
Model-A	n/a		×	n/a	
Model-B	n/a		×	n/a	×
Model-C	n/a	×		n/a	×
Δ(Model-A, Model-B)	×		×		
Δ(Model-A, Model-C)	×	×	×		×
Δ(Model-B, Model-C)	×	×	×		×

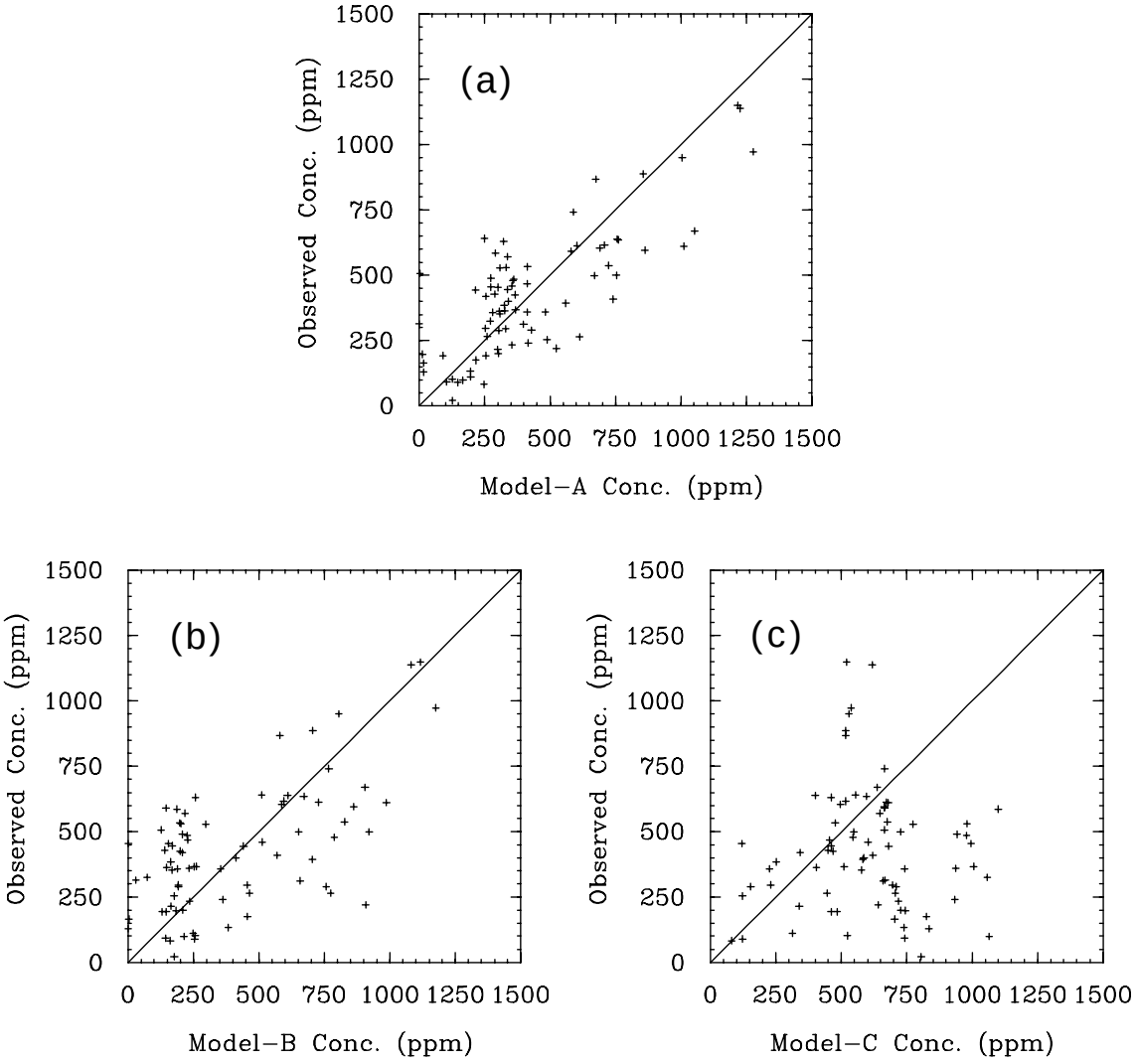

Figure 1. Scatter plots of observed versus predicted concentrations for the three models listed in Table 1. (a) Model-A, (b) Model-B, and (c) Model-C.

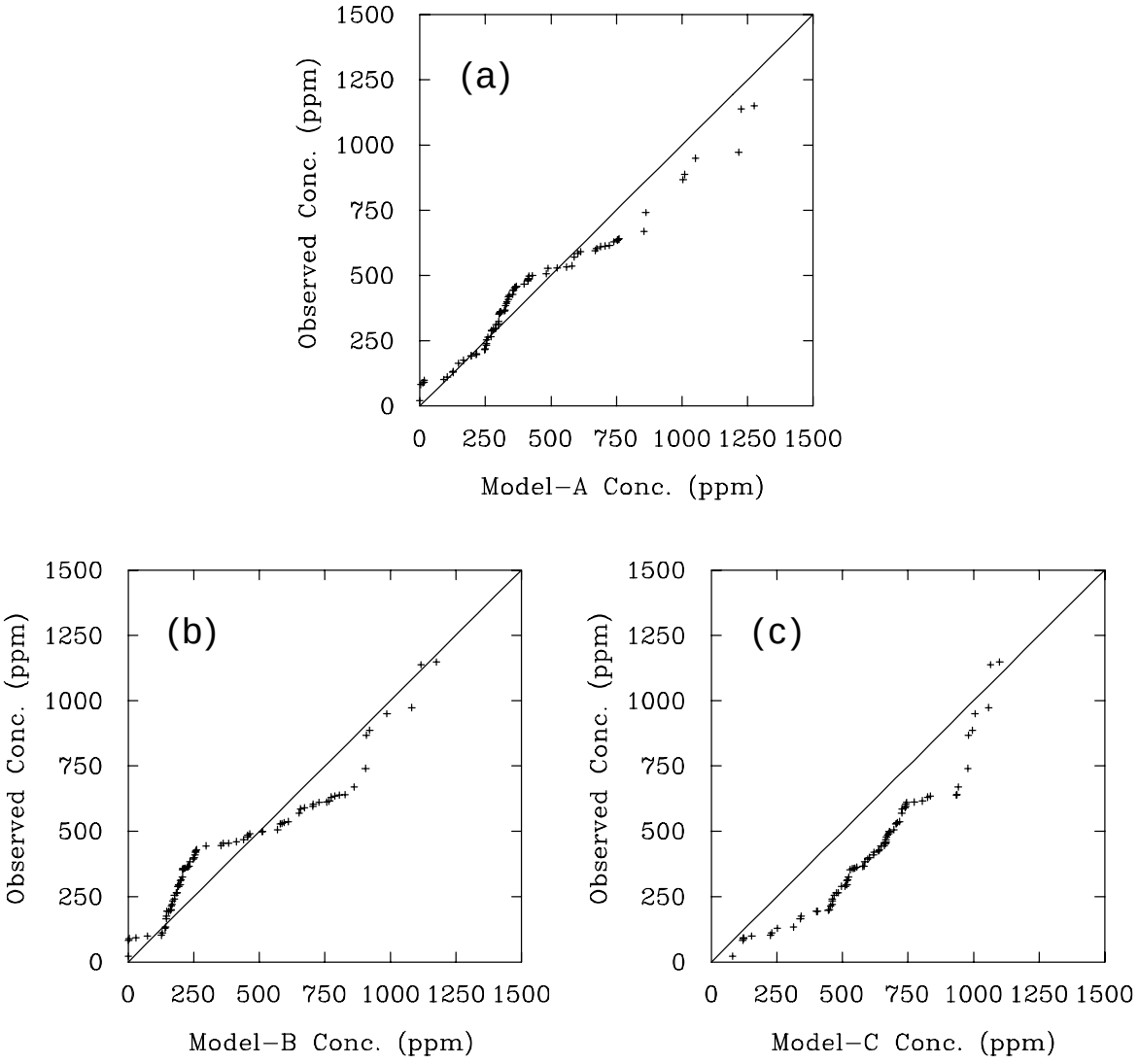

Figure 2. Quantile-quantile plots of observed versus predicted concentrations for the three models listed in Table 1, where predicted and observed concentrations are separately ranked. (a) Model-A, (b) Model-B, and (c) Model-C.

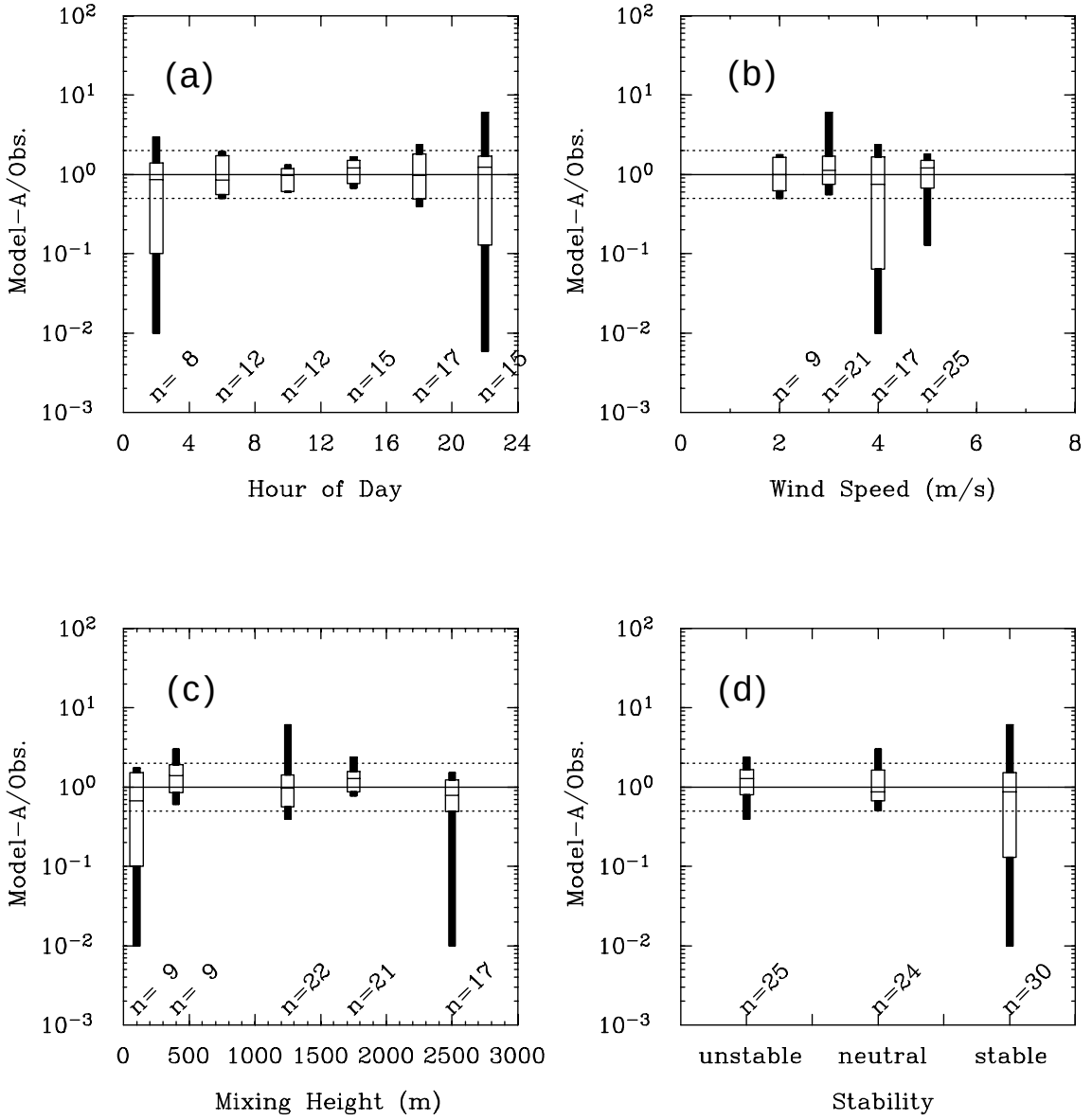

Figure 3. Box plots of model residuals (ratios of predicted to observed concentrations) for Model-A as functions of (a) time of day (local hour), (b) ambient wind speed (m s^{-1}), (c) mixing height (m), and (d) atmospheric stability, where "unstable", "neutral", and "stable" refer to stability classes 1 through 3, 4, and 5 through 6, respectively. The significant points for each box indicate the 2nd, 16th, 50th, 84th, and 98th percentiles of the cumulative distribution of the n points considered in the bin of data used in the box. Dashed lines indicate factor-of-two scatter. See Table 1 for a listing of the data used.

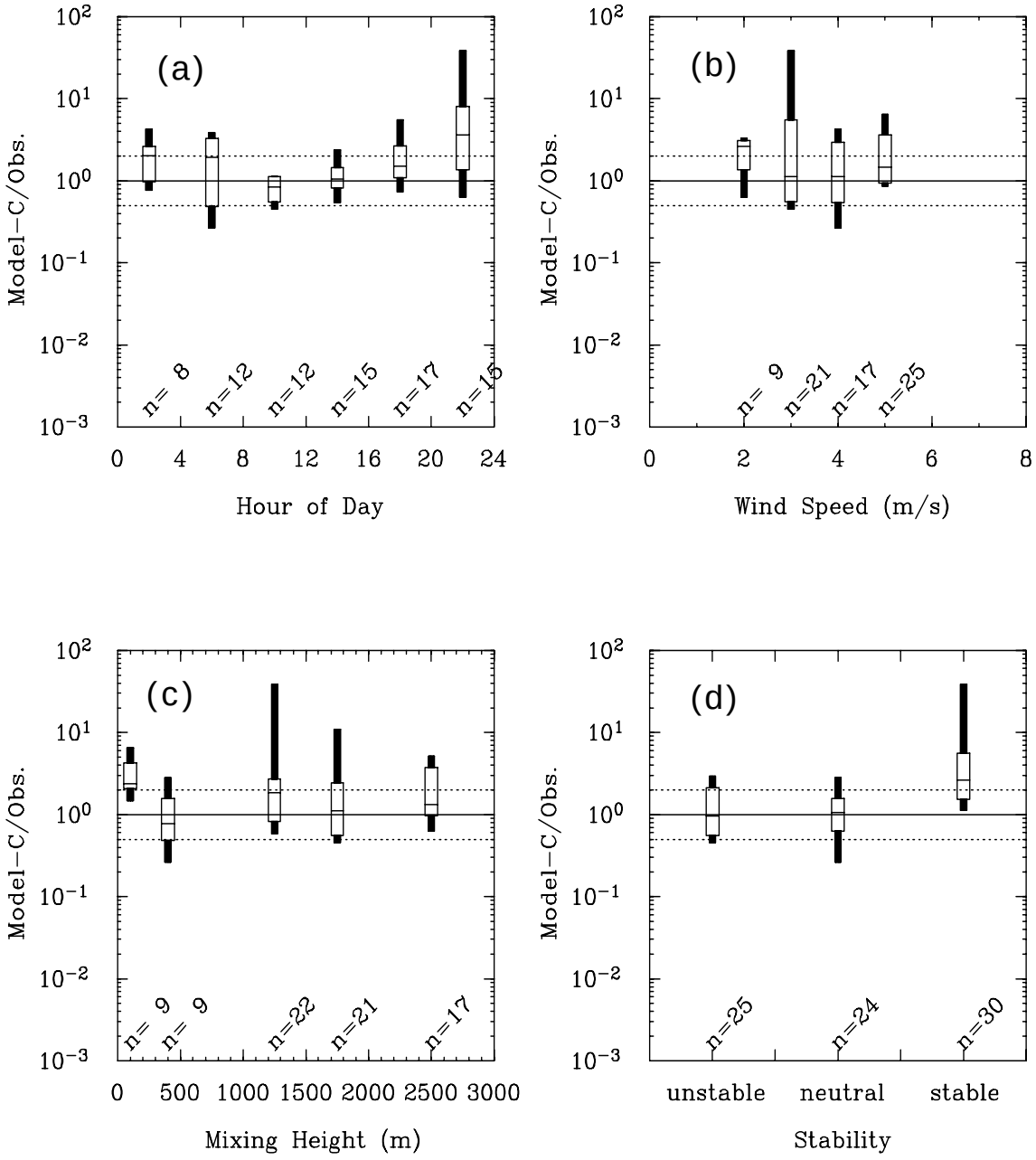

Figure 4. Box plots of model residuals (ratios of predicted to observed concentrations) for Model-C as functions of (a) time of day (local hour), (b) ambient wind speed (m s^{-1}), (c) mixing height (m), and (d) atmospheric stability, where "unstable", "neutral", and "stable" refer to stability classes 1 through 3, 4, and 5 through 6, respectively. The significant points for each box indicate the 2nd, 16th, 50th, 84th, and 98th percentiles of the cumulative distribution of the n points in the bin of data used in the box. Dashed lines indicate factor-of-two scatter. Notice the slight trends with time of day and atmospheric stability. See Table 1 for a listing of data used.

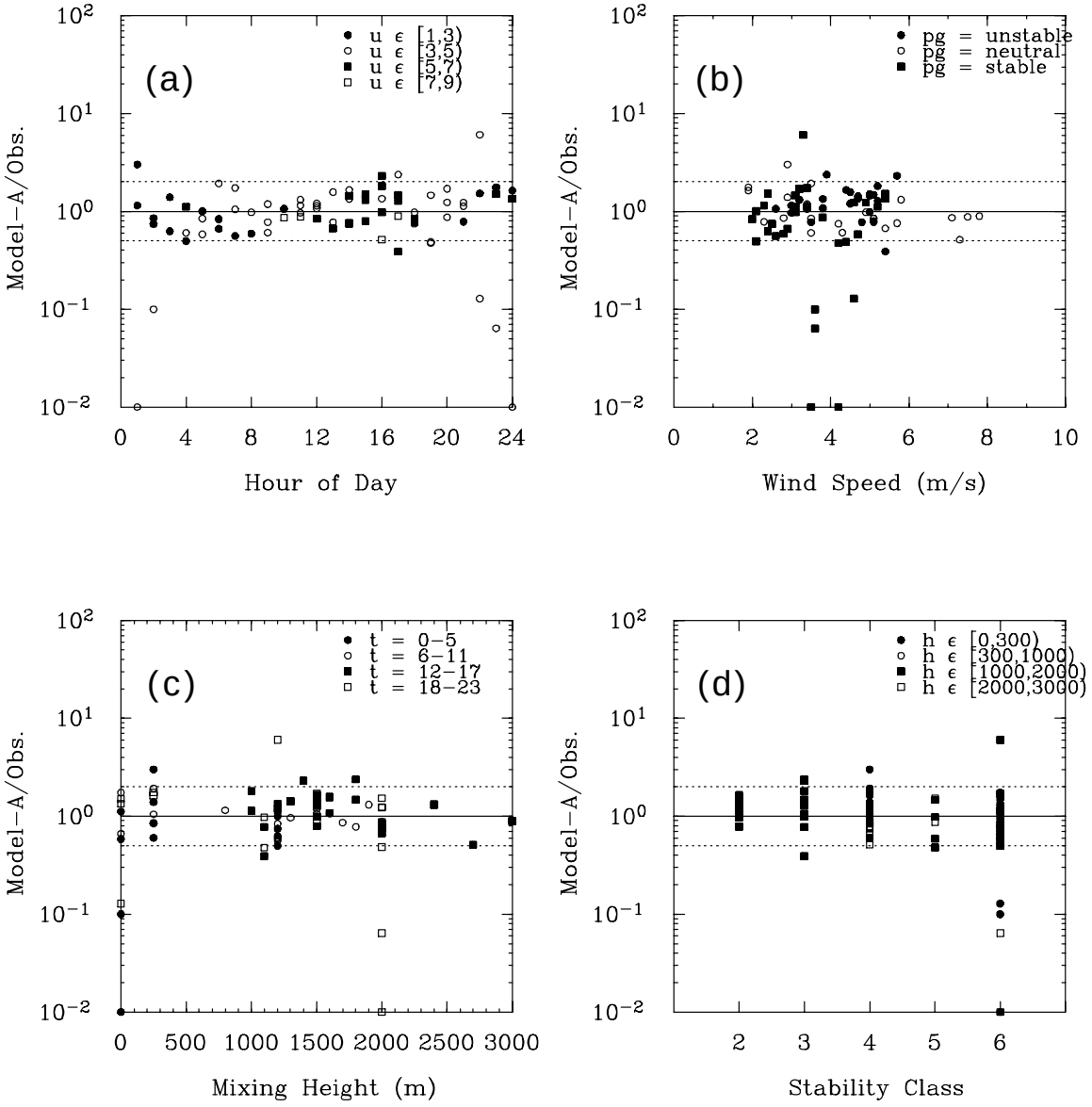

Figure 5. Conditional scatter plots of model residuals (ratios of predicted to observed concentrations) for Model-A as functions of (a) time of day (local hour), (b) ambient wind speed (m s^{-1}), (c) mixing height (m), and (d) stability class, where different symbols are used to indicate different ranges of a second independent variable. Dashed lines indicate factor-of-two scatter. See Table 1 for a listing of the data used.

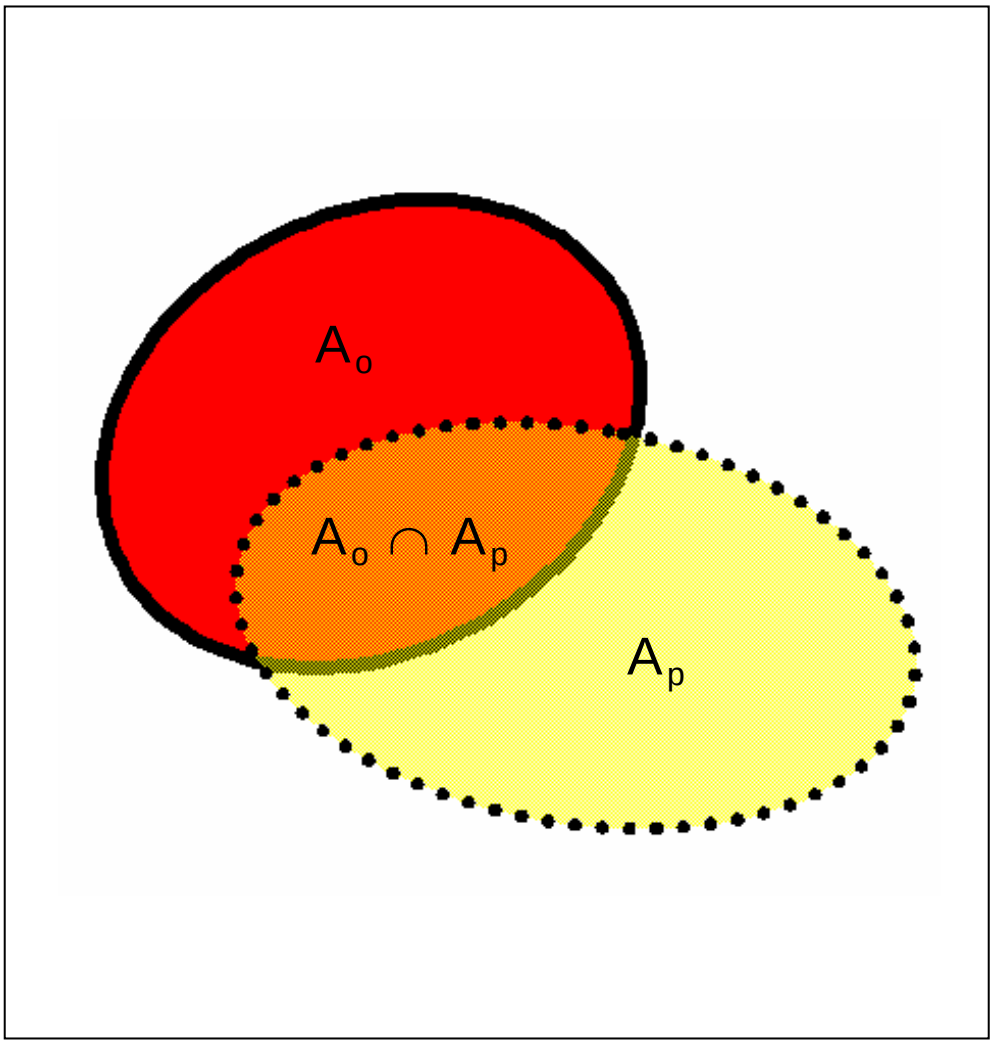

Figure 6. Schematic diagram illustrating the areas used to calculate the Figure of Merit in Space (FMS), $A_p \cap A_o / A_p \cup A_o$, where A_p (area enclosed by thick dotted line) is the predicted contour area, and A_o (area enclosed by thick solid line) is the observed contour area. The contour can be defined, for example, by a concentration threshold for the released chemical for dispersion modeling, or by areas of precipitation for weather forecast. The orange area is the intersection area ($A_p \cap A_o$), the red area is the false-negative (or underpredicting) area, and the yellow area is the false-positive (or overpredicting) area.

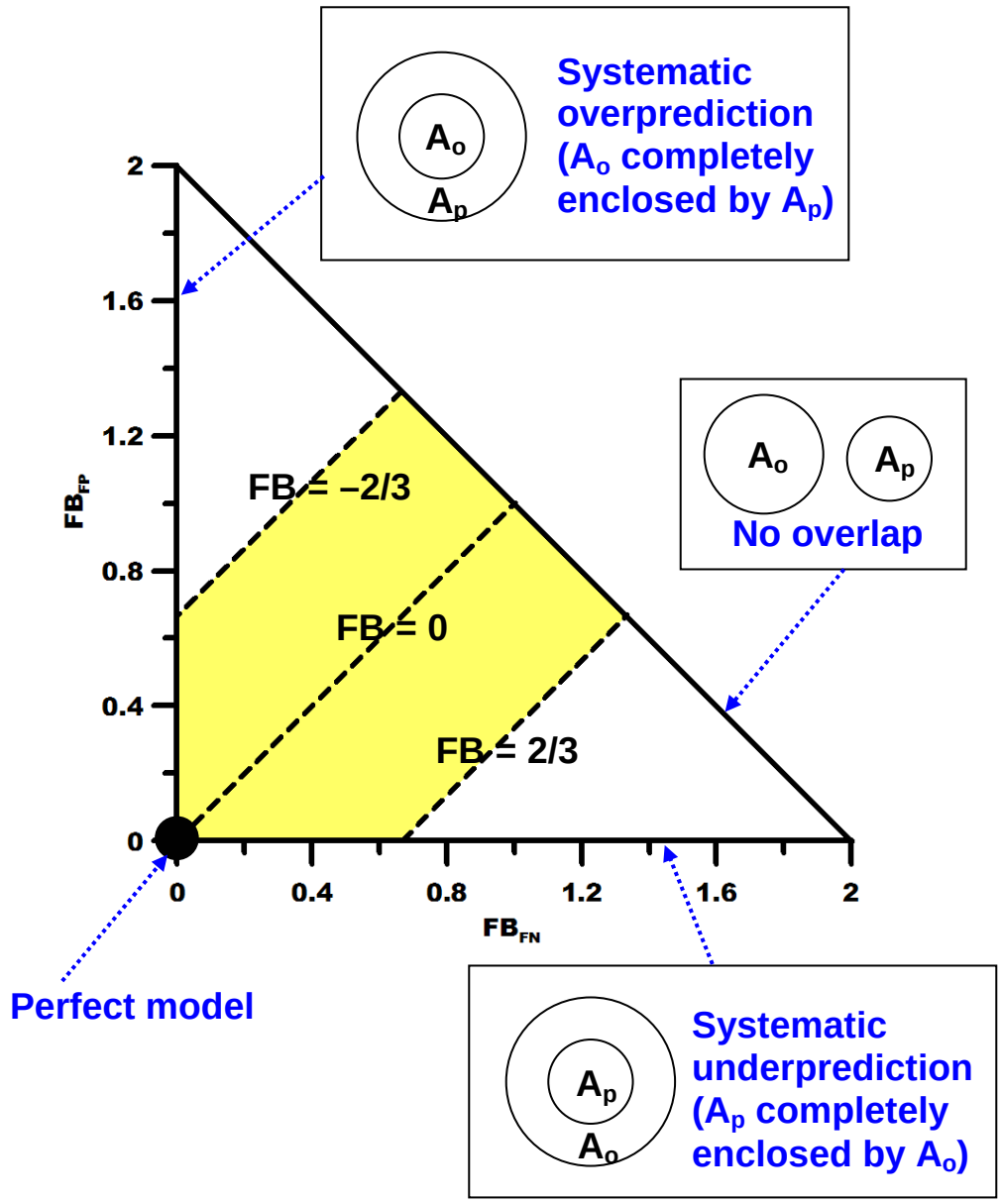

Figure 7. Two-dimensional fractional bias (FB) diagram, with the x- and y-coordinates (FB_{FN} and FB_{FP}) defined in Eqs. (9) and (10), or Eqs. (21) and (22). A perfect model would be located at $FB_{FN} = FB_{FP} = 0$. The x-axis indicates systematic underprediction *(i.e.*, no false positive), the y-axis indicates systematic overprediction (*i.e.*, no false negative), and the hypotenuse of the larger triangle indicates no overlap between predictions and observations *(i.e.*, predictions are zero whenever observations are finite, and vice versa). The three dashed lines correspond to FB (Eq. (8)) = 0, 2/3, and -2/3; and the shaded area represents $|FB| < 2/3$, or a mean bias that is less than a factor of two.

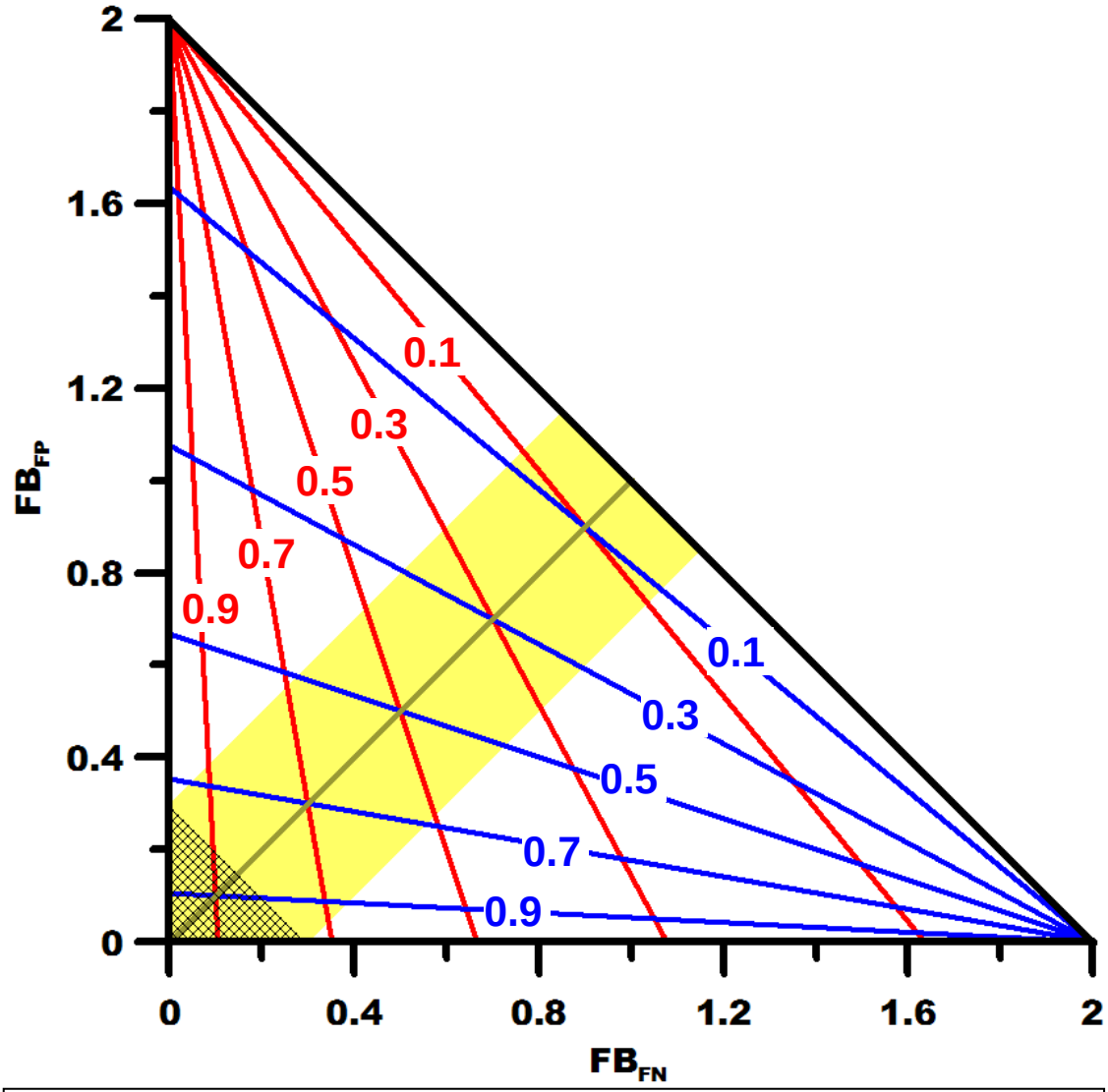

Figure 8. Contour lines of the two-dimensional Measure of Effectiveness (MOE_{FN} in red and MOE_{FP} in blue) as a function of the false-negative (underpredicting) and false-positive (overpredicting) components of the Fractional Bias (FB) based on Eqs. (26) and (27). For example, (FB_{FN}, FB_{FP}) = (0.4, 0.2) would correspond to (MOE_{FN}, MOE_{FP}) = (0.64, 0.78). The shaded area corresponds to $|FB| < 0.3$, *i.e.*, $|FB_{FN} - FB_{FP}| < 0.3$. The solid line in the middle of the shaded area corresponds to FB = 0, *i.e.*, $FB_{FN} = FB_{FP}$ and $MOE_{FN} = MOE_{FP}$. The cross hatched area corresponds to $|AFB| < 0.3$, *i.e.*, $FB_{FN} + FB_{FP} < 0.3$, where AFB is the Absolute Fractional Bias (Eq. 11).

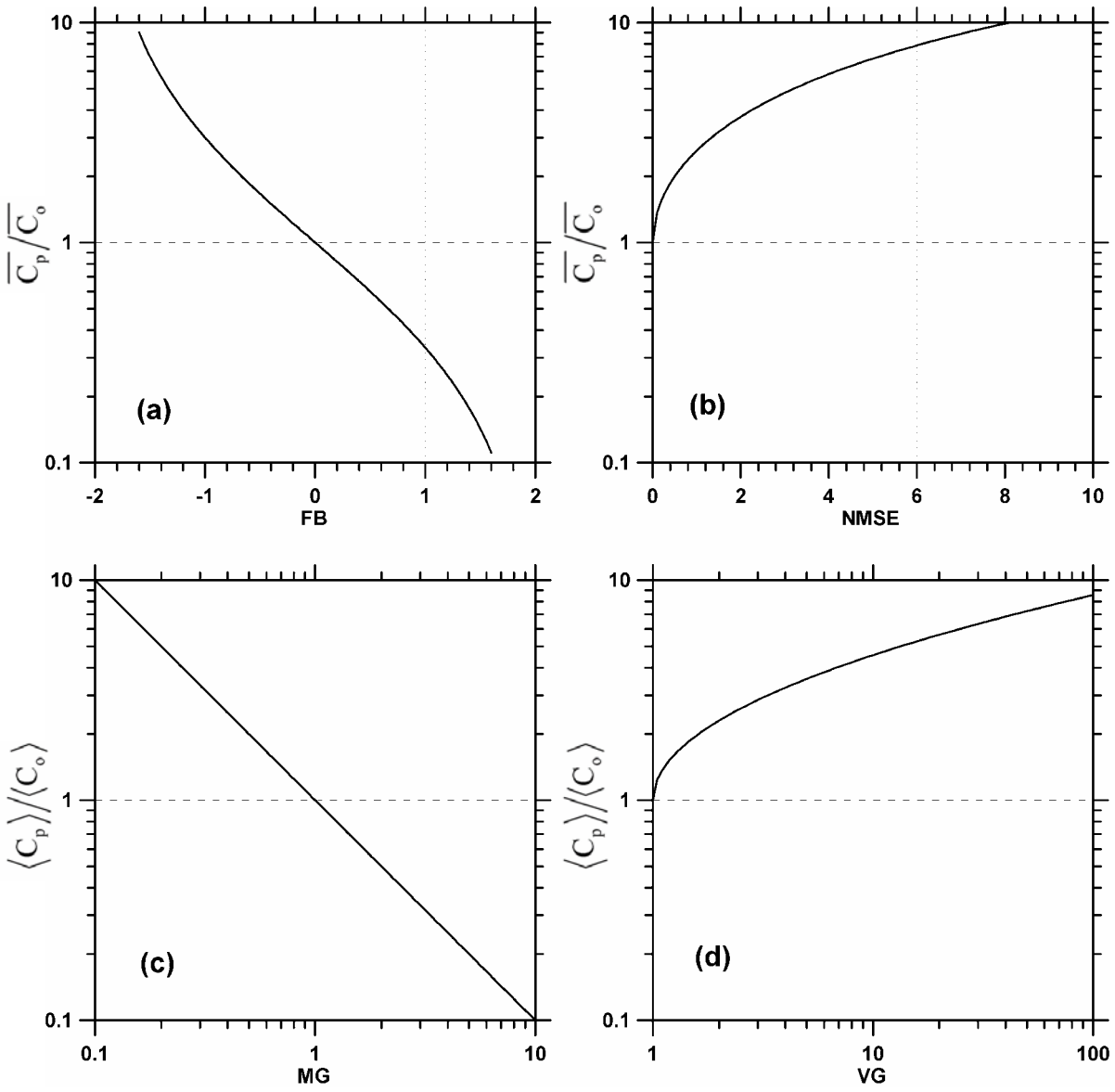

Figure 9. Relationships between $\overline{C_p}/\overline{C_o}$ and (a) FB (Eq. (31)) and (b) NMSE (Eq. (32)), and between $\langle C_p \rangle / \langle C_o \rangle$ and (c) MG (Eq. (33)) and (d) VG (Eq. (34)), where C_p is the model prediction, C_o is the observation, the overbar represents the linear average, and the angle brackets represent the geometric average. The relationships in (a) and (c) are exact, whereas the relationships in (b) and (d) are based on assumptions that ignore the random scatter between C_p and C_o. The relationships in (b) and (d) can also be interpreted as the reciprocal of $\overline{C_p}/\overline{C_o}$ and $\langle C_p \rangle / \langle C_o \rangle$, respectively, because NMSE and VG involve the square of errors. From Chang (2002).

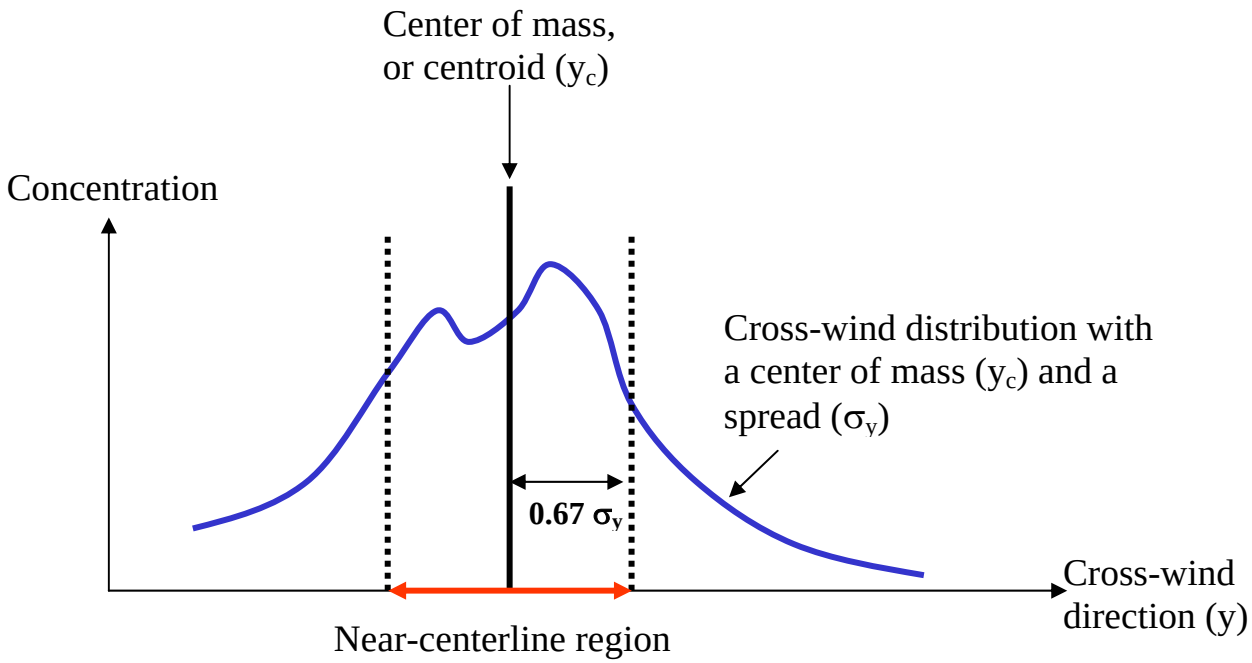

Figure 10. Schematic diagram illustrating the definition of a near-centerline region used by the ASTM procedure. A sample cross-wind concentration distribution is shown. The first moment of the distribution defines the center-of-mass (or centroid) location, y_c. The second moment defines the spread, σ_y, of the distribution. The region, marked by dotted lines, that is within 0.67 σ_y from the centroid location is the near-centerline region.

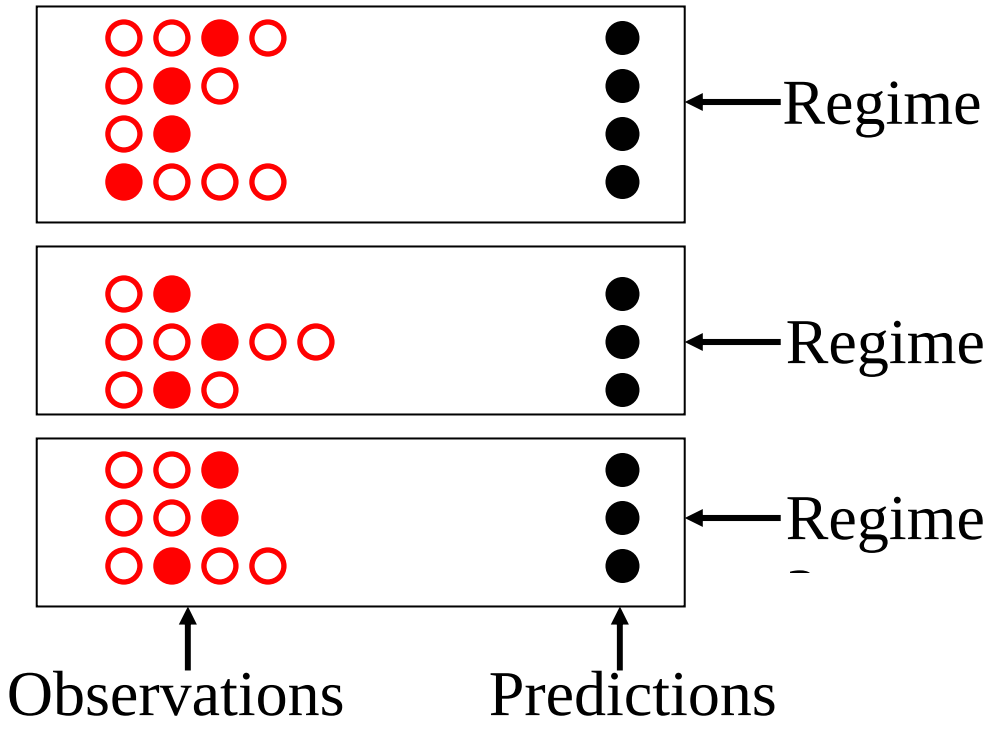

Figure 11. Schematic for the ASTM model evaluation procedure. The example has ten experiments that are further grouped into three regimes. Black solid circles represent the ten centerline predictions for the ten experiments. Red open and solid circles represent near-centerline observations that are considered representative of centerline concentrations. Red solid circles represent the maximum observed concentrations for each experiment. Traditional model evaluation compares black solid with red solid circles. The ASTM procedure compares black solid with red solid *and* open circles. See text and Fig. 10 for the definition of the near-centerline region.

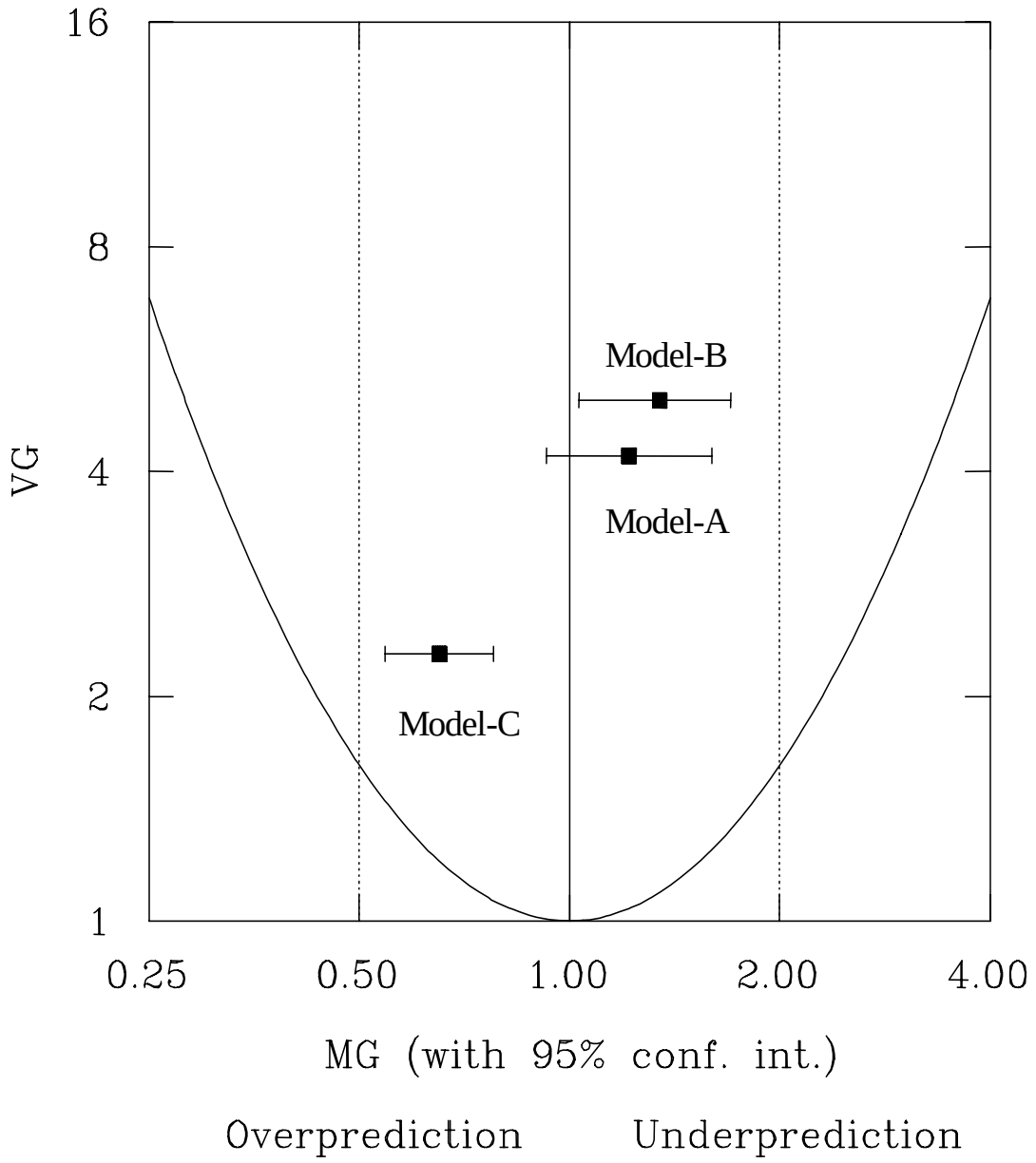

Figure 12. The MG (Eq. (2)) and VG (Eq. (4)) for the sample dataset and the three models listed in Table 3. The 95% confidence limits for MG based on the bootstrap resampling (the percentile confidence limits) are also indicated by thin horizontal bars. The solid parabola represents the minimum VG for a given value of MG, assuming all scatter is due solely to the mean bias (see Eq. (30)). Dotted lines represent a plus and minus factor-of-two mean bias for predictions. A perfect model would have MG = VG = 1.0, *i.e.*, located at the bottom center of the diagram.