Designation: ~~D6589 – 05 (Reapproved 2015)~~ D6589 – 23

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

This standard is issued under the fixed designation D6589; 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 applies to gaussian plume models; it may not be applicable to non-point sources, heavy gas models from evaporation from pool (for example, liquid spills), as well as near-field receptors.

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

Current edition approved ~~April 1, 2015~~Sept. 1, 2023. Published ~~April 2015~~September 2023. Originally approved in 2000. Last previous edition approved in ~~2010~~2015 as D6589 – 05 ~~(2010)~~(2015).~~$^{\varepsilon 1}$.~~ DOI: ~~10.1520/D6589-05R15.~~10.1520/D6589-23.

1.6 The values stated in SI units are to be regarded as standard. No other units of measurement are included in this guide.

1.7 *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, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.8 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

D1356 Terminology Relating to Sampling and Analysis of Atmospheres

3. Terminology

3.1 *Definitions*—For definitions of terms used in this guide, refer to Terminology D1356.

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 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]

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

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

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 procedures 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 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 ensure 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 (**21**). Each input should be varied over a reasonable range likely to be encountered. The traditional sensitivity studies (**21**) 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 (**22**), especially for models simulating chemically reactive species where there are strong nonlinear couplings between the model input and output (**23**). 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:

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

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:

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

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; $c(\Delta c)$ represents the effects of measurement uncertainty, and $c(\alpha,\beta)$ represents ignorance in β (unresolved deterministic processes and stochastic fluctuations) (**14, 24**).

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:

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

where:

$d(\Delta\alpha)$ represents the effects of uncertainty in specifying the model inputs, and $f(\alpha)$ 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 (**22, 23**), 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 judgment 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 preceding 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 $FB_i = \dfrac{2(P_i - O_i)}{(P_i + O_i)}$.

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 $AFB_i = \frac{2|P_i - O_i|}{(P_i + O_i)}$

8.2.4 As a measure of gross error resulting from both bias and scatter, the root mean squared error, RMSE, is often used:

$$RMSE = \sqrt{\overline{(P_i - O_i)^2}} \tag{10}$$

8.2.5 Another measure of gross error resulting from both bias and scatter, the normalized mean squared error, NMSE, often is used:

$$NMSE = \frac{\overline{(P_i - O_i)^2}}{\bar{P}\,\bar{O}} \tag{11}$$

The advantage of the NMSE over the RMSE is that the normalization allows comparisons between experiments with vastly different average values. The disadvantage of the NMSE versus RMSE is that uncertainty in the observation of low concentration values will make the value of the NMSE so uncertain that meaningful conclusions may be precluded from being reached.

8.2.6 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:

$$m = \frac{N\sum P_i O_i - \left(\sum P_i\right)\left(\sum O_i\right)}{N\sum P_i^2 - \left(\sum P_i\right)^2} \tag{12}$$

$$b = \frac{\left(\sum O_i\right)\left(\sum P_i^2\right) - \left(\sum P_i O_i\right)\left(\sum P_i\right)}{N\sum P_i^2 - \left(\sum P_i\right)^2} \tag{13}$$

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

$$r = \frac{\sum\left(P_i - \bar{P}\right)\left(O_i - \bar{O}\right)}{\left[\sum\left(P_l - \bar{P}\right)^2 \cdot \sum\left(O_l - \bar{O}\right)^2\right]^{1/2}} \tag{14}$$

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 7.3.5, 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)**:

$$f = 100\,\%\,(\rho - 0.4)/N, \text{ for } \rho < N/2 \tag{15}$$

$$f = 100\,\% - 100\,\%\,(N - \rho + 0.6)/N, \text{ for } \rho > N/2 \tag{16}$$

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:

$$RHC = C(R) + \Theta{*}ln\left(\frac{3R-1}{2}\right) \tag{17}$$

where:

Θ = average of the *R*-1 largest values minus C(R), 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, RMSE 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~~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 $RMSE_1$), is significantly different when compared to that computed for another model (say Model No. 2, $RMSE_2$) 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 $RMSE_1^{\ b}$ and $RMSE_2^{\ b}$. The difference $\Delta^{*b} = RMSE_1^{*b} - RMSE_2^{*b}$ 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:

$$t = \frac{\bar{\Delta}}{\sigma_\Delta} \tag{18}$$

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 19 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 RMSE 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~~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 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 $C_{\max} = C^Y/\left(\sigma_y\sqrt{2\pi}\right)$, 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

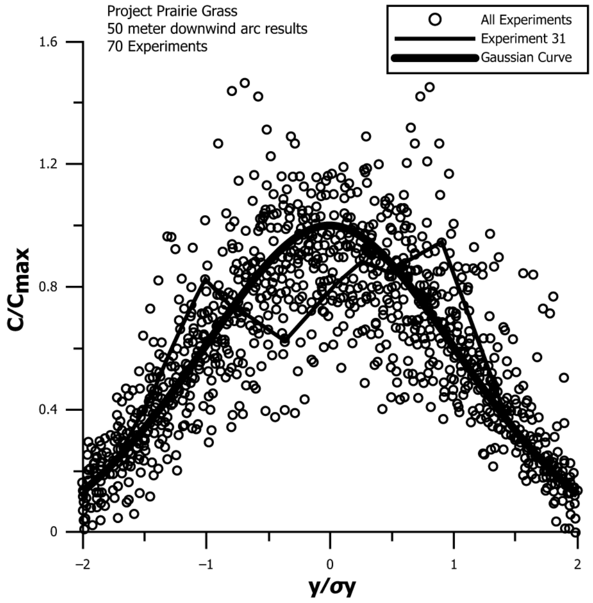

FIG. 1 Illustration of Effects of Natural Variability on Crosswind Profiles of a Plume Dispersing Downwind (Grouped in a Relative Dispersion Context)

lateral profile for individual experimental results might well vary from the ideal Gaussian shape. It must be recognized 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,~~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 2. 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 2.

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 RMSE (other comparison metrics may also provide useful insights). Define the model having the lowest value for the RMSE as the base-model. Then to assess the relative skill of the other models, the null hypotheses would be that the RMSE 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.