Designation: D 6589 – 00

Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance[1]

This standard is issued under the fixed designation D 6589; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ϵ) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This guide provides techniques that are useful for the comparison of modeled air concentrations with observed field data. Such comparisons provide a means for assessing a model's performance, for example, bias and precision or uncertainty, relative to other candidate models. Methodologies for such comparisons are yet evolving; hence, modifications will occur in the statistical tests and procedures and data analysis as work progresses in this area. Until the interested parties agree upon standard testing protocols, differences in approach will occur. This guide describes a framework, or philosophical context, within which one determines whether a model's performance is significantly different from other candidate models. It is suggested that the first step should be to determine which model's estimates are closest on average to the observations, and the second step would then test whether the differences seen in the performance of the other models are significantly different from the model chosen in the first step. An example procedure is provided in Appendix X1 to illustrate an existing approach for a particular evaluation goal. This example is not intended to inhibit alternative approaches or techniques that will produce equivalent or superior results. As discussed in Section 6, statistical evaluation of model performance is viewed as part of a larger process that collectively is referred to as model evaluation.

1.2 This guide has been designed with flexibility to allow expansion to address various characterizations of atmospheric dispersion, which might involve dose or concentration fluctuations, to allow development of application-specific evaluation schemes, and to allow use of various statistical comparison metrics. No assumptions are made regarding the manner in which the models characterize the dispersion.

1.3 The focus of this guide is on end results, that is, the accuracy of model predictions and the discernment of whether differences seen between models are significant, rather than operational details such as the ease of model implementation or the time required for model calculations to be performed.

1.4 This guide offers an organized collection of information or a series of options and does not recommend a specific course of action. This guide cannot replace education or experience and should be used in conjunction with professional judgment. Not all aspects of this guide may be applicable in all circumstances. This guide is not intended to represent or replace the standard of care by which the adequacy of a given professional service must be judged, nor should it be applied without consideration of a project's many unique aspects. The word "Standard" in the title of this guide means only that the document has been approved through the ASTM consensus process.

1.5 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety and health practices and to determine the applicability of regulatory limitations prior to use.*

2. Referenced Documents

2.1 *ASTM Standards:*

D 1356 Terminology Relating to Sampling and Analysis of Atmospheres[2]

3. Terminology

3.1 *Definitions*—For definitions of terms used in this guide, refer to Terminology D 1356.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *atmospheric dispersion model*, *n*—an idealization of atmospheric physics and processes to calculate the magnitude and location of pollutant concentrations based on fate, transport, and dispersion in the atmosphere. This may take the form of an equation, algorithm, or series of equations/algorithms used to calculate average or time-varying concentration. The model may involve numerical methods for solution.

3.2.2 *dispersion*, *absolute*, *n*—the characterization of the spreading of material released into the atmosphere based on a coordinate system fixed in space.

3.2.3 *dispersion*, *relative*, *n*—the characterization of the spreading of material released into the atmosphere based on a

[1] This guide is under the jurisdiction of ASTM Committee D22 on Sampling and Analysis of Atmospheres and is the direct responsibility of Subcommittee D22.11 on Meteorology.

Current edition approved Sept. 10, 2000. Published November 2000.

[2] *Annual Book of ASTM Standards*, Vol 11.03.

coordinate system that is relative to the local median position of the dispersing material.

3.2.4 *evaluation objective*, *n*—a feature or characteristic, which can be defined through an analysis of the observed concentration pattern, for example, maximum centerline concentration or lateral extent of the average concentration pattern as a function of downwind distance, which one desires to assess the skill of the models to reproduce.

3.2.5 *evaluation procedure*, *n*—the analysis steps to be taken to compute the value of the evaluation objective from the observed and modeled patterns of concentration values.

3.2.6 *fate*, *n*—the destiny of a chemical or biological pollutant after release into the environment.

3.2.7 *model input value*, *n*—characterizations that must be estimated or provided by the model developer or user before model calculations can be performed.

3.2.8 *regime*, *n*—a repeatable narrow range of conditions, defined in terms of model input values, which may or may not be explicitly employed by all models being tested, needed for dispersion model calculations. It is envisioned that the dispersion observed should be similar for all cases having similar model input values.

3.2.9 *uncertainty*, *n*—refers to a lack of knowledge about specific factors or parameters. This includes measurement errors, sampling errors, systematic errors, and differences arising from simplification of real-world processes. In principle, uncertainty can be reduced with further information or knowledge **(1)**[3].

3.2.10 *variability*, *n*—refers to differences attributable to true heterogeneity or diversity in atmospheric processes that result in part from natural random processes. Variability usually is not reducible by further increases in knowledge, but it can in principle be better characterized **(1)**.

4. Summary of Guide

4.1 Statistical evaluation of dispersion model performance with field data is viewed as part of a larger process that collectively is called model evaluation. Section 6 discusses the components of model evaluation.

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

4.3 In assessing the performance of air quality models to characterize a particular evaluation objective, one should consider what the models are capable of providing. As discussed in Section 7, most models attempt to characterize the ensemble average concentration pattern. If such models should provide favorable comparisons with observed concentration maxima, this is resulting from happenstance, rather than skill in the model; therefore, in this discussion, it is suggested a model be assessed on its ability to reproduce what it was designed to produce, for at least in these comparisons, one can be assured that zero bias with the least amount of scatter is by definition good model performance.

4.4 As an illustration of the principles espoused in this guide, a procedure is provided in Appendix X1 for comparison of observed and modeled near-centerline concentration values, which accommodates the fact that observed concentration values include a large component of stochastic, and possibly deterministic, variability unaccounted for by current models. The procedure provides an objective statistical test of whether differences seen in model performance are significant.

5. Significance and Use

5.1 Guidance is provided on designing model evaluation performance precedures and on the difficulties that arise in statistical evaluation of model performance caused by the stochastic nature of dispersion in the atmosphere. It is recognized there are examples in the literature where, knowingly or unknowingly, models were evaluated on their ability to describe something which they were never intended to characterize. This guide is attempting to heighten awareness, and thereby, to reduce the number of "unknowing" comparisons. A goal of this guide is to stimulate development and testing of evaluation procedures that accommodate the effects of natural variability. A technique is illustrated to provide information from which subsequent evaluation and standardization can be derived.

6. Model Evaluation

6.1 *Background*—Air quality simulation models have been used for many decades to characterize the transport and dispersion of material in the atmosphere **(2-4)**. Early evaluations of model performance usually relied on linear least-squares analyses of observed versus modeled values, using traditional scatter plots of the values, **(5-7)**. During the 1980s, attempts have been made to encourage the standardization of methods used to judge air quality model performance **(8-11)**. Further development of these proposed statistical evaluation procedures was needed, as it was found that the rote application of statistical metrics, such as those listed in **(8)**, was incapable of discerning differences in model performance **(12)**, whereas if the evaluation results were sorted by stability and distance downwind, then differences in modeling skill could be discerned **(13)**. It was becoming increasingly evident that the models were characterizing only a small portion of the observed variations in the concentration values **(14)**. To better deduce the statistical significance of differences seen in model performance in the face of large unaccounted for uncertainties and variations, investigators began to explore the use of bootstrap techniques **(15)**. By the late 1980s, most of the model

[3] The boldface numbers in parentheses refer to the list of references at the end of this standard.

performance evaluations involved the use of bootstrap techniques in the comparison of maximum values of modeled and observed cumulative frequency distributions of the concentrations values **(16)**. Even though the procedures and metrics to be employed in describing the performance of air quality simulation models are still evolving **(17-19)**, there has been a general acceptance that defining performance of air quality models needs to address the large uncertainties inherent in attempting to characterize atmospheric fate, transport and dispersion processes. There also has been a consensus reached on the philosophical reasons that models of earth science processes can never be validated, in the sense of claiming that a model is truthfully representing natural processes. No general empirical proposition about the natural world can be certain, since there will always remain the prospect that future observations may call the theory in question **(20)**. It is seen that numerical models of air pollution are a form of a highly complex scientific hypothesis concerning natural processes, that can be confirmed through comparison with observations, but never validated.

6.2 *Components of Model Evaluation*—A model evaluation includes science peer reviews and statistical evaluations with field data. The completion of each of these components assumes specific model goals and evaluation objectives (see Section 10) have been defined.

6.3 *Science Peer Reviews*—Given the complexity of characterizing atmospheric processes, and the inevitable necessity of limiting model algorithms to a resolvable set, one component of a model evaluation is to review the model's science to confirm that the construct is reasonable and defensible for the defined evaluation objectives. A key part of the scientific peer review will include the review of residual plots where modeled and observed evaluation objectives are compared over a range of model inputs, for example, maximum concentrations as a function of estimated plume rise or as a function of distance downwind.

6.4 *Statistical Evaluations with Field Data*—The objective comparison of modeled concentrations with observed field data provides a means for assessing model performance. Due to the limited supply of evaluation data sets, there are severe practical limits in assessing model performance. For this reason, the conclusions reached in the science peer reviews (see 6.3) and the supportive analyses (see 6.5) have particular relevance in deciding whether a model can be applied for the defined model evaluation objectives. In order to conduct a statistical comparison, one will have to define one or more evaluation objectives for which objective comparisons are desired (Section 10). As discussed in 8.4.4, the process of summarizing the overall performance of a model over the range of conditions experienced within a field experiment typically involves determining two points for each of the model evaluation objectives: which of the models being assessed has on average the smallest combined bias and scatter in comparisons with observations, and whether the differences seen in the comparisons with the other models statistically are significant in light of the uncertainties in the observations.

6.5 *Other Tasks Supportive to Model Evaluation*—As atmospheric dispersion models become more sophisticated, it is not easy to detect coding errors in the implementation of the model algorithms. And as models become more complex, discerning the sensitivity of the modeling results to input parameter variations becomes less clear; hence, two important tasks that support model evaluation efforts are verification of software and sensitivity and Monte Carlo analyses.

6.5.1 *Verification of Software*—Often a set of modeling algorithms will require numerical solution. An important task supportive to a model evaluation is a review in which the mathematics described in the technical description of the model are compared with the numerical coding, to insure that the code faithfully implements the physics and mathematics.

6.5.2 *Sensitivity and Monte Carlo Analyses*—Sensitivity and Monte Carlo analyses provide insight into the response of a model to input variation. An example of this technique is to systematically vary one or more of the model inputs to determine the effect on the modeling results **(22)**. Each input should be varied over a reasonable range likely to be encountered. The traditional sensitivity studies **(22)** were developed to better understand the performance of plume dispersion models simulating the transport and dispersion of inert pollutants. For characterization of the effects of input uncertainties on modeling results, Monte Carlo studies with simple random sampling are recommended **(23)**, especially for models simulating chemically reactive species where there are strong nonlinear couplings between the model input and output **(24)**. Results from sensitivity and Monte Carlo analyses provide useful guidance on which inputs should be most carefully prescribed because they account for the greatest sensitivity in the modeling output. These analyses also provide a view of what to expect for model output in conditions for which data are not available.

7. A Framework for Model Evaluations

7.1 This section introduces a philosophical model for explaining how and why observations of physical processes and model simulations of physical processes differ. It is argued that observations are individual realizations, which in principle can be envisioned as belonging to some ensemble. Most of the current models attempt to characterize the average concentration for each ensemble, but there are under development models that attempt to characterize the distribution of concentration values within an ensemble. Having this framework for describing how and why observations differ from model simulations has important ramifications in how one assesses and describes a model's ability to reproduce what is seen by way of observations. This framework provides a rigorous basis for designing the statistical comparison of modeling results with observations.

7.2 The concept of "natural variability" acknowledges that the details of the stochastic concentration field resulting from dispersion are difficult to predict. In this context, the difference between the ensemble average and any one observed realization (experimental observation) is ascribed to natural variability, whose variation, σ_n^2, can be expressed as:

$$\sigma_n^2 = \overline{(C_o - \overline{C_o})^2} \tag{1}$$

where:

C_o = the observed concentration (or evaluation objective, see 10.3) seen within a realization; the overbars represent averages over all realizations within a given ensemble, so that $\overline{C_o}$ is the estimated ensemble average. The "*o*" subscript indicates an observed value.

7.2.1 The ensemble in Eq 1 refers to the ideal infinite population of all possible realizations meeting the (fixed) characteristics associated with an ensemble. In practice, one will have only a small sample from this ensemble.

7.2.2 Measurement uncertainty in concentration values in most tracer experiments may be a small fraction of the measurement threshold, and when this is true its contribution to σ_n can usually be deemed negligible; however, as discussed in 9.2 and 9.4, expert judgment is needed as the reliability and usefulness of field data will vary depending on the intended uses being made of the data.

7.3 Defining the characteristics of the ensemble in Eq 1 using the model's input values, α, one can view the observed concentrations (or evaluation objective) as:

$$C_o = C_o(\alpha,\beta) = \overline{C_o}(\alpha) + c(\Delta c) + c(\alpha,\beta) \qquad (2)$$

where

β are the variables needed to describe the unresolved transport and dispersion processes, the overbar represents an average over all possible values of β for the specified set of model input parameters α; *c* (Δ*c*) represents the effects of measurement uncertainty, and *c* (α,β) represents ignorance in β (unresolved deterministic processes and stochastic fluctuations) **(14,21)**.

7.3.1 Since $\overline{C_o}(\alpha)$ is an average over all β, it is only a function of α, and in this context, $\overline{C_o}(\alpha)$ represents the ensemble average that the model ideally is attempting to characterize.

7.3.2 The modeled concentrations, C_m, can be envisioned as:

$$C_m = \overline{C_o}(\alpha) + d(\Delta\alpha) + f(\alpha) \qquad (3)$$

where:

d (Δα) represents the effects of uncertainty in specifying the model inputs, and *f* (α) represents the effects of errors in the model formulations. The "*m*" subscript indicates a modeled value.

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

7.3.4 This process is not without problems. The variance in observed concentration values due to natural variability is of order of the magnitude of the regime averages **(17,25)**, hence small sample sizes in the groups will lead to large uncertainties in the estimates of the ensemble averages. The variance in modeled concentration values due to input uncertainty can be quite large **(23,24)**, hence small sample sizes in the groups will lead to large uncertainties in the estimates of the deterministic error in each group. Grouping data together for analysis requires large data sets, of which there are few.

7.3.5 The observations and the modeling results come from different statistical populations, whose means are, for an unbiased model, the same. The variance seen in the observations results from differences in realizations of averages, that which the model is attempting to characterize, plus an additional variance caused by stochastic variations between individual realizations, which is not accounted for in the modeling.

7.3.6 As the averaging time increases in the concentration values and corresponding evaluation objectives, one might expect the respective variances in the observations and the modeling results would increasingly reflect variations in ensemble averages. As averaging time increases, one might expect the variance in the concentration values and corresponding evaluation objectives to decrease; however, as averaging time increases, the magnitude of the concentration values also decreases. As averaging time increases, it is possible that the modeling uncertainties may yet be large when compared to the average modeled concentration values, and likewise, the unexplained variations in the observations yet may be large when compared to the average observed concentration values.

7.4 It is recommended that one goal of a model evaluation should be to assess the model's skill in predicting what it was intended to characterize, namely $\overline{C_o}(\alpha)$, which can be viewed as the systematic (deterministic) variation of the observations from one regime to the next. In such comparisons, there is a basis for believing that a well-formulated model would have zero bias for all regimes. The model with the smallest deviations on average from the regime averages, would be the best performing model. One always has the privilege to test the ability of a model to simulate something it was not intended to provide, such as the ability of a deterministic model to provide an accurate characterization of extreme maximum values, but then one must realize that a well-formulated model may appear to do poorly. If one selects as the best performing model, the model having the least bias and scatter, when compared with observed maxima, this may favor selection of models that systematically overestimate the ensemble average by a compensating bias to underestimate the lateral dispersion. Such a model may provide good comparisons with short-term observed maxima, but it likely will not perform well for estimating maximum impacts for longer averaging times. By assessing performance of a model to simulate something it was not intended to provide, there is a risk of selecting poorly-formed models that may by happenstance perform well on the few experiments available for testing. These are judgement decisions that model users will decide based on the anticipated uses and needs of the moment of the modeling results. This guide has served its purpose, if users better realize the ramifications that arise in testing a model's performance to simulate something that it was not intended to characterize.

8. Statistical Comparison Metrics and Methods

8.1 The preceeding section described a philosophical framework for understanding why observations differ with model simulation results. This section provides definitions of the comparison metrics methods most often employed in current air quality model evaluations. This discussion is not meant to be exhaustive. The list of possible metrics is extensive **(8)**, but it has been illustrated that a few well-chosen simple-to-understand metrics can provide adequate characterization of a model's performance **(14)**. The key is not in how many metrics are used, but is in the statistical design used when the metrics are applied **(13)**.

8.2 *Paired Statistical Comparison Metrics*—In the following equations, O_i is used to represent the observed evaluation objective, and P_i is used to represent the corresponding model's estimate of the evaluation objective, where the evaluation objective, as explained in 10.3, is some feature that can be defined through the analysis of the concentration field. In the equations, the subscript "*i*" refers to paired values and the "overbar" indicates an average.

8.2.1 Average bias, *d*, and standard deviation of the bias, σ_d, are:

$$d = \overline{d_i} \tag{4}$$

$$\sigma_d^2 = \overline{(d_i - d)^2} \tag{5}$$

where:
d_i = $(\mathrm{P_i} - \mathrm{O_i})$.

8.2.2 Fractional bias, FB, and standard deviation of the fractional bias, σ_{FB}, are:

$$FB = \overline{FB_i} \tag{6}$$

$$\sigma_{FB}^2 = \overline{(FB_i - FB)^2} \tag{7}$$

where $FB_i = \dfrac{2(\mathrm{P_i} - \mathrm{O_i})}{(\mathrm{P_i} + \mathrm{O_i})}$.

8.2.3 Absolute fractional bias, *AFB*, and standard deviation of the absolute fractional bias, σ_{AFB}, are:

$$AFB = \overline{AFB_i} \tag{8}$$

$$\sigma_{AFB}^2 = \overline{(AFB_i - AFB)^2} \tag{9}$$

where $AFB_i = \dfrac{2|\mathrm{P_i} - \mathrm{O_i}|}{(\mathrm{P_i} + \mathrm{O_i})}$.

8.2.4 As a measure of gross error resulting from both bias and scatter, the normalized mean squared error, NMSE, often is used (the normalization by $\overline{P}\,\overline{\mathrm{O}}$ assures that NMSE treat equally bias to over-predict or under-predict):

$$NMSE = \frac{\overline{(P_i - O_i)^2}}{\overline{P}\,\overline{O}} \tag{10}$$

8.2.5 For a scatter plot, where the predictions are plotted along the horizontal x-axis and the observations are plotted along the vertical y-axis, the linear regression (method of least squares) slope, m, and intercept, b, between the predicted and observed values are:

$$m = \frac{N\sum P_iO_i - (\sum P_i)(\sum O_i)}{N\sum P_i^2 - (\sum P_i)^2} \tag{11}$$

$$b = \frac{(\sum O_i)(\sum P_i^2) - (\sum P_iO_i)(\sum P_i)}{N\sum P_i^2 - (\sum P_i)^2} \tag{12}$$

8.2.6 As a measure of the linear correlation between the predicted and observed values, the Pearson correlation coefficient often is used:

$$r = \frac{\sum(P_i - \overline{P})(O_i - \overline{O})}{[\sum(P_l - \overline{P})^2 \cdot \sum (O_l - \overline{O})^2]^{1/2}} \tag{13}$$

8.3 *Unpaired Statistical Comparison Metrics*—If the observed and modeled values are sorted from highest to lowest, there are several statistical comparisons that are commonly employed. The focus in such comparisons usually is on whether the maximum observed and modeled concentration values are similar, but one can substitute for the word "concentration," any evaluation objective that can be expressed numerically. As discussed in 7.3.7, the direct comparison of individual observed realizations with modeled ensemble averages is the comparison of two different statistical populations with different sources of variance; hence, there are fundamental philosophical problems with such comparisons. As mentioned in 7.4, such comparisons are going to be made, as this may be how the modeling results will be used. At best, one can hope that such comparisons are made by individuals that are cognizant of the philosophical problems involved.

8.3.1 The quantile-quantile plot is constructed by plotting the ranked concentration values against one another, for example, highest concentration observed versus the highest concentration modeled, etc. If the observed and modeled concentration frequency distributions are similar, then the plotted values will lie along the 1:1 line on the plot. By visual inspection, one can easily see if the respective distributions are similar and whether the observed and modeled concentration maximum values are similar.

8.3.2 Cumulative frequency distribution plots are constructed by plotting the ranked concentration values (highest to lowest) against the plotting position frequency, *f* (typically in percent), where ρ is the rank (1=highest), *N* is the number of values and f is defined as **(26)**:

$$f = 100\,\%\,(\rho - 0.4)/N, \text{ for } \rho < N/2 \tag{14}$$

$$f = 100\,\% - 100\,\%\,(N - \rho + 0.6)/N, \text{ for } \rho > N/2 \tag{15}$$

As with the quantile-quantile plot, a visual inspection of the respective cumulative frequency distribution plots (observed and modeled), usually is sufficient to suggest whether the two distributions are similar, and whether there is a bias in the model to over- or under-estimate the maximum concentration values observed.

8.3.3 The Robust Highest Concentration (RHC) often is used where comparisons are being made of the maximum concentration values and is envisioned as a more robust test statistic than direct comparison of maximum values. The RHC is based on an exponential fit to the highest R-1 values of the cumulative frequency distribution, where R typically is set to be 26 for frequency distributions involving a year's worth of values (averaging times of 24 h or less) **(16)**. The RHC is computed as:

$$RHC = C(R) + \Theta * ln\left(\frac{3R - 1}{2}\right) \tag{16}$$

where:
Θ = average of the *R*-1 largest values, and
C(R) = the R^{th} largest value.

NOTE 1—The value of *R* may be set to a lower value when there are fewer values in the distribution to work with, see **(16)**. The RHC of the observed and modeled cumulative frequency distributions are often compared using a FB metric, and may or may not involve stratification of the values by meteorological condition prior to computation of the RHC values.

8.4 *Bootstrap Resampling*—Bootstrap sampling can be used to generate estimates of the sampling error in the statistical metric computed **(15,16,27)**. The distribution of some statistical metrics, for example, NMSE and RHC, are not necessarily easily transformed to a normal distribution, which is desirable when performing statistical tests to see if there are statistically significant differences in values computed, for example, in the comparison of RHC values computed from the 8760 values of 1-h observed and modeled concentration values for a year.

8.4.1 Following the description provided by **(27)**, suppose one is analyzing a data set $x_1,x_2,...x_n$, which for convenience is denoted by the vector $x=(x_1,x_2,...x_n)$. A bootstrap sample $x^*=(x_1^*,x_2^*,...x_n^*)$ is obtained by randomly sampling n times, with replacement, from the original data points $x=(x_1,x_2,...x_n)$. For instance, with n=7 one might obtain $x^*=(x_5,x_7,x_5,x_4,x_7,x_3,x_1)$. From each bootstrap sample one can compute some statistics (say the median, average, RHC, etc.). By creating a number of bootstrap samples, B, one can compute the mean, $\bar{s}$, and standard deviation, σ_s, of the statistic of interest. For estimation of standard errors, B typically is on the order of 50 to 500.

8.4.2 The bootstrap resampling procedure often can be improved by blocking the data into two or more blocks or sets, with each block containing data having similar characteristics. This prevents the possibility of creating an unrealistic bootstrap sample where all the members are the same value **(15)**.

8.4.3 When performing model performance evaluations, for each hour there is not only the observed concentration values, but also the modeling results from all the models being tested. In such cases, the individual members, x_i, in the vector $x=(x_1,x_2,...x_n)$ are in themselves vectors, composed of the observed value and its associated modeling results (from all models, if there are more than one); thus the selection of the observed concentration x_2 also includes each model's estimate for this case. This is called "concurrent sampling." The purpose of concurrent sampling is to preserve correlations inherent in the data **(16)**. These temporal and spatial correlations affect the statistical properties of the data samples. One of the considerations in devising a bootstrap sampling procedure is to address how best to preserve inherent correlations that might exist within the data.

8.4.4 For assessing differences in model performance, one often wishes to test whether the differences seen in a performance metric computed between Model No. 1 and the observations (say the NMSE_1), is significantly different when compared to that computed for another model (say Model No. 2, NMSE_2) using the same observations. For testing whether the difference between statistical metrics is significant, the following procedure is recommended. Let each bootstrap sample be denoted, x^{*b}, where * indicates this is a bootstrap sample (8.4.1) and b indicates this is sample b of a series of bootstrap samples (where the total number of bootstrap samples is B). From each bootstrap sample, x^{*b}, one computes the respective values for $\text{NMSE}_1{}^b$ and $\text{NMSE}_2{}^b$. The difference $\Delta^{*b} = \text{NMSE}_1{}^{*b} - \text{NMSE}_2{}^{*b}$ then can be computed. Once all B samples have been processed, compute from the set of B values of $\Delta^* = (\Delta^{*1}, \Delta^{*2},...\Delta^{*B})$, the average and standard deviation, $\bar{\Delta}$ and σ_Δ. The null hypothesis is that $\bar{\Delta}$ is greater than zero with a stated level of confidence, η, and the t-value for use in a Student's-t test is:

$$t = \frac{\bar{\Delta}}{\sigma_\Delta} \tag{17}$$

For illustration purposes, assume the level of confidence is 90 % (η = 0.1). Then, for large values of B, if the t-value from Eq 17 is larger than Student's-$t_{\eta/2}$ equal to 1.645, it can be concluded with 90 % confidence that $\bar{\Delta}$ is not equal to zero, and hence, there is a significant difference in the NMSE values for the two models being tested.

9. Considerations in Performing Statistical Evaluations

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

9.2 *Understanding Observed Concentrations*:

9.2.1 It is not necessary for a user of concentration observations to know or understand all details of how the observations were made, but some fundamental understanding of the sampler limitations (operational range), background concentration value(s), and stochastic nature of the atmosphere is necessary for developing effective evaluation procedures.

9.2.2 All samplers have a detection threshold below which observed values either are not provided, or are considered suspect. It is possible that there is a natural background of the tracer, which either has been subtracted from the observations, or needs to be considered in using the observations. Data collected under a quality assurance program following consensus standards are more credible in most settings than data whose quality cannot be objectively documented. Some samplers have a saturation point which limits the maximum value that can be observed. The user of concentration observations should address these, as needed, in designing the evaluation procedures

9.2.3 Atmospheric transport and dispersion processes include stochastic components. The transport downwind follows a serpentine path, being influenced by both random and periodic wind oscillations, composed of both large and small scale eddies in the wind field. Fig. 1 illustrates the observed concentrations seen along a sampling arc at 50-m downwind and centered on a near-surface point-source release of sulfur-dioxide during Project Prairie Grass **(28)**. Fig. 1 is a summary

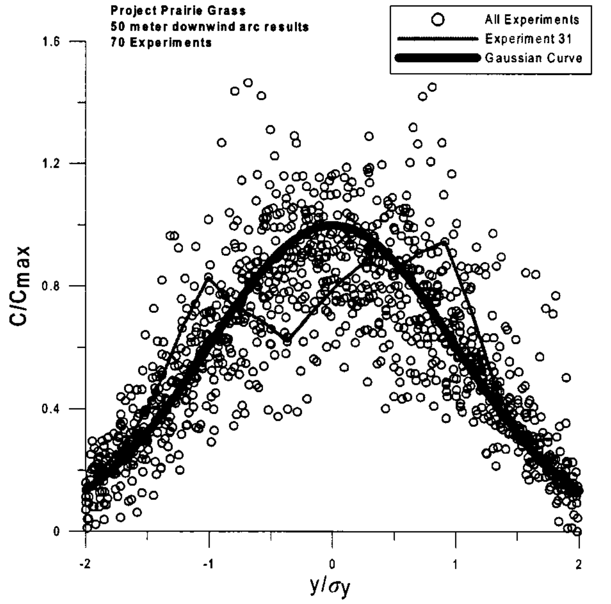

FIG. 1 Illustration of Effects of Natural Variability on Crosswind Profiles of a Plume Dispersing Downwind (Grouped in a Relative Dispersion Context)

over all 70 experiments. For each experiment the crosswind receptor positions, *y*, relative to the observed center of mass along the arc have been divided by σ_y, which is the second-moment of the concentration values seen along each arc, that is, the lateral dispersion which is a measure of the lateral extent of the plume. The observed concentration values have been divided by $C_{max} = C^Y/(\sigma_y\sqrt{2\pi})$, where C^Y is the crosswind integrated concentration along the arc. The crosswind integrated concentration is a measure of the vertical dilution the plume has experienced in traveling to this downwind position. To assume that the crosswind concentration distribution follows a Gaussian curve, which is implicit in the relationship used to compute C_{max}, is seen to be a reasonable approximation when all the experimental results are combined. As shown by the results for Experiment 31, a Gaussian profile may not apply that well for any one realization, where random effects occurred, even though every attempt was made to collect data under nearly ideal circumstances. Under less ideal conditions, as with emissions from a large industrial power plant stack of order 75 m in height and a buoyant plume rise of order 100 m above the stack, it is easy to understand that the observed lateral profile for individual experimental results might well vary from the ideal Gaussian shape. It must be recognised that features like double peaks, saw-tooth patterns and other irregular behavior are often observed for individual realizations.

9.3 *Understanding the Models to be Evaluated*:

9.3.1 As in other branches of meteorology, a complete set of equations for the characterization of the transport and fate of material dispersing through the atmosphere is so complex that no unique analytical solution is known. Approximate analytical principles, such as mass balance, are frequently combined with other concepts to allow study of a particular situation **(29)**. Before evaluating a model, the user must have a sufficient understanding of the basis for the model and its operation to know what it was intended to characterize. The user must know whether the model provides volume average concentration estimates, or whether the model provides average concentration estimates for specific positions above the ground. The user must know whether the characterizations of transport, dispersion, formation and removal processes are expressed using equations that provide ensemble average estimates of concentration values, or whether the equations and relationships used provide stochastic estimates of concentration values. Answers to these and like questions are necessary when attempting to define the evaluation objectives (10.3).

9.3.2 A mass balance model tracks material entering and leaving a particular air volume. Within this conceptual framework, concentrations are increased by emissions that occur within the defined volume and by transport from other adjacent volumes. Similarly, concentrations are decreased by transport exiting the volume, either by removal by chemical/physical sinks within the volume, for example, wet and dry deposition, and for reactive species, or by conversion to other forms. These relationships can be specified through a differential equation quantifying factors related to material gain or loss **(29)**. Models of this type typically provide ensemble volume-average concentration values as a function of time. One will have to consult the model documentation in order to know whether the concentration values reported are averaged over some period of time, such as 1-h, or are the volume-average values at the end of time periods, such as at the end of each hour of simulation.

9.3.3 Some models are entirely empirical. A common example **(30)** involves analysis and characterization of the concentration distributions using measurements under different conditions across a variety of collection sites. Empirical models are strictly-speaking only applicable to the range of measurement conditions upon which they were developed.

9.3.4 Most atmospheric transport and dispersion models involve the combination of theoretical and empirical parameterizations of the physical processes **(31)**, therefore, even though theoretical models may be suitable to a wide range of applications in principle, they are limited to the physical processes characterized, and to the inherent limitations of empirically derived relationships embedded within them.

9.3.5 Generally speaking, as model complexity grows in terms of temporal and spatial detail, the task of supplying appropriate inputs becomes more demanding. It is not a given that increasing the complexity in the treatment of the transport and fate of dispersing material will provide less uncertain predictions. As the number of model input parameters increases, more sources are provided for development of model uncertainty, *d* ($\Delta\alpha$) in Eq 3. Understanding the sensitivity of the modeling results to model input uncertainty should affect the definition of evaluation objectives and associated procedures. For instance, specifying the transport direction of a dispersing plume is highly uncertain. It has been estimated that the uncertainty in characterizing the plume transport is on the order of 25 % of the plume width or more **(17)**. If one attempts to define the relative skill of several models with the modeling results and observations paired in time and space, the uncertainties in positioning a plume relative to the receptor positions will cause there to be no correlation between the model results and observations, when in fact some of the models may be performing well, once uncertainties resulting from plume transport are mitigated **(13,17)**.

9.4 *Choosing Data Sets for Model Evaluation*:

9.4.1 In principle, data used for the evaluation process should be independent of the data used to develop the model. If independent data cannot be found, there are two choices. Either use all available data from a variety of experiments and sites to broadly challenge the models to be evaluated, or collect new data to support the evaluation process. Realistically, the latter approach is only feasible in rare circumstances, given the cost to conduct full-scale comprehensive field studies of atmospheric dispersion.

9.4.2 The following series of steps should be used in choosing data sets for model evaluation: select evaluation field data sets appropriate for the applications for which the model is to be evaluated; note the model input values that require estimation for the selected data sets; determine the required levels of temporal detail, for example, minute-by-minute or hour-by-hour, and spatial detail, for example, vertical or horizontal variation in the meteorological conditions, for the models to be evaluated, as well as the existence and variations of other sources of the same material within the modeling domain; ensure that the samplers are sufficiently close to one another and in sufficient numbers for definition of the evaluation objectives; and, find or collect appropriate data for estimation of the model inputs and for comparison with model outputs.

9.4.3 In principle, the information required for the evaluation process includes not only measured atmospheric concentrations but also measurements of all model inputs. Model inputs typically include: emission release characteristics (physical stack height, stack exit diameter, pollutant exit temperature and velocity, emission rate), mass and size distribution of particulate emissions, upwind and downwind fetch characteristics, for example, land-cover, surface roughness length, daytime and nighttime mixing heights, and surface-layer stability. In practice, since suitable data for all the required model inputs are rarely, if ever, available, one resorts to one or more of the following alternatives: compress the level of temporal and spatial detail for model application to that for which suitable data can be obtained; provide best estimates for model inputs, recognizing the limitations imposed by this particular approach; or, collect the additional data required to enable proper estimation of inputs. A number of assumptions are usually made when modeling even the simplest of situations. These assumptions, and their potential influence on the modeling results, should be identified in the evaluation process.

10. Statistical Procedures and Data Analysis

10.1 *Establishing Evaluation Goals*—Assuming suitable observational data are available, the evaluation goals may be to assess the performance of the model on its ability to characterize what it was intended to characterize or on its ability to characterize something different than it was intended to characterize. There are consequences in choosing the latter, as is mentioned in 7.4. This guide recommends including in the evaluation, an assessment of how well the model performs, when used to characterize quantities it was intended to characterize, namely $\overline{C_o}(\alpha)$ of Eq 3.

10.1.1 When the intent is to test a model on its ability to perform as intended, the evaluation goal for each evaluation objective can be to determine which of several models has the lowest combination of bias and scatter when modeling results are compared with observed values of evaluation objectives defined within the observed and modeled $\overline{C_o}(\alpha)$ patterns. For this assessment, this guide recommends using at least the NMSE (other comparison metrics may also provide useful insights). Define the model having the lowest value for the NMSE as the base-model. Then to assess the relative skill of the other models, the null hypotheses would be that the NMSE values computed for the other models significantly is different when compared to that computed for the base-model (see 8.4.4).

10.1.2 Given that verification of the truth of any model is an impossible task, this guide recommends viewing model performance in relative terms. Testing one model using results from one field experiment provides little insight into its performance. This guide anticipates that models are going to be used for situations for which there is no evaluation data; hence, it is always best to test several models in their ability to performing certain desired tasks best over a variety of circumstances. Then, the task becomes to eliminate those models whose performance is significantly different from the apparent best performing model, given the unexplained variations seen within the observations. As new field data becomes available the apparent best performing model may change, as the models may be tested for new conditions and in new circumstances. This argues for using a variety of field data sets, to provide hope for development of robust conclusions as to which of several models can be deemed to be performing best.

10.2 *Establishing Regimes (Stratification)*—As mentioned in 7.3.3, this guide recommends sorting the available concentration data into regimes, or groups of data having similar model input, α, prior to performing any statistical comparisons. If one chooses to stratify the evaluation data into regimes, this may affect the evaluation objectives, their definition, and the procedures used to compute their values, hence "regimes" will be discussed now, before discussing evaluation objectives and evaluation procedures.

10.2.1 By stratifying the data into regimes, one mitigates the possibility for offsetting biases in the model's performance to compensate. By stratifying the data into regimes and analyzing all the data within a group together, comparisons can be made of the ability of a deterministic model to replicate without bias the regime's characteristics, for example, average" centerline" concentration, average lateral extent, average time a puff takes to pass a particular position, average horizontal extent. If a stochastic model were being evaluated, the evaluation objectives might be the average variance in the "centerline" concentration values, or the average variance in the lateral extent.

10.2.2 The goal in grouping data together is to use such strata as needed to capture the essence of the physics being characterized, such that model performance can be quantified. As discussed in **(32)**, the aim in stratification is to break up the universe into classes, or regimes, that are fundamentally different in respect to the average or level of some quality-characteristic. In theory, the stratification is based on properties of the various regimes that govern the variance of the estimate of the mean or the total variance of the universe **(32)**. A consideration in defining the strata is that there should be a reasonable number of realizations within each stratum, of order five or more **(33)**. The ability to describe model performance as conditions change will argue for many regimes, while the limits of data available for comparison will limit the number of regimes possible.

10.2.3 Specific criteria, as to numbers of cases needed in each regime, or on desired tolerances on how much the model input values, α, can vary for data being grouped together can not be provided at this time. What can be reported is that even rather simplistic sorting of the data by stability and distance has been shown to reveal differences in model performance **(13)**, where with identical evaluation data and modeling results no differences were detected when the data were not sorted **(12)**. In an assessment of modeling multiple point source emissions in an urban area **(34)**, a stratification by Pasquill stability categories was used, which revealed an informative pattern of model bias as a function of stability. An investigation was undertaken with the example evaluation procedures discussed in Appendix X1 to this guide **(33)**. It was found that even with minimal sorting of the data at a specified distance downwind into as few as two stability classes, namely all cases with Zi/L <– 50 and all cases with Zi/L > –50 (Zi is mixing height and L is Monin-Obukhov length), differences in model performance were detectable. These results are admittedly anecdotal, but they provide evidence that overly tight criteria are likely not needed in sorting the data into regimes.

10.2.4 Besides atmospheric stability and transport distance, one could consider other sorting criteria. Time of day may prove to be useful for evaluating model performance when land-sea breeze circulations are present. Cloud amount and presence of precipitation may prove to be useful for evaluating model performance when the fate of the dispersing material is strongly affected by the presence of moisture and cloud processes, such as the fate and transport of sulfates **(35)**.

10.2.5 As discussed in **(32)**, a common misconception regarding stratification is that a particular sample is invalidated because some of the elements were “misclassified.” The real universe is dynamic, and the information that is used for classification is always to some extent uncertain. Moreover, in any real stratification a few blunders may occur; they ought not to, but they do. Misclassification is thus expected as a natural course of events. The point to be made is not that misclassifications occur, but to understand that such occurrences will increase the sampling error, and thus reduce the overall precision in discerning differences in model performance. At very worse, stratification will provide no better discernment than if the universe was left unstratified. It is sometimes proposed that stratification will always bring gains in precision, but this will only occur if the regime averages, or quality-characteristic, either modeled or observed, are indeed different. It is during such circumstances, when modeled or observed regime quality-characteristics differ, that the gains of stratification are great **(32)**.

10.3 *Establishing and Defining Evaluation Objectives*—In order to perform statistical comparisons, this guide recommends defining those evaluation objectives (features or characteristics) within the pattern of observed and modeled concentration values that are of interest to compare. As yet, no one feature or characteristic has been found that can be defined within a concentration pattern that will fully test a model’s performance. For instance, the maximum surface concentration may appear unbiased through a compensation of errors in estimating the lateral extent of the dispersing material and in estimating the vertical extent of the dispersing material. Adding into consideration that other biases that may exist, for example, in treatment of the chemical and removal processes during transport, in estimating buoyant plume rise, in accounting for wind direction changes with height, in accounting for penetration of material into layers above the current mixing depth, in systematic variation in all of these biases as a function of atmospheric stability), one appreciates that there are many ways that a model can falsely give the appearance of good performances.

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

10.3.2 Developing evaluation objectives, involves having a sense of what analysis procedures might be employed. This involves a combination of understanding the modeling assumptions, knowledge of possible comparison measures, and knowledge of the success of previous practices. For example, to assess performance of a model to simulate the pattern of a dispersing puff from a comparison of isolated measurements with the estimated concentration pattern, **(36)** used a procedure developed for measuring the skill of mesoscale meteorological models to forecast the pressure pattern of a tropical cyclone when only isolated pressure measurements are availabe for comparison **(37)**. In particular, the surface area where concentrations were predicted to be above a certain threshold was compared to a surface area deduced from the available monitoring data. The lesson here is that evaluation objectives and procedures developed in other earth sciences can often be adapted for evaluating air dispersion models.

10.3.3 This guide recommends that an evaluation should attempt to include in its comparisons, test of whether the model is performing well when used as intended. This would entail developing evaluation objectives (features or characteristics) from the pattern of observed and modeled $\overline{C_o}(\alpha)$ values for comparison.

10.3.4 For each of the example evaluation objectives listed in 10.2.1, one will have to provide a definition of what is meant by the observed and modeled pattern of $\overline{C_o}(\alpha)$. It will be found that these definitions will have to be specified in terms of the nature and scope of available field data for analysis, and whether one has sorted the data into regimes.

NOTE 2—For instance, if one is testing models on their ability to provide regime average centerline concentration values, then criteria for sorting the data into regimes will be needed. If the model input values are used for the regime criteria, a resolution will be needed for cases when different models have different input values for the same parameter for the same hour. If one is testing models on their ability to reproduce centerline concentration values, a procedure will be needed to determine which of the available observations are representative of centerline concentration values. If one is testing models to produce estimates of the average or the variance in the centerline concentration values, the averaging time to be associated with these estimates will need to be stated.

10.4 *Establishing Evaluation Procedures*—Having selected evaluation objectives for comparison, the next step would be to define an analysis procedure or series of procedures, which define how each evaluation objective will be derived from the available information.

10.4.1 Development of evaluation procedures begins by defining the terminology used in the goal statement. For instance let us suppose that one of the evaluation goals is to test the ability of models to replicate the average centerline concentration as a function of transport downwind and as a function of atmospheric stability. The stated goal involves several items which will require definition, namely average centerline concentration, transport downwind, and stability. The last two may appear innocent, but when viewed in the context of the evaluation data, other terms or problems will surface for resolution. During near-calm wind conditions, when transport may have favored more than one direction over the sampling period, "downwind" is not well described by one direction. If plume models are being tested, one might exclude near-calm conditions, since plume models are not meant to provide meaningful results during such conditions. If puff models or grid models are being tested, one might sort the near-calm cases into a special regime for analysis. For surface releases, surface-layer Monin-Obukhov length (L) has been found to adequately define stability effects, whereas, for elevated releases Zi/L, where Zi is the mixing depth, has been found a useful parameter for describing stability effects. Each model likely has its own meteorological processor. It is a likely circumstance that different processors will have different values for L and Zi for each of the evaluation cases. There is no one best way to deal with this problem. One solution might be to sort the data into regimes using each of the model's input values, and see if the conclusions reached as to best performing model are affected. Given a sampling arc of concentration values, a decision is needed of whether the centerline concentration is the maximum value seen anywhere along the arc, or whether the centerline concentration is that seen near the center of mass of the observed lateral concentration distribution. If one chooses the latter concept, then definition is needed of how near one has to be, to be representative of a centerline concentration value. If one chooses the latter concept, one might decide to select all values within a specific range (nearness to the center of mass). In such a case, either a definition or a procedure will be needed to define how this specific range will be determined. If one is grouping data together for which the emission rates are different, one might choose to resolve this by normalizing the concentration values by dividing by the respective emission rates. To divide by the emission rate requires either a constant emission rate over the entire release, or the downwind transport is sufficiently obvious that one can compute an emission rate based on travel time that is appropriate for each downwind distance. This discussion is not meant to be exhaustive but to be illustrative. It provides an illustration of how the thought process might evolve. It is seen that in defining terms, other questions arise that when resolved eventually will develop an analysis that will compute the evaluation objective from the available data. There may be no one best answer to the questions that develop, and this may cause the evaluation procedures to develop multiple paths to the same goal. If the same set of models is chosen as the best performing models, regardless of which path is chosen, one can likely be assured that the conclusions reached are robust.

10.4.2 Appendix X1 contains an example evaluation procedure for computing the average centerline maximum concentration value from tracer field data. It illustrates an approach that has been tested and shown to be effective, but has yet to reach a consensus of acceptance **(38,39)**. In the example approach in Appendix X1, an example procedure is outlined for definition of centerline concentration values that is robust to the effects of variations in the atmospheric conditions and modeling input (α-variations).

10.4.3 Providing technical definitions of terminology is the basis upon which one defines and develops the evaluation procedures. In some cases, there is no one correct answer to some of the questions that one might pose. What is important is to define what is being evaluated and how terms are to be defined, and it is recommended that these definitions be expressed within the context of the evaluation framework discussed in Section 7. This requires one to understand the consequences of the evaluation framework, within the context of the analyses being conducted.

10.4.4 It is beyond the scope of this guide to address comprehensively how evaluation procedures should be defined. The example procedure in Appendix X1 might be adapted to other evaluation goals, but it is narrowly limited to a particular class of problems. Dispersion models are used in a variety of ways, and the discussion in Appendix X1 is not meant to provide a template from which all evaluation procedures can be developed successfully. As experience is gained in the development and testing of evaluation procedures, more specific guidance may become available.

11. Summarizing Model Performance

11.1 There are two methods commonly used for summarizing model performance. The first involves the comparison of observed and modeled values (evaluation objectives) as a function of one or more model input variables and is used to detect trends in modeling bias. The second is a summary of one or more comparison metrics over all conditions experienced within a particular field experiment and is used to select the set of best performing models for a given application. These comparisons can be done with or without stratification of the evaluation data into regimes. This guide suggests that performing these summary comparisons using the modeled and observed regime averages, improves the likelihood of detecting bias in the models' ability to perform as intended. As noted below, references to observed and modeled values are referring to the observed and model model evaluation objectives.

11.2 *Detecting Trends in Modeling Bias*:

11.2.1 A plot of the observed and modeled values as a function of one of the model input parameters is a direct means for detecting model bias. Such comparison have been recommended and employed in a variety of investigations **(8, 17, 25)**. In some cases the comparison is the ratio, formed by dividing the modeled value by the observed value, plotted as a function of one or more of the model input parameters.

11.2.2 If the data have been stratified into regimes, one also can display the standard error estimates on the respective modeled and observed regime averages. If the respective averages are encompassed by the error bars, typically ± two times the standard error estimates, one can assume the differences are not significant **(38)**. As described by **(15)**, this a seductive inference. A more robust assessment of the significance of the differences would be to use the analysis discussed in 8.4.4, but in practice conclusions reached using such seductive inferences usually are confirmed when the more rigorous procedure is employed.

11.3 *Overall Summary of Performance*:

11.3.1 The design for statistically summarizing model performance over several regimes can be envisioned as a five step procedure. In Step 1, form a replicate sample using concurrent sampling of the observed and modeled values for each regime. In Step 2, compute the average of observed and modeled values for each regime. In Step 3, compute the NMSE using the computed regime averages, and store off the value of the NMSE computed for this pass of the bootstrap sampling. In Steps 4 and 5, repeat steps 1 through 3 for all B bootstrap sampling passes and implement the procedure described in 8.4.4 to detect which model has the lowest computed NMSE value (call this the base model), and which models have NMSE values that are significantly different from the base model.

NOTE 3—As described in Appendix X1 forming the replicate samples involves judgement in order to preserve correlations that might exist within the data, and to insulate the data samples from being unduly affected by grouping data together within a regime that invariably are affected by variations in dispersive conditions (variations in α, model input parameters).

NOTE 4—When developing pooled (summary) statistics across all regimes, steps should be taken to avoid confounding the conclusions reached. No realization should appear in more than one regime, and it is helpful towards avoiding Simpson's Paradox **(40)** to have a similar number of realizations in each regime.

NOTE 5—The NMSE is used in the discussion above, but any of a number of comparison metrics may usefully be substituted, for example, linear regression slope and intercept results, Pearson correlation coefficient, etc. These paired comparison metrics are difficult to interpret and have less value when raw observations are compared to modeling results, as then the comparison is between different populations that are only vaguely correlated. When comparing regime averages, the paired comparison metrics are anticipated to have greater usefulness, because the averaging process filters out, hopefully most, of the observational and modeled fluctuations which are uncorrelated.

11.3.2 The summary procedure described above and used in Appendix X1 was constructed using that introduced by **(16)** as a template. In **(16)**, the data were sorted into regimes (defined in terms of Pasquill stability category and low/high wind speed classes), bootstrap resampling was used to develop standard error estimates on the comparisons, and the comparison metric was the RHC (computed from the raw observed cumulative frequency distribution). This procedure can be improved if the regimes are defined in terms of stability defined by Zi/L and transport distance, and if the comparison metric is the NMSE computed from the modeled and observed regime averages of centerline concentration values. As discussed by **(31)**, Zi/L better captures the boundary-layer stability effects for elevated releases than Pasquill stability category (which is essentially another definition of L). Using downwind distance avoids good comparisons to result from offsetting model bias in the characterizations of the vertical dilution of the material during transport downwind. The RHC was a comparison of the highest concentration values (maxima), which most models do not contain the physics to simulate.

11.3.3 A major limitation with the Appendix X1 procedure is that it can only be applied to intensive tracer field studies, with a dense receptor network, whereas **(16)** attempted to make use of very sparse receptor networks having one or more years of sampling results. With dense receptor networks, attempts can be made to compare modeled and observed centerline concentration values, but there are only a few experiments that have sufficient data to allow stratification of the data into regimes for analysis. With sparse receptor networks, there are more data for analysis, but there is insufficient information to define the observed maxima relative to the dispersing plume's center of mass; thus, it will not be known whether the observed maxima are representative of centerline concentration values. As seen in Fig. 1, observed concentrations near the center of mass can easily vary by a factor of two in magnitude. Also seen in Fig. 1, observed concentrations away from the center of mass, easily can be of the same magnitude as the average centerline concentration value. It is not obvious that the average of the *N* (say 25) observed maximum hourly concentration values (for a particular distance downwind and narrowly defined stability range) is the ensemble average centerline concentration the model is attempting to model; in fact, one might anticipate that it is likely higher than the ensemble average of the centerline concentration. As discussed in 7.4, testing a model in this manner may favor as best performing poorly-formed models that routinely underestimate the lateral dispersion. This in turn, may bias the performance of such models in their ability to correctly characterize concentration patterns for longer averaging times. Given a well-formulated model (which was selected using field data from an intensive network of receptors), evaluations using field data from sparse networks are seen as a useful extension to further explore the performance of a well-formulated model for other environs and for use of the model for other purposes.

12. Conclusions

12.1 The statistical evaluation of model performance with field data is a valuable exercise, but will be limited in scope by the availability of suitable data for comparison; thus, the statistical comparison results are best interpreted within the context of a peer review of the model algorithms and supportive analyses (code verification, sensitivity analyses, and Monte Carlo analyses).

12.2 The development of a design for performing a statistical evaluation begins with the formation of evaluation goals, which will allow construction of null hypotheses, which can be tested. This guide recognizes that one may compose evaluation goals that test the ability of a model to replicate something it was not designed to characterize, as this may be the real-world application. This guide recommends that the evaluation goals include statistical evaluation of the ability of the models to perform as designed. The evaluation goals provide a basis for the definition of specific evaluation objectives and associated procedures for computing the evaluation objectives. This guide recommends that the evaluation objectives and associated procedures be described within the context of the model evaluation framework described in Section 7.

12.3 There are evaluations in the literature that implicitly assume that what is observed is what is being modeled, when this is not the case. This guide is attempting to heighten awareness and thereby promote statistical model evaluations to at least document when the models are being tested on their ability to simulate something they were not intended to provide. It would be worthwhile also for such evaluations to discuss the consequences and possible limitations of such evaluations.

12.4 Given that verification of the truth of any model is an impossible task, this guide recommends viewing a model's performance in relative terms, in comparison to several available models over a variety of circumstances. As new field data becomes available the selection of which of several models is performing best may change, as the models may be tested for new conditions and in new circumstances. This argues for using a variety of field data sets, to provide hope for development of robust conclusions as to which of several models can be deemed to be currently performing best.

12.5 As experience increases in the use of this guide, it is hoped that consensus will be reached in certain evaluation goals, evaluation objectives and associated evaluation procedures, and data sets to be employed. The ultimate goal will be to define a standard test method for each evaluation goal.

13. Keywords

13.1 air quality; model; model performance; qualitative evaluation; quantitative evaluation; statistical

APPENDIXES

(Nonmandatory Information)

X1. EXAMPLE PRACTICE FOR STATISTICAL PERFORMANCE EVALUATION OF THE CHARACTERIZATION OF THE AVERAGE CENTERLINE CONCENTRATION VALUES BY LOCAL-SCALE DISPERSION MODELS FOR DISPERSION FROM AN ISOLATED POINT SOURCE

X1.1 This appendix provides the logic and analysis steps to conduct a statistical performance evaluation for local-scale atmospheric dispersion models. The logic is the same, regardless of whether the evaluation objective is the average of the centerline concentration values, or the average variance of the centerline concentration values, or the average lateral extent of the plume as a function of downwind distance, or any other average feature that can be derived from a collection of observations within a regime.

X1.2 *Evaluation Goal*—To conduct a statistical performance evaluation of the ability of several models to simulate without bias the near-surface average centerline concentration value (1 h or less) for continuous releases from an isolated point source as a function of transport downwind and atmospheric stability.

X1.2.1 *Definition of Centerline Concentration Values*—It is assumed that those receptors that are suitably close (to be defined within the associated procedure, see X1.5) to the observed center of mass along each arc, are close enough to the true plume centerline position to be representative of what would be observed at the plume centerline.

X1.2.2 *Definition of Transport Downwind*—It is assumed that field studies have near-surface receptors placed along arcs (or pseudo-arcs) at prescribed distances downwind and centered on the pollutant (or tracer) release position. It is assumed that the downwind transport of the pollutant (or tracer) is fairly steady, and situations having light and variable winds are not considered in this assessment. These assumptions allow testing of dispersion models that are typically used in routine model assessments. Use of field data with poorly formed sampling arcs or unsteady transport will degrade the usefulness of these evaluation procedures.

X1.2.3 *Definition of Atmospheric Stability*—For near-surface releases, the Monin-Obukhov length (L) has been shown to explain well the stability effects to atmospheric dispersion. For elevated releases, the ratio Zi/L, where Zi is the local mixing height, is anticipated to explain the stability effects to atmospheric dispersion.

X1.2.4 *Definition of Regimes*—The goal is to detect bias in the average value of the centerline concentration. This is interpreted to indicate that the intent is to test the ability of models to simulate without bias the ensemble mean centerline concentration. Pseudo-ensembles (regimes) will be formed by sorting all data seen along each arc into separate stability groups. This will allow comparison of observed and modeled group averages at each distance downwind. Each regime is defined by distance downwind and a range of stability conditions.

X1.3 *Evaluation Procedure for Sorting Data into Regimes*

(Pseudo-Ensembles)—Judgement will be involved in sorting the data by stability, as it will not always be obvious whether L or Zi/L is best, and the ranges in stability will vary depending on the number of arcs of data available at each downwind distance. As described in 10.2.3, exact criteria on defining the range of allowable variation in stability are likely not needed, as past experiences have shown that most any segregation will prove helpful. Numerical studies **(33)** indicate characterization of the centerline concentration average is best achieved by having at least five, preferably seven or more, arcs of data within each stability regime for an arc.

X1.4 *Evaluation Procedures for Combining Data Within a Regime for Analysis*—There are several practical problems to be solved. A decision is needed of how to group data together having different transport directions and different emission rates, but belonging to the same regime. Stochastic variations will make uncertain the definition of the mean transport direction, as well as, the centerline position of the dispersing plume. In the following, a rationale is provided for the processing steps that could be taken to address these problems.

X1.4.1 *Rationale for Grouping Experiments with Different Emission Rates*—For dispersion in which the transport downwind does not involve wind direction reversals or near-calm conditions, the observed concentration is directly proportional to the emission rate. In order to jointly analyze experiments within a regime, or across several regimes, the observed and modeled concentration values will be divided by the emission rate. For a continuous point source, the normalized concentration values then have the units of (mass/volume) (time/mass) = (time/volume). For a puff release, the normalized concentration values then have the units of (mass/volume) (1/mass) = (1/volume).

X1.4.2 *Rationale for Definition of Centerline Position*—The stochastic fluctuations in the concentration values within the near-centerline portion of dispersing plumes have been estimated to have a somewhat log-normal distribution with a standard geometric deviation of order 1.5 to 2 **(33)**. Most of the time, the dispersing plume expands to an angle on the order of 10 to 50° during transport downwind. With such a narrow plume, even a 4° error in estimating the transport direction can cause very large differences between predicted and observed concentrations paired in time and space **(13,17)**. As a result of wind shifts and random fluctuations, the crosswind profile of concentrations observed along a sampling arc for individual realizations likely will not appear Gaussian (see Fig. 1). These departures from the ideal Gaussian shape will result in uncertainty in the definition of the centerline position. To avoid the effects of transport uncertainty, and assuming the observed concentration values have a sampling time of 1-h or less, concentrations taken from different experiments, but for the same regime and for the same experimental site, can be jointly analyzed using the center of mass to provide an approximate position of the centerline. By using the center of mass, computed for each experiment and arc, as a common reference point, data that actually may have been transported in different directions can be grouped together. The assumption being made here is that the rate of dispersion for different transport directions is not anticipated to be different, all other factors being equal.

X1.4.3 *Data Files for Analyses*—The first two steps, X1.4.1 and X1.4.2, can be accomplished and the results stored for use in later analyses. It has been assumed in the discussion that follows in Appendix X2, that two data files have been constructed. The first file contains a listing of results obtained through an analysis of the observed concentration values for each arc of the field study. Each arc has been analyzed to determine the center of mass, and the receptor positions have been expressed relative to the position of the arc's center of mass, either in units of angular degrees or distance. The concentration values are listed and have been divided by the tracer emission rate. Sufficient information is included in this file to allow sorting the arcs of data into specified regimes. The second file contains the modeled centerline concentration values for each arc for each experiment. Note, that the modeling results are not listed at positions where observations were taken, since there is only one value listed for each experiment for each arc, namely the modeled centerline concentration value divided by the emission rate.

X1.5 *Evaluation Procedure for Determining Which Receptors Provide Observations of Centerline Concentration Values*—If the crosswind dispersion were Gaussian, one would expect observed concentration values would be on average, at least within 20 % of the centerline maximum concentration value if $|y_i/Sy|$ were less than or equal to 0.67, where y_i is the distance from the centerline and Sy is the standard deviation of the crosswind concentration distribution for the regime (collection of experiments). Given the anticipated large fluctuations to be seen (substantially greater than ±20 %), a receptor that is near the centerline (center of mass) is assumed to provide a representative sample of a centerline concentration value. With this in mind, near centerline concentrations are defined as coming from those receptors for which $|y_i/Sy| \leq 0.67$. There are results that suggest that due to large variations in the lateral extent of the individual arcs within a regime, the $|y_i/Sy| \leq 0.67$ criteria may not be sufficiently restrictive **(41)**; therefore, a study was conducted to refine the selection criteria by further limiting the selection criteria **(39)**. The results are specific to particular tracer field studies. It was found that a better representation of the centerline concentration values can be obtained by limiting the selection to only the 1 (for sparse sampling along an arc) to 3 (for dense sampling along an arc) receptors that are nearest to the center of mass receptors that still satisfy the criteria of $|y_i/Sy| \leq 0.67$. When saving off those observations selected as being near the centerline position, record which receptors are adjacent to one another, and link the modeled concentrations with the observations selected.

X1.6 *Evaluation Procedure for Sampling of Observed Centerline Concentration Values During Bootstrap Resampling*—The procedure described in X1.5, and discussed in more detail in X2.3, may result in an unequal number of values being selected from each realization (sampling arc of data) within a regime. When developing the bootstrap samples, treat each arc as equally likely. This argues for the same number of values to be selected from each realization during the bootstrap resampling. There also is every likelihood that concentration values from adjacent receptors will be somewhat autocorrelated. Autocorrelation will affect the computed variance. Two methods could be used in the construction of the bootstrap samples. Method 1, when an arc is selected for sampling, at random one value would be selected from all the near-centerline values available. Method 2, when an arc is selected for sampling, at random one pair of adjacent near-centerline values, see X2.6.3, would be selected from all possible pairs of adjacent values. The consequence of computing variances using Method 1 versus Method 2 was tested **(42)**, and it was concluded that autocorrelation effects were present, and were sufficient to warrant consideration; hence, sampling by pairs (Method 2) is recommended.

X2. GUIDANCE ON PROGRAMING LOGIC OF EXAMPLE PRACTICE

X2.1 In Section X2.3 a series of steps is outlined. Steps 1 and 2, implement the procedures defined in X1.5 for selecting concentration values deemed representative of what would be observed near the centerline position of a plume dispersing downwind. Data files as discussed in X1.4 are assumed to have been created for use. Steps 3 through 6, implement the procedures defined in X1.6 for sampling pairs of observations in the bootstrap replicate sampling.

X2.2 *Begin Loop on B Bootstrap Samples (say B = 500)*—See X2.6.1.

X2.3 *Begin Loop on K Regimes*:

X2.3.1 *Step 1*—Compute *Sy*, for this regime. Read in all arcs of observations for this regime, and compute the regime's lateral dispersion, *Sy*, as:

$$S_y^2 = \frac{\sum_{r=1}^{R} \sum_{i=1}^{l_r} y_{ri}^2 \, C_n^o}{\sum_{r=1}^{R} \sum_{i=1}^{l_r} C_n^o} \tag{X2.1}$$

where I_r is the number of observed concentrations on arc *r* of which there are *R* arcs to be processed, *i* represents the sampling position along the sampling arc, and y_{ri} is the distance of sampling position from the center of mass (computed using only the I_r concentration values for this experiment). It has been assumed that the observed concentration values, ${C_{ri}}^o$, have been divided by the emission rate appropriate for each experiment.

NOTE X2.1—The preceding actions can be completed and the results stored off for the first bootstrap sample. For all succeeding bootstrap samples, the results obtained and previously stored are used. It is anticipated in the following discussion, that a regime contains arcs of data that are at the same distance (or nearly so) downwind from the release.

X2.3.2 *Step 2*—Identify the set of observed centerline concentrations. Treat each arc separately. Store for later analysis, the *N*filter values of ${C_{ri}}^o$ that have the smallest values for $|y_{ri}|$ for which $|y_{ri}/S_y| \leq 0.67$, where *N*filter is an integer, likely in the range from 1 to 3, that will be defined for each field study using judgment, see analyses discussed in **(39)**. As values are stored for later use, it is important also to retain the relative position of the receptors along the arc, as it will be needed to know which values are adjacent to one another.

NOTE X2.2—For the purposes of later discussion, let s_r represent the number of values selected from arc r (assumed to be representative of an observed centerline concentration value). Then the total number of observed values selected as being representative of centerline concentrations from all arcs in this regime, *N*, equals $\sum_{r=1}^{R} S_r$.

NOTE X2.3—In the sampling discussion presented, the method with sample pairs of values were employed because studies **(42)** suggest that there is some positive autocorrelation between adjacent observed near centerline concentration values. When an experiment is selected for sampling, values are selected not only from the observations, but also from all models, as well. Sampling in this fashion, called concurrent sampling, attempts to address correlation effects that might exist between the modeling results and the experimental conditions present during a selected experiment.

X2.3.3 *Step 3*—Define number of pairs to sample. Draw *N*'/2 pairs of samples, with replacement, where *N*' = 2**INT* (*N*/2), where *INT* represents the truncated integer result of dividing *N* by 2.

X2.3.4 *Step 4*—Each arc is considered equally likely. Select one of the arcs at random. It is necessary to sample these same arcs in developing the samples of modeled centerline concentrations for each model, see discussion in X2.6.2. One method to accomplish this would be to store off a pair of modeled centerline concentration values for each model, just before or after the pair of observed values (see Step 5 below) is selected.

X2.3.5 *Step 5*—Select a pair. If $S_r > 1$, then there is more than one observed concentration value ${C_{rj}}^o$ that was deemed to meet the criteria for providing concentrations that are representative of centerline values. At random select a pair of values that are adjacent to one another. If $S_r = 1$, then there is only one value ${C_{rj}}^o$ to sample and the pair is defined to be this one value, selected twice. For further discussion of sampling of pairs see X2.6.3.

X2.3.6 *Step 6*—Repeat Steps 4 and 5 above *N*'/2 times.

NOTE X2.4—In 8.4.2, mention was made that blocking can be used to avoid the development of unrealistic bootstrap samples, where all the values are identical. The addition of blocking in Steps 4 and 5 is a possible future enhancement that could be made to this procedure.

X2.3.7 *Step 7*—At the completion of Step 6, a bootstrap sample of observed and modeled centerline concentration values for this regime will have been constructed.

X2.3.7.1 Compute the average ${C_b}^o$ of the *N*' samples of observed concentration values, where b represents the bootstrap sample index and the *o* superscript denotes that the average was derived using observed concentration values.

X2.3.7.2 Loop on all models and compute the average ${C_b}^m$ of the *N*' samples of modeled concentration values for each model.

X2.4 *End Loop on Regimes*:

X2.4.1 At the completion of generating and analyzing all K regimes, there is available for analysis K values of $C_b{}^o$ and K values for each model of $C_b{}^m$. There are any number of paired statistical analyses that one could insert at this step. For illustration of the logic and completion of the evaluation goal being discussed, compute the NMSE for each model with the K regime averages of observed and modeled centerline concentration values. For notation purposes let NMSE^{bm}, represent the normalized mean square error for bootstrap sample b and model m.

X2.5 *End Loop on Bootstrap Sampling*—At the completion of generating and analyzing all B bootstrap samples, there will be available for analysis B values of the NMSE for each model.

X2.6 *Notes on Considerations in Performance Evaluation Procedure*:

X2.6.1 *Determination of Bootstrap Sample Size*—Once the data to be used in each regime have been defined, one can conduct the following numerical experiment to help answer the question of how many bootstrap samples, B, are needed **(42)**. Run the analysis outlined in Steps 1 through 7, ten times, with ten bootstrap samples for each regime. From each run, we have an estimate for each regime for the standard deviation of the observed concentration values. Using the ten estimates for the standard deviation of the observed concentration values, for each regime compute the average of these ten standard deviations, Avg(std), and the variance of these ten standard deviations, Var(std). The ratio $\sqrt{Var(std)}$/Avg(std) is expected to be proportional to 1 / $\sqrt{B}$, where B is the bootstrap sample size. There will be a different ratio for each regime. One can solve for the value of B such that the ratio $\sqrt{Var(std)}$/Avg(std) is less than some desired tolerance, say 0.05, for all regimes. The sample size B needed, typically varies from one regime to the next. In a previous study **(42)**, it was found that if B was at least 500, the criterion was met in all regimes.

X2.6.2 *Concurrent Sampling*—It is assumed that the replicate sampling conducted in the "loop on samples" is such that the exact same sequence of cases will be chosen for every model. For instance, if sample one contains five cases, and those selected are cases numbered 1, 2, 4, 5 and 1. Then, every model for sample one will consist of cases numbered 1, 2, 4, 5 and 1. This preserves any model to model correlations in the differences to be seen between the modeling results and the observations. Ideally, the observations and modeling results would all be sampled at the same time, but this may not be feasible as it maximizes the computer memory requirements. One may be able to take advantage of the fact that some random number generators will generate the same sequence of random numbers given the same seed. This allows one to sample the observations and the modeling results from each model separately, taking care to always start with the same seed. This may result in a substantial saving of memory requirements for the storage of data arrays. Another solution would be to list the case numbers selected, when sampling the observations, to a side file and reuse these results in developing the samples for each model.

X2.6.3 *Determination of a Random Pair of Values*—Assume that four values from one arc are found to satisfy the criteria of being sufficiently close to the center of mass that they will be considered as representative of centerline concentration values. Number these four values as 1, 2, 3, and 4. Value 1 is adjacent to only value 2, whereas, value 2 is adjacent to both value 1 and value 3. There are three pairs possible: (1,2), (2,3), (3,4). Pair (1,2) is considered to be the same as pair (2,1). In this example, there are three pairs possible, and in the replicate sampling each of the three pairs would be considered equally likely.

X3. GUIDANCE ON PROGRAMMING LOGIC FOR DETERMINATION OF BEST PERFORMING MODELS

X3.1 *Determine Base Case Model*:

X3.1.1 The base case model is defined as that model, which typically exhibits the smallest deviations (bias plus scatter) from the average of the observed values. Loop on all models and compute the average NMSE^m for each model m from the B values of NMSE^{bm}.

NOTE X3.1—The procedure for determining the best performing models is outlined in 8.4.4. It is worth mentioning here that although one might show that model-to-model differences are statistically significant, such differences are typically of little practical importance. What is of real concern is whether the model-to-observation differences, from one model comparison metric to the next, are significantly different.

X3.1.2 The model with the lowest value for NMSE is defined as the base case model. Let s denote this model's number in the sequence of M models, that is NMSE^s.

X3.2 *Determine Set of Best Performing Models*:

X3.2.1 Loop on all M models, excluding model s.

X3.2.1.1 Compute from the B values of NMSE^{bs} and NMSE^{bm}, the B differences $\Delta_b{}^{sm} = \text{NMSE}^{bs} - \text{NSME}^{bm}$.

X3.2.1.2 Compute the mean $\overline{\Delta^{sm}}$ and the standard deviation $\sigma_\Delta{}^{sm}$ from the B values of $\Delta_b{}^{sm}$.

X3.2.1.3 Compute the t-statistic for comparison with $t_{\eta/2}$ as, $t^{sm} = \overline{\Delta^{\text{sm}}}/\sigma_\Delta^{\text{sm}}$.

X3.2.1.4 Reject any model for which t^{sm} is greater than or equal to $t_{\eta/2}$. The null Hypothesis will be that the difference, NMSE^s– NMSE^m, is not statistically significant for a stated level of significance η. For illustration purposes say $\eta = 0.1$, then the Student-t for use in testing the null Hypothesis (assume large number of values) is $t_{\eta/2} = 1.645$.

X3.2.1.5 End loop on models.

X3.3 The set of best performing models is composed of all those not rejected (above) plus the base case model.

REFERENCES

(**1**) U.S. Environmental Protection Agency,“ Guiding Principles for Monte Carlo Analysis,” EPA/630/R-97/001, Office of Research and Development, Washington DC, 1997, 40 pp.

(**2**) Pasquill, F., “The Estimation of the Dispersion of Windborne Material,” *Meteorological Magazine*, Vol 90, 1961, pp. 33–49.

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

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

(**5**) Clarke, J.F.,“ A Simple Diffusion Model for Calculating Point Concentrations from Multiple Sources,” *Journal of the Air Pollution Control Association*, Vol 14, No. 9, 1964, pp. 347–352.

(**6**) Martin, D.O., “An Urban Diffusion Model for Estimating Long Term Average Values of Air Quality,” *Journal of the Air Pollution Control Association*, Vol 21, No. 1, 1971, pp. 16–19.

(**7**) Hanna, S.R., “A Simple Method of Calculating Dispersion from Urban Area Sources, *Journal of the Air Pollution Control Association*, Vol 21, No. 22, 1971, pp. 774–777.

(**8**) Fox, D.G., “Judging Air Quality Model Performance,” *Bulletin of the American Meteorological Society*, Vol 62, 1981, pp. 599–609.

(**9**) Willmot, C.J., “Some Comments on the Evaluation of Model Performance,” *Bulletin of the American Meteorological Society*, Vol 63, 1982, pp. 1309–1313.

(**10**) Venkatram, A., “A Framework for Evaluating Air Quality Models,” *Boundary-Layer Meteorology*, Vol 24, 1982, pp. 371–385.

(**11**) Fox, D.G., “Uncertainty in Air Quality Modeling,” *Bulletin of the American Meteorological Society*, Vol 65, 1984, pp. 27–36.

(**12**) Smith, M.E., “Review of the Attributes and Performance of 10 Rural Diffusion Models,” *Bulletin of the American Meteorological Society*, Vol 65, 1984, pp. 554–558.

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

(**14**) Hanna, S.R.,“ Air Quality Model Evaluation and Uncertainty,” *Journal of the Air Pollution Control Association*, Vol 38, No. 4, 1988, pp. 406–412.

(**15**) Hanna, S.R., “Confidence Limits for Air Quality Model Evaluations, as Estimated by Bootstrap and Jackknife Resampling Methods,” *Atmospheric Environment*, Vol 23, No. 6, 1989, pp. 1385–1398.

(**16**) Cox, W.M. and Tikvart, J.A., “A Statistical Procedure for Determining the Best Performing Air Quality Simulation Model,” *Atmospheric Environment*, Vol 24A, 1990, pp. 2387–2395.

(**17**) Weil, J.C., Sykes, R.I., and Venkatram, A., “Evaluating Air-Quality Models: Review and Outlook,” *Journal of Applied Meteorology*, Vol 31, 1992, pp. 1121–1145.

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

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

(**20**) Oreskes, N., Shrader-Frechette, K., and Belitz, K., “Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences,” *Science*, Vol 263, 1994, pp. 641–646.

(**21**) Venkatram, A., “Topics in Applied Modeling,” In *Lectures on Air Pollution Modeling*, A. Venkatram and J.C. Wyngaard (editors). American Meteorological Society, Boston, MA, 1988, pp. 267–324.

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

(**23**) Irwin, J.S., Rao, S.T., Petersen, W.B., and Turner, D.B., “Relating error bounds for maximum concentration predictions to diffusion meteorology uncertainty,” *Atmospheric Environment*, Vol 21, 1987, pp. 1927–1937.

(**24**) Hanna, S.R., Chang, J.C., and Fernau, M.E., “Monte Carlo Estimates of Uncertainties in Predictions by a Photochemical Grid Model (UAM-IV) Due to Uncertainties in Input Variables,” *Atmospheric Environment*, Vol 32, 1998, pp. 3617–3628.

(**25**) Hanna, S.R., “Uncertainties in Air Quality Model Predictions,” *Boundary Layer Meteorology*, Vol 62, 1993, pp. 3–20.

(**26**) Larsen, R.I., “A New Mathematical Model of Air Pollution Concentration Averaging Time and Frequency,” *Journal of the Air Pollution Control Association*, Vol 19, 1969, pp. 24–30.

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

(**28**) Barad, M.L. (editor), “Project Prairie Grass, A Field Program in Diffusion,” Geophysical Research Paper, No. 59, Vol I and II, Report AFCRC-TR-58-235, Air Force Cambridge Research Center, 1954, 439 pp.

(**29**) Stull, R.B., “*An Introduction to Boundary Layer Meteorology*,” Kluwer Academic Publishers, Dordrecht, the Netherlands, 1988, pp. 75–114.

(**30**) Benarie, M.M.,“ *Urban Air Pollution Modeling Without Computers*,” EPA-600/4-76-055, U.S. Environmental Protection Agency, Research Triangle Park, NC, 1976, pp. 83.

(**31**) Gryning, S.E., Holtslag, A.A.M., Irwin, J.S., and Sivertsen, B., “Applied Dispersion Modeling Based on Meteorological Scaling Parameters,” *Atmospheric Environment*, Vol 21, 1987, pp. 79–89.

(**32**) Deming, W.E.,“ *Some Theory of Sampling*,” Dover Publications, Inc. New York, NY, 1966, 602 pp.

(**33**) Irwin, J.S., “Effects of Concentration Fluctuations on Statistical Evaluations of Centerline Concentration Estimates by Atmospheric Dispersion Models,” Sixth International Conference on Harmonisation with Atmospheric Dispersion Modelling for Regulatory Purposes, October 11–14, 1999, Rouen, France (Preprints, to be published in *International Journal of Environment and Pollution*), 1999, 8 pp.

(**34**) Turner, D.B. and Irwin, J.S., “The Relation of Urban Model Performance to Stability,” in: *Air Pollution Modeling and Its Application IV*, C. De Wispelaere, ed. Plenum Pub. Corp., New York, 1985, pp. 721–732.

(**35**) Dennis, R.L., McHenry, J.N., Barchet, W.R., Binkowski, F.S., and Byun, D.W., “Correcting RADM’s sulfate underprediction: discovery and correction of model errors and testing the corrections though comparisons against field data,” *Atmospheric Environment*, Vol 27A, 1993, pp. 975–997.

(**36**) Brost, R.A., Haagenson, P.L., and Kuo, Y-H., “The effect of diffusion on tracer puffs simulated by a regional scale eulerian model,” *Journal of Geophysical Research*, Vol 93, No. D3, 1988, pp. 2389–2404.

(**37**) Anthes, R.A., “Review—regional models of the atmosphere in middle latitudes,” *Monthly Weather Review*, Vol 111, 1983, pp. 1306–1335.

(**38**) Irwin, J.S., “Statistical evaluation of atmospheric dispersion models,” Fifth International Conference on Harmonisation with Atmospheric Dispersion Modelling for Regulatory Purposes, May 18–21, 1998, Rhodes, Greece (Preprints pp. 46–53, to be published in *International Journal of Environment and Pollution)*, 1998, 8 pp.

(**39**) Irwin, J.S., “Comparison of Two Sampling Procedures for the Statistical Evaluation of the Performance of Atmospheric Dispersion Models in Estimating Centerline Concentration Values,” Proceedings of the Millennium NATO/CCMS International Technical Meeting and its Application, May 15–19, 2000, Boulder, CO, 2000, 8 pp.

(**40**) Mittal, Y., "Homogeneity of Subpopulations and Simpson's Paradox," *Journal of the American Statistical Society*, Vol 86(413), 1991, pp. 167–171.

(**41**) Olesen, H.R., "Model Validation Kit—Status and Outlook," Fifth International Conference on Harmonisation with Atmospheric Dispersion Modelling for Regulatory Purposes, May 18–21, 1998, Rhodes, Greece (Preprints pp. 63–70, to be published in *International Journal of Environment and Pollution*), 1998, 8 pp.

(**42**) Irwin, J.S., and Rosu, R., "Comments on a draft practice for statistical evaluation of atmospheric dispersion models," 10th Joint Conference on the Applications of Air Pollution Meteorology with the A&WMA, January 11–16, 1998, *American Meteorology Society*, 1998, pp. 6–10.

ASTM International takes no position respecting the validity of any patent rights asserted in connection with any item mentioned in this standard. Users of this standard are expressly advised that determination of the validity of any such patent rights, and the risk of infringement of such rights, are entirely their own responsibility.

This standard is subject to revision at any time by the responsible technical committee and must be reviewed every five years and if not revised, either reapproved or withdrawn. Your comments are invited either for revision of this standard or for additional standards and should be addressed to ASTM International Headquarters. Your comments will receive careful consideration at a meeting of the responsible technical committee, which you may attend. If you feel that your comments have not received a fair hearing you should make your views known to the ASTM Committee on Standards, at the address shown below.