Evaluating Air-Quality Models: Review and Outlook*

J. C. Weil†

National Center for Atmospheric Research,* Boulder, Colorado

R. I. Sykes

ARAP Group, California Research and Technology Division, Titan Corporation, Princeton, New Jersey

A. Venkatram

ENSR Corporation, Camarillo, California

(Manuscript received 4 December 1990, in final form 30 December 1991)

ABSTRACT

Over the past decade, much attention has been devoted to the evaluation of air-quality models with emphasis on model performance in predicting the high concentrations that are important in air-quality regulations. This paper stems from our belief that this practice needs to be expanded to 1) evaluate model physics and 2) deal with the large natural or stochastic variability in concentration. The variability is represented by the root-mean-square fluctuating concentration (σₐ) about the mean concentration (C) over an ensemble—a given set of meteorological, source, etc. conditions. Most air-quality models used in applications predict C, whereas observations are individual realizations drawn from an ensemble. For σₐ ∼ C, large residuals exist between predicted and observed concentrations, which confuse model evaluations.

This paper addresses ways of evaluating model physics in light of the large σₐ; the focus is on elevated point-source models. Evaluation of model physics requires the separation of the mean model error—the difference between the predicted and observed C—from the natural variability. A residual analysis is shown to be an effective way of doing this. Several examples demonstrate the usefulness of residuals as well as correlation analyses and laboratory data in judging model physics.

In general, σₐ models and predictions of the probability distribution of the fluctuating concentration (c'), ω(c'), are in the developmental stage, with laboratory data playing an important role. Laboratory data from point-source plumes in a convection tank show that ω(c') approximates a self-similar distribution along the plume center plane, a useful result in a residual analysis. At present, there is one model—ARAP—that predicts C, σₐ, and ω(c) for point-source plumes. This model is more computationally demanding than other dispersion models (for C only) and must be demonstrated as a practical tool. However, it predicts an important quantity for applications—the uncertainty in the very high and infrequent concentrations. The uncertainty is large and is needed in evaluating operational performance and in predicting the attainment of air-quality standards.

1. Introduction

Air-quality models are important tools in air-pollution control and regulatory decision making. They are used to determine the sites of new pollution sources, source emission limits, and air-pollution control strategies. Given this and the associated control costs, it is important to know the accuracy and limitations of models. Model evaluation generally has proceeded by addressing the following questions:

1) How well does the model predict the high ground-level concentrations (GLCs) that are necessary in assessing compliance with air-quality regulations? This is termed operational performance evaluation.

2) Is the model based on sound physical principles and does it give good predictions for the “right” reasons? This is termed model physics evaluation.

3) How can one distinguish the performance of two or more models?

This paper is aimed at addressing these questions. In some applications, one also needs to know how well a

© 1992 American Meteorological Society
model predicts the entire range or probability distribution of concentrations.

Air-quality managers are mostly concerned with operational performance and generally consider model physics to be a secondary point. However, there is an important practical reason for good physics—to have faith in model predictions beyond the range of existing experimental data and confidence in modeling new situations with different dispersion climatologies (section 5d).

Under a cooperative agreement with the Environmental Protection Agency (EPA) and administered through a small steering committee, the American Meteorological Society (AMS) conducted a peer review of ten Gaussian plume models for tall stacks (Smith 1984). The evaluation relied on performance measures recommended by the 1980 AMS Woods Hole workshop (Fox 1981). As Smith reported, the evaluation showed no clear superiority of one model over another, and the volume of statistics was so large as to confuse the intended purpose of judging models. Similar conclusions were reached in reviews of urban dispersion models (White 1984) and complex terrain models (White 1985).

The difficulty in distinguishing performance could have been attributed to several causes (Smith 1984): 1) the models were all of a similar content, being generally out of touch with current knowledge of turbulence and dispersion, so that one would not expect them to perform differently, 2) the database was inadequate, and 3) the comparisons were plagued by the "natural" variability in concentration, which arises from the stochastic nature of dispersion. Air-quality models predict the mean concentration for a given set of conditions (i.e., an ensemble), whereas observations are individual realizations drawn from the ensemble. The natural variability is the random concentration fluctuation about the mean and is large (of the order of the mean; section 2). The steering committee considered the natural variability to be very significant in hampering the performance evaluation. In addition, there are three other sources of concentration variance that must be considered: 1) errors in the model physics, that is, in the mathematical idealization of dispersion, 2) uncertainties or errors in the model input variables, and 3) errors in concentration measurements (Fox 1984; Venkatram 1982). These sources are discussed further in section 4.

While EPA and AMS are to be commended for conducting these reviews, there has been dissatisfaction with these performance evaluations both within and outside the steering committee. It was felt that the Woods Hole recommendations were followed too literally and without trial and error, and that there was too much emphasis on operational performance. Peer reviewers and steering committee members recommended more testing of the model physics and a reduction in the list of performance measures to a manageable number that would highlight the performance (Smith 1984). The EPA has made progress on the second recommendation; see Cox and Tikhvart (1990) and the discussion in section 3b.

In response to these concerns, the AMS steering committee organized a review of model evaluation practices including those just mentioned as well as studies conducted outside the EPA. This paper is the result of that review. Although we initially considered a range of problems, we narrowed our focus to elevated point-source models because 1) they are widely used in air-quality applications, 2) a significant amount of information exists on evaluating these models with field observations (e.g., Irwin and Smith 1984; Smith 1984; Sykes 1988; Venkatram 1988), and 3) controlled laboratory experiments have been conducted to investigate the mean and fluctuating concentration fields in point-source plumes. In addition, we restricted attention to dispersion over ~1-h averaging times and to distances less than about 30 km. Our main purpose was to review and recommend ways of evaluating models in the presence of a large random variability in concentration, that is, to distinguish the "model error" from this variability. This has been aimed at a model physics evaluation (sections 4 and 5).

It is hoped that the issues and approaches discussed in this paper will have application to model evaluations for other problems that have recently received attention. For example, evaluations have been conducted for 1) regional-scale dispersion and acid deposition models (Clark et al. 1989), 2) regional- and global-scale models with application to the Chernobyl accident (Hass et al. 1990), 3) long-range trajectory models for accidental releases (Draxler 1991; Haagenson et al. 1990), and 4) dense and hazardous gas models (Hanna et al. 1991).

The outline of the paper is as follows. We begin with a background discussion of two topics: 1) the natural variability in concentration (section 2) and 2) the current approaches for evaluating the operational performance of models (section 3). These sections are included primarily for review and as introductory material for new practitioners of dispersion modeling; readers familiar with these topics may wish to skip to section 4. In section 4, we extend Fox's (1984) analysis for separating the model error or bias from the natural variability and other sources of concentration variance; the analysis is intended for application to a model physics evaluation. We then apply the basic concepts of the analysis to point-source models, discuss the evaluation of mean and fluctuating concentration models, and consider the relevance of this to regulatory applications (section 5).

2. The natural variability in concentration

a. Stochastic nature of dispersion

Dispersion in the planetary boundary layer (PBL) is controlled by the mean and turbulence structure in
that layer. Like turbulence, dispersion is a random or stochastic phenomenon. This means that the concentration observed at a particular time and location downwind of a pollution source cannot be predicted in general with precision (Chatwin 1982). Concentration is a random variable and should be described statistically through a probability distribution (Chatwin 1982; Csanady 1973). In principle, one could find the distribution at some location by repeating the same experiment many times to generate an ensemble of concentrations. The ensemble is defined by a set of meteorological conditions that control turbulence and dispersion—the mean wind speed, surface heat and momentum fluxes, PBL depth, etc. (Venkatram and Weynagard 1988)—and thus is determined by the physics of the problem; these conditions should be fixed in every realization of the experiment.

In practice, an ensemble of dispersion experiments is difficult to conduct because the set of meteorological conditions that define it do not repeat with sufficient frequency; moreover, they are expensive. Such experiments are more feasible to do in the laboratory, and some were recently conducted to simulate the dispersion of point-source plumes in the convective boundary layer (CBL) (Deardoff and Willis 1988). They show that the probability distribution of concentration can be described by a shape function (e.g., a gamma distribution), the ensemble-mean concentration C, and the root-mean-square (rms) concentration \( \sigma_c \), which characterizes the width of the distribution.

In this paper, we call \( \sigma_c \) the natural or stochastic variability in concentration:

\[
\sigma_c = \left\langle \left( c(x, t) - C(x, t) \right)^2 \right\rangle^{1/2}
\]

where the concentration \( c(x, t) \) in any realization of an experiment and \( C(x, t) \) are time-averaged quantities in general, \( x \) (a vector) is the location downwind of the source, \( t \) is time, and the brackets indicate an ensemble average, that is, \( C = \langle c \rangle \). The natural variability, also called the inherent uncertainty (Fox 1984; Venkatram 1982), is caused by PBL turbulence. It arises because the details of the velocity field are not the same in each realization of a turbulent flow (Chatwin 1982; Tennekes and Lumley 1972) and because of the finite-averaging time of the concentration. Generally \( \sigma_c \) decreases with an increase in the averaging time [Eq. (2.13)]. In addition to \( x, t \), and the averaging time, \( \sigma_c \) depends on the meteorological and source conditions defining the ensemble.

Due to limitations in replicating field experiments, \( \sigma_c \) is difficult to obtain from them and generally must be simulated in laboratory experiments or modeled.

For many PBL dispersion problems, the natural variability is large, \( \sigma_c \sim C \), for averaging times of about 1 h and for distances of 30 km or less. A prime example is the \( \sigma_c \) in tall stack plumes. Laboratory experiments show that \( \sigma_c/C \) can be as large as 6 near the surface for simulated averaging times of a few minutes (Deardorff and Willis 1988; Fackrell and Robins 1982a.b); the averaging time for the CBL experiments is discussed in section 5b (see probability distribution of \( c_c/\sigma_c \)). For hourly averaged concentrations, we estimate that \( \sigma_c/C \) would be of the order of 1 assuming that \( \sigma_c^2 \) varies inversely with the averaging time (Sykes 1984). Large \( \sigma_c/C \) ratios have also been deduced from lidar aerosol measurements in elevated plumes (Lewellen and Sykes 1986). Another example with \( \sigma_c/C \sim 1 \) is multiple-source dispersion in an urban area (Hanna 1982).

Air-quality models used in regulatory applications predict \( C \) and with few exceptions ignore \( \sigma_c \). A key question is: How can such models be evaluated in the presence of a large natural variability? In principle, this is a sampling problem and should be addressed by collecting a large number of observations under the same conditions to minimize the uncertainty in \( C \). For example, if \( c \) followed a Gaussian distribution, the uncertainty in the observed \( C \) would be 1.96\( \sigma_c N^{-1/2} \) at the 95% confidence level, where \( N \) is the number of independent observations. To achieve a 20% accuracy in \( C \), we would require \( N = 100 \) if \( \sigma_c/C = 1 \) and \( N = 25 \) if \( \sigma_c/C = 0.5 \). The main point is that the \( N \) in typical field experiments is small, one or a few cases, if we restrict the ensemble definition to narrow ranges of the variables controlling dispersion. In section 4, we discuss ways of overcoming this problem.

b. Concentration fluctuations in point-source plumes

In the following, we cite examples of concentration fluctuations, many of which are from plumes in the CBL. From a practical viewpoint, the CBL is an important case because tall stack plumes usually produce their highest GLCs during convection and with \( \sigma_c \sim C \) (see Venkatram 1988; Weil 1985).

In the bulk of the CBL, the large convective eddies scale in size with the CBL depth \( h \), and have velocities that are proportional to the convective velocity scale \( w_\ast \) (Deardoff 1972):

\[
w_\ast = \left( \frac{gw \theta_0 h}{\theta_0} \right)^{1/3},
\]

where \( g \) is the gravitational acceleration, \( w \theta_0 \) is the surface heat flux, and \( \theta_0 \) is the mean potential temperature. At midday over land, typical values of \( w \) and \( h \) are 1 to 2 km and 2 m s\(^{-1}\).

1) Physics of fluctuations

Concentration fluctuations in plumes arise from two principal mechanisms (Csanady 1973): 1) the "instantaneous" plume entrains ambient air due to small-scale (inertial-range) turbulence, which leads to concentration fluctuations over distance scales less than the plume width, and 2) a small plume meanders due to the large-scale (energy-containing) eddies in a turbulent flow. These eddies are of the order of the turbulence integral length scale \( \Lambda \) (e.g., Sykes 1988). Meandering causes concentrations at a fixed receptor
to vary essentially between "in-plume" peaks and the zero value in the environment; it is an important source of fluctuations as long as the instantaneous plume width \( \sigma_t \) is less than \( \Lambda \). The inequality \( \sigma_t < \Lambda \) holds for \( t < T_L \), where \( t \) is time and \( T_L \) is the Lagrangian integral time scale.

Figure 1a illustrates the highly fluctuating concentration profile inside an instantaneous plume as well as the meander. If the concentration is averaged over a finite time period in an Eulerian or fixed reference frame, much of the meandering and internal fluctuations are removed and a smoother concentration profile results (Fig. 1b). Averaging the concentration over a much longer time period or a large number of realizations approximates the ensemble-averaged pattern (Fig. 1c), which is even smoother.

Field observations have been useful for understanding concentration fluctuations in plumes. For example, Fig. 2a shows a sequence of crosswind \( \text{SO}_2 \) GLC profiles in a plume downwind of the Dickerson (Maryland) power plant. These "quasi-instantaneous" profiles were obtained from a mobile instrumented van that traversed the plume six times during an approximate 1-h interval; they were obtained during a convective period with \( h \approx 400 \) m. The profiles were constructed by sampling significant points from a raw concentration trace and thus do not show the high-frequency fluctuations over small spatial scales; for example, compare Fig. 2a with the unsmoothed laboratory profiles in

**Fig. 2.** (a) "Quasi-instantaneous" ground-level concentration profiles measured from repeated passes by a mobile van along a measurement route transverse to a stack plume downwind from the Dickerson power plant, (b) average concentration profile for the six passes, (c) ratio of the root-mean-square concentration to the average value. From Weil and Jepson (1977).

**Fig. 3.** The profiles (Fig. 2a) show the lateral meander and the change in the maximum concentration during the averaging period. Figure 2b shows that the average of the six profiles is not Gaussian as might be expected; this may be due to plume movement during the traverses, the finite number of profiles, etc. Figure 2c shows the concentration fluctuation intensity for the
1-h period, \((c'_{1/2} / \sigma^2)^{1/2}\), where the overbar denotes an average over the six profiles and \(c' = c - \bar{c}\). Note that the minimum intensity occurs near the maximum concentration, and it increases at the plume edges.

More detailed field observations of concentration fluctuations have been made in plumes from near-surface sources (e.g., Dinar et al. 1988; Hanna and Insley 1989; Mylne and Mason 1991; Sawford 1987) but they were generally limited to downwind distances \(< 100\) m. An exception is the study by Mylne and Mason, who reported measurements out to 1 km from the source. In addition, their results as well as those of Sawford (1987) demonstrated the importance of lateral meandering on the probability distribution of concentrations and on the intensity, \(\sigma_c / C\), of concentration fluctuations.

Laboratory experiments also have provided insight into the behavior of concentration fluctuations in plumes. Figure 3 shows several realizations of the crosswind concentration profile in plumes from the convection-tank experiments of Deardorff and Willis (1984). The source height was 0.13\(h\), and the profiles were obtained for both nonbuoyant and buoyant releases. These profiles show the highly fluctuating concentration inside the plume as well as the meander about the mean position \((y = 0)\).

Fackrell and Robins (1982a,b) conducted wind-tunnel experiments in a neutral boundary layer and found that the fluctuation intensity \((c'^2)^{1/2} / C\) was generally much greater for an elevated release than for a ground-level source; this is shown in Fig. 4. For the elevated plume, the large intensity is due to the small plume size \((\sigma_r \ll \Lambda)\), the significant plume meander \((\sigma_z \ll \Lambda)\), and the distance of the ground from the mean plume axis. For these conditions, the plume only occasionally or intermittently reaches the ground, but it does so with a high concentration (due to the small \(\sigma_r\)), thus creating ‘spikes’ in the GLC distribution and a large concentration variance (see also Venkatram 1979). The \(\sigma_r\) is governed by turbulence in the inertial subrange (Batchelor 1950), which for small times \((t \ll T_z)\) is dependent upon the source diameter. The elevated source data (Fig. 4) show that the fluctuation intensity systematically increases with a decrease in the source size (i.e., smaller initial \(\sigma_r\)) and decreases with downwind distance (due to the increase in \(\sigma_r\)).

For a surface release, the plume dimension is of the order of the local eddy size or integral scale, and vertical meandering is not an important source of concentration fluctuations (Sykes 1988). However, as discussed previously, lateral meandering can lead to significant fluctuations in concentration.

The Deardorff and Willis (1984) results for the CBL are similar to those of Fackrell and Robins. Figure 5 shows the fluctuation intensity near the surface for a nonbuoyant plume in the CBL as a function of the dimensionless downwind distance \(X\):

\[
X = \frac{w_s x}{U h},
\]

where \(x\) is the downwind distance and \(U\) is the mean wind speed (averaged over the CBL). The intensity decreases significantly from its near-source peak with increasing distance; the same trend occurs for the buoyant-plume data (not shown). Again, this is attrib-
Fig. 5. Concentration fluctuation intensity at ground level due to an elevated, nonbuoyant source \((z_{\ast}/h = 0.1)\) as a function of the dimensionless downwind distance in a laboratory convectively mixed layer. \(\Delta\) represents an average over \(|y| < \sigma_y/2\); \(\bigcirc\) is an average over \(\sigma_y/2 < |y| < \sigma_y\). From Deardorff and Willis (1984).

The physical parameter \(\sigma_y/\Lambda\), or more specifically \(\sigma_y/h\), with distance (in the CBL, \(\Lambda \approx 1.5h\); Kaimal et al. 1976).

2) PROBABILITY DENSITY FUNCTION OF CONCENTRATION

Functional forms of the probability density function (pdf) of concentration, \(p(c)\), have been suggested for point-source plumes based on laboratory and field measurements. These include the lognormal, gamma, exponential, and clipped normal pdf’s.

The lognormal pdf is given by

\[
p(c) = \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left[ -\frac{\ln(c/c_0)^2}{2\sigma_c^2} \right]
\]

where \(c_0\) and \(\sigma_c\) are the geometric-mean concentration and the logarithmic mean standard deviation (Csanyi 1973). Csanyi (1973) gave a simple plausibility argument for this pdf based on an assumed series of diluting impulses. The pdf applies only to in-plume (i.e., nonzero) concentrations. Csanyi suggested that the situation with zero concentrations could be described by including the intermittency. Nettlerville (1979) followed this suggestion and found the result to be in approximate agreement with his wind-tunnel measurements of an elevated plume.

The gamma pdf is given by

\[
p(c) = \frac{\lambda(c/C)^{\lambda-1}}{\Gamma(\lambda)C^\lambda} \exp\left[ -\frac{\lambda c}{C} \right]
\]

where

\[
\lambda = \frac{C^2}{\sigma_c^2},
\]

and \(\Gamma(\lambda)\) is the gamma function (e.g., see Deardorff and Willis 1988). Here, \(C\) and \(\sigma_c\) are the mean and rms values of in-plume or nonzero concentrations. As discussed by Wilson and Simms (1985), this pdf is an attractive compromise between the lognormal and exponential pdf’s [Eq. (2.6)]. For large relative concentrations, \(c/C > 1\), the gamma pdf is similar to the lognormal, and when \(\lambda = 1\), it reduces to the exponential form.

The exponential pdf is a special case of the gamma and is given by

\[
p(c) = \frac{c}{C} \exp\left( -\frac{c}{C} \right)
\]

where \(\gamma\) is the intermittency factor that measures the fraction of the time that the plume is present; that is, \(c\) exceeds a small threshold value. Wilson (1982) argued that the exponential pdf corresponds to a mixing process in which the concentration “spikes” are produced by eddies separated by a sufficiently long time that they are independent of one another. Barry (1977) and Hanna (1984) showed that this form fit observed pdf’s downwind of near-surface releases in flat terrain and in a building wake (Barry).

Deardorff and Willis (1988) compared the gamma pdf to observed pdf’s of nonbuoyant tracers released from elevated sources in their laboratory experiments. The measurements were of near-instantaneous concentrations exceeding a threshold of \(0.01Q/\pi h^2\) and were obtained in the region \(0 < z/h < 0.2\) and \(|y|/\sigma_y < 1\), that is, along the plume center plane; here, \(Q\) is the source emission rate. Figure 6 shows their observed pdf’s, which are plotted as histograms within logarithmically spaced intervals on the abscissa. The histograms cover dimensionless distances \(X\) ranging from 0.22 to 4.78. The gamma pdf is shown as a solid curve and is a good approximation to the data for \(X > 1\). For smaller \(X\), the approximation is less good and may be due to an insufficient number \((N)\) of samples in the observed pdf’s.

Deardorff and Willis also measured the pdf’s for buoyant effluents released from the same source height. The buoyancy was characterized by the dimensionless flux \(F_\ast\) defined by

\[
F_\ast = \frac{F_b}{U\delta^2 h^{\frac{3}{2}}},
\]

where \(F_b\) is the stack buoyancy flux, which is proportional to the stack heat flux (Briggs 1984). For a weakly buoyant plume, \(F_\ast = 0.03\), they reported that the gamma pdf did not fit the data as well as it did for neutral plumes. For a strongly buoyant plume, \(F_\ast = 0.26\), the pdf’s could not be characterized by Eq. (2.5a) or any other simple function. However, in section 5, we show that the cumulative distribution functions (cdf’s) of \((c - C)/\sigma_c\) are about the same for neutral and highly buoyant plumes. (The similarity in the cdf’s may result from the smoothing of differences in the pdf’s due to the integration of them.)

The clipped normal pdf was proposed by Lewellen and Sykes (1986) based on an analysis of aerosol and
3) Models for $\sigma_e$

Models for $\sigma_e$ are required to characterize the pdf of $c$. There are two existing models that we consider adaptable to applications—Gifford’s (1959) meandering-plume model and ensemble-closure approaches. The latter include Csanady’s (1967) similarity model and the Sykes et al. (1984, 1986) model. The meandering model describes the fluctuations due to the large-scale eddies alone and is applicable when $\Lambda > \sigma_e$. In contrast, Csanady’s model treats the fluctuations due to in-plume turbulence only and is strictly valid for $\Lambda < \sigma_e$. Both models are limited to homogeneous turbulence. The Sykes et al. model applies to both $\sigma_e$ regimes and has been adapted to inhomogeneous turbulence.

In Gifford’s model, the spatial distribution of $c$ in an instantaneous plume is assumed to be nonrandom and to have a Gaussian distribution about the instantaneous centroid position $y_c$, $z_c$—in the crosswind $(y)$ and vertical $(z)$ directions. It is

$$c(x, y, z; y_c, z_c) = \frac{Q}{2\pi U\sigma_r^2} \exp \left[ -\frac{(y - y_c)^2 + (z - z_c)^2}{2\sigma_r^2} \right]$$

(2.9)

where the concentration distribution is assumed to be axisymmetric; that is, $\sigma_{yr} = \sigma_{zp} = \sigma_r/\sqrt{2}$. Gifford assumed the centroid coordinates to be random uncorrelated variables and to have a Gaussian pdf. To find $C$, Gifford averaged the $c$ in (2.9) over the pdf of $y_c$, $z_c$ and obtained

$$C(x, y, z) = \frac{Q}{2\pi U(\sigma_r^2 + \sigma_{me}^2)} \exp \left[ -\frac{y^2 + z^2}{2(\sigma_r^2 + \sigma_{me}^2)} \right],$$

(2.10)

where $\sigma_{me}$ is the rms displacement of the meander, which is assumed to be the same in the $y$ and $z$ directions. Equation (2.10) is the familiar Gaussian plume model routinely used in dispersion applications.

By averaging the quantity $c^2$ found using Eq. (2.9), Gifford obtained

$$\langle c^2(x, y, z) \rangle = \frac{Q^2}{(2\pi U\sigma_r)^2(2\sigma_{me}^2 + \sigma_r^2)} \times \exp \left[ -\frac{y^2 + z^2}{2\sigma_{me}^2 + \sigma_r^2} \right],$$

(2.11)

and then found the concentration variance to be

$$\sigma_{e\lambda}^2 = \left\langle c^2 \right\rangle - C^2,$$

(2.12)

where subscript $\lambda$ denotes the instantaneous plume.

Fackrell and Robins (1982a) presented one of the first tests of Gifford’s model and used data from wind-tunnel simulations in a neutral boundary layer. They estimated the total spread $(\sigma_r^2 + \sigma_{me}^2)^{1/2}$ and $\sigma_e$ from the models of Hay and Pasquill (1959) and Smith and Hay (1961); $\sigma_e$ was based on plume growth in the inertial subrange.
Figure 7 shows the intensity of the fluctuations, \( \langle c'^2 \rangle_{m^2}/C_m \), where subscript \( m \) denotes the maximum value over the plume at distance \( x \). As can be seen, there is reasonable agreement between the theory (lines) and the experimental data for the elevated source located at 0.19h; \( \Lambda_2 \) is the integral length scale normal to the wind direction. The peak fluctuations are larger for the smaller sources (smaller \( \sigma_r/\Lambda_2 \), where \( \sigma_r \) is the source diameter) due to the higher concentrations in the narrower plumes from these sources.

The main advantages of Gifford's model are its simplicity and its applicability to the near-source region \((t < T_a)\) for an elevated source where the fluctuations are largest. However, it has four main limitations. 1) It neglects fluctuations within the instantaneous plume, which become more important farther downstream as \( \sigma_r \) becomes comparable to \( \sigma_{me} \). 2) It is restricted to homogeneous turbulence. 3) It is limited to nonbuoyant plumes. 4) It applies only to the concentration variance of an instantaneous plume. In many applications, interest is in the variance of time-averaged concentrations, which is smaller than \( \sigma_{e}^2 \).

Sykes (1984) gave a useful extension of Gifford's model for the \( \sigma_{e}^2 \) of time-averaged concentrations. Assuming a stationary flow, he used the following expression relating \( \sigma_{e}^2 \) to \( \sigma_{e}^2 \) (see Tennekes and Lumley 1972):

\[
\sigma_{e}^2(T_a) = 2 \frac{\rho^2}{T_a} \int_0^{T_a} (1 - \frac{t'}{T_a}) \rho(t') dt'
\]

(2.13)

where \( T_a \) is the averaging time, and \( \rho(t') \) is the autocorrelation function of the concentration at a fixed location. Sykes developed an expression for \( \rho(t') \) by assuming that the centroid position has an exponential autocorrelation with an integral time scale equal to the velocity time scale. His model is limited to \( t < T_a \) but is useful because that is where the fluctuations are largest.

Venkatram (1984) gave another model for \( \sigma_r \) that is applicable to elevated sources in the CBL. The key idea is that the concentration variance is caused by the unknown variation of the pdf’s of the vertical and lateral velocity fluctuations over the sampling period. The model predicts that \( \sigma_r \) decreases with downstream distance in accord with the meandering-plume model, but it has not been tested with observations.

In contrast to the statistical nature of the meandering-plume model, the basis of Csanady's (1967) Eulerian model is a differential equation for the concentration variance, which was developed from the diffusion equation. In addition to adopting eddy-diffusion theory, Csanady assumed similarity of the mean concentration and variance (spatial) fields in order to obtain simple solutions to the differential equation. Csanady (1973) showed that the model described data from laboratory experiments and a lake dye study. A number of simplifications and adaptations of this model have been made (Hanna 1984; Wilson et al. 1982; Nettles et al. 1979) in which the simplest yields a Gaussian spatial distribution for \( \sigma_{e}^2 \).

Sykes et al. (1984) developed a second-order closure model based on a differential equation for \( \sigma_{e}^2 \) as in Csanady's approach. However, recognizing the importance of meandering at early times, they modified this approach to incorporate the description of a narrow wandering plume with a dimension \( \sigma_r \). At short times, \( \sigma_r \propto t^{1/2} \) as predicted by inertial-subrange theory (Batchelor 1950). Sykes et al. found good agreement between their \( \sigma_{e}^2 \) predictions and the Fackrell and Robins (1982b) wind-tunnel data. Sykes et al. (1986) developed a simpler version of the full model using an integral method, that is, by integrating the complete equations over a plume cross section. This model also agreed well with the Fackrell and Robins data.

As this discussion suggests, \( \sigma_r \) models are in the developmental stage and more work is necessary to address buoyancy and the complications of PBL turbulence; for further review see Sykes (1988). Extension of the models to more general conditions may follow from state-of-the-art research efforts. For example, Sykes et al. (1988) numerically calculated \( \sigma_r \) for elevated sources in the CBL using large-eddy simulation (LES) for the large-scale, resolvable turbulence and a puff model for the unresolved, subgrid-scale motion. More recently, Sykes and Henn (1992) applied this approach to elevated sources in the neutral boundary layer and obtained good agreement with the Fackrell and Robins (1982b) data. Another approach is the Lagrangian two-particle model recently developed by Thomson (1990); it also reproduced the Fackrell and Robins data. This approach is currently limited to homogeneous conditions, but it could be coupled with LES to simulate concentration fluctuations in the inhomogeneous turbulence field of the PBL.

![Figure 7. Predictions of the concentration fluctuation intensity from Gifford's (1959) model (solid lines) compared with the wind-tunnel data shown in Fig. 4; predictions given only for elevated sources (\( x/h = 0.19 \)) where \( x \) is source diameter and \( \Lambda_2 \) is the integral length scale normal to the mean wind direction. From Fackrell and Robins (1982a).](image-url)
3. Current practice in operational performance evaluation

The performance measures recommended by the 1980 AMS Woods Hole workshop (Fox 1981) are now routinely used in model evaluation. In the following, we review these measures and their application and discuss bootstrapping, a statistical resampling technique for estimating confidence intervals.

a. Woods Hole performance measures

The 1980 Woods Hole workshop recommended that a set of statistical measures be determined for all dispersion models in a given category (e.g., for tall stacks) in the hope of discriminating the better from the poorer models. Their main application was intended for comparisons between observed, \( c_o \), and predicted, \( c_p \), concentrations although they also could be applied to other variables, such as the plume widths (\( \sigma_y \) and \( \sigma_z \)). The measures characterized the residual or the difference \( d \) between \( c_o \) and \( c_p \)

\[
d = c_o - c_p,
\]

and their correlation. (Here, \( c_o \) and \( c_p \) represent individual realizations of observed and predicted concentrations that are subject to uncertainties; that is, they are not the ensemble-mean values denoted by \( C_o \) and \( C_p \).)

The difference measures included the average or bias,

\[
\langle d \rangle = \frac{1}{N} \sum_{i=1}^{N} d_i
\]

and measures of the variability in \( d \). These consisted of three alternatives: the variance \( \sigma_d^2 \), the mean-square error \( \langle d^2 \rangle \), and the average absolute error \( \langle |d| \rangle \):

\[
\sigma_d^2 = \frac{1}{N-1} \sum_{i=1}^{N} (d_i - \langle d \rangle)^2
\]

\[
\langle d^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} d_i^2
\]

\[
\langle |d| \rangle = \frac{1}{N} \sum_{i=1}^{N} |d_i|
\]

where \( N \) is the number of observations.

Tests for the correlation between observations and predictions were recommended for time, space, and combined time–space pairings of \( c_o \) and \( c_p \). In addition to determining the correlation and difference measures for all paired data, that is, paired in space and time, the measures were recommended for maximum concentrations only, for example, the highest 25 1-h averaged \( c_o \) and \( c_p \) from a year of measurements. This kind of comparison was considered for a small number (approximately five) of monitors around an isolated tall stack and was probably included for two reasons.

First, the small monitoring network was typical of the experimental data available around tall stacks. Experience showed that large \( d \) values occurred when one compared \( c_o \) and \( c_p \) paired in space and time and that this comparison yielded little or no information. The main problem was the uncertainty in the plume transport direction (see Appendix) and the inability to resolve it with only a small number of monitors. Second, the primary regulatory concern was with the maximum concentrations, over specified averaging times, irrespective of where or on which they occurred. It was not important that the predictions and observations be correlated well in space and time, only that the models predict the magnitude of the highest concentrations. However, the participants agreed that this kind of comparison said little or nothing about the physics or the believability of the model.

The participants in the Woods Hole workshop expressed several caveats concerning the performance measures. Some important ones were:

1) For an elevated point source, the measures were most applicable to a well-described concentration field as obtained from a dense array of approximately 100–200 fixed, ground-level monitors; for example, the tracer experiment sponsored by the Electric Power Research Institute (EPRI) at the Kincaid power plant (Bowne et al. 1983). Another approach was the use of a mobile, instrumented van in which average crosswind concentration profiles were obtained from repeated van transects of the plume (Weil and Jepsen 1977). In either approach, the plume centerline as defined by the locus of the maximum concentration on crosswind arcs could be adequately determined as well as the concentration along it. Therefore, one could remove one of the major sources of uncertainty—the plume direction.

2) Statistical tests performed with the difference measures would be most meaningful if \( d \) were normally distributed. Thus, it would be useful to transform \( c \) to a variable that was nearly normal, such as \( \ln c \) for the point-source problem.

3) The statistical evaluation alone could not determine the best models; a scientific evaluation might be the only approach to determine them.

The EPA has used the foregoing measures extensively in model evaluation but generally with little or mixed success in distinguishing model performance. In addition to Clifty Creek (Smith 1984), EPA evaluated point-source models at two other power plants—Muskingum River in Ohio (Cox and Moss 1985) and Kincaid in Illinois (Cox et al. 1986). These evaluations also were conducted with year-long records of SO\(_2\) GLCs at stationary monitors with attention on 1-, 3-, and 24-h averaging periods. The evaluations emphasized comparisons between the 25 highest \( c_o \) and \( c_p \) values unpaired in space or time. The results were generally the same as summarized by Smith (1984) with the exception of a new model, PPSP, added to the group.
PPSP (Weil and Brower 1984) is an updated Gaussian plume model with dispersion and stability formulations based on recent understanding of the CBL. Weil and Brower evaluated the model with an intensive dataset including centerline SO₂ GLCs obtained from a mobile van. They found that PPSP performed much better than the CRSTER model (EPA 1977), which is based on the Pasquill–Gifford–Turner stability classification and dispersion curves. However, the total record length of their dataset was much less than a year.

In contrast, the performance evaluations at the Muskingum River, Kincaid, and Clifty Creek (Tikvar and Cox 1984) plants showed that the PPSP model typically overestimated the maximum concentrations by a factor of 3 and was much more biased than the other models. The highest predictions were made for relatively shallow boundary layers (<500 m) and light winds (~2 m s⁻¹) when buoyant plumes tend to remain near the top of the boundary layer and resist downward mixing. This situation was not revealed in the initial evaluation with the Maryland dataset (Weil and Brower 1984) but was in the later one (Weil and Corio 1985; see section 4); the problem was corrected to a large extent in a revised model (Weil and Corio 1988).

In summary, the Woods Hole measures generally have been of limited use in discriminating the operational performance of models. However, our discussion has shown that models developed and tested with intensive datasets of limited duration should be evaluated with routine monitoring records to capture all combinations of meteorological conditions. Thus, it brings out the value of the routine data records (see also Irwin and Smith 1984).

b. Simplification of performance measures

In line with the AMS steering committee recommendations (section 1), Cox and Tikvar (1986) proposed simple measures—the “fractional” bias FB and the “fractional” scatter FS—to highlight model performance at the upper end of the concentration distribution. They applied them to the 25 highest \( \bar{c}_p \) and \( \bar{c}_o \) values, which were unpaired in space or time. The FB and FS are defined by

\[
FB = \frac{2(\bar{c}_o - \bar{c}_p)}{\bar{c}_o + \bar{c}_p} \quad (3.6)
\]

\[
FS = \frac{2(\sigma_{co} - \sigma_{cp})}{\sigma_{co} + \sigma_{cp}} \quad (3.7)
\]

where the overbar here indicates an average over the 25 highest concentrations and \( \sigma_c \) is the standard deviation of the concentrations about the average. The fractional bias was used earlier by Irwin and Smith (1984) to summarize model performance.

The FB and FS indicate how well the model reproduces the average and spread, respectively, of the highest observed concentrations. The ideal values of these quantities are zero, but they can range from -2 to +2; a value of ±0.67 corresponds to a prediction within a factor of 2 of the observation.

More recently, Cox and Tikvar (1990) presented a statistical method for evaluating model performance in predicting extreme concentrations (in the upper tail of the cdf) for a typical 1-yr period. The method uses FB at a specified percentile (e.g., 99th) as the basic performance measure and an exponential tail fit to the cdf. An approach is given for combining results from different averaging times, sources, and monitors to generate a composite performance measure. The uncertainty in this measure as well as the statistical significance of differences between model results is assessed by the bootstrap method, and an example application is discussed.

c. Bootstrapping

Bootstrapping is a statistical resampling technique (Efron 1982) used to make nonparametric estimates of the mean, variance, or other statistic of the cdf (e.g., the 60th-percentile value). By nonparametric, we mean that the method does not depend on an assumed sample distribution such as a Gaussian. Instead, the method uses the sampled or observed distribution to obtain the estimates. In addition to estimating statistical properties of the cdf, one can place confidence limits on the estimates. This makes it possible to determine the statistical significance of differences between a modeled mean concentration and observations, or of differences between concentrations predicted by various models. Bootstrapping is numerically intensive and has been made possible by plentiful and inexpensive computer power. (Another resampling method is the jackknife; see Tukey 1987.)

The key assumptions in bootstrapping are that the sampled (observed) distribution is unbiased and that it is comprised of independent members. Efron (1982) demonstrated that the approach yields reasonable estimates of means and standard deviations of known distributions—Gaussian, gamma, and lognormal.

Bootstrapping has become popular in the evaluation of air-quality models as a means of testing the significance of model differences. However, there are two problems that must be considered in applications.

1) To have an unbiased distribution, each member of the distribution should be drawn from the same ensemble of events, that is, those with the same meteorology and source conditions. In many bootstrapping applications, the hourly averaged GLCs from a year of monitoring data are lumped together, and thus, the ensemble is really a mixture of different ensembles. A comparison between measurements and predictions from such a mixture may be a useful test of operational performance, but it does not constitute a test of model physics—that the model gives good predictions for the right reasons. To some extent, this problem has been handled by blocking the observations—separating the
data into blocks or groups representing different conditions. In data resampling, care is taken to ensure that the different blocks are appropriately represented (see Cox and Tikvart 1990; Tukey 1987).

As discussed by Efron (1982, section 10), the estimation of confidence intervals for a statistic is more sensitive to the observed distribution than is the estimate of that statistic; the confidence limit is sensitive to the tails of the observed distribution. Misleading confidence limits—generally too small—can be obtained for skewed distributions such as an exponential. Efron showed that the combination of bootstrapping with an exponential tail fit gave better estimates of confidence intervals than the bootstrap alone (see also Cox and Tikvart 1990). However, even with this fit, the estimation of confidence limits for the 98th-percentile value and above is uncertain and can be in significant error (see Rao et al. 1985).

4. Separating model error from natural variability

a. Preferred approach

To test models and their physics rigorously, we must separate the model error—the difference between the observed and predicted ensemble-mean concentrations—from the natural variability and other sources of concentration variance. A key issue is the definition of an ensemble of realizations for determining the observed concentration statistics. As discussed earlier, this should follow from the physics of the dispersion problem and should include the meteorological variables characterizing dispersion as well as the source conditions. The variables would include the mean wind profile, surface heat and momentum fluxes, PBL depth, averaging time, etc. Assuming that the model is based on sound physics, these variables would be included in the model input array.

This led Venkatram (1982) to suggest that an ensemble be defined in terms of the model inputs. The AMS workshop on model uncertainty (Fox 1984) implicitly adopted this definition in determining the contributions to the mean-square error \( \langle d^2 \rangle \) between observations and predictions. Using somewhat different notation, we briefly summarize and expand their analysis for this choice of an ensemble. The purpose is to isolate the model error from other contributions to \( \langle d^2 \rangle \). In the following, the model inputs are denoted by \( \xi \) and an uncertainty in them by \( \xi' \); for example, \( \xi \) could be the mean wind, \( \xi \) the PBL depth, etc.

The model prediction is decomposed as

\[
C_p(\xi, \xi') = C_p(\xi) + c'_p(\xi, \xi')
\]  \hspace{1cm} (4.1)

where \( C_p \) is the predicted ensemble-mean concentration and \( c'_p \) is a prediction uncertainty due to \( \xi' \) with \( c'_p(\xi, 0) = 0 \). The observed concentration (subscript \( o \)) is decomposed as

\[
c_o(\xi, \mu_m) = c_o(\xi) + c'_o(\xi, \mu_m)
\]  \hspace{1cm} (4.2)

where \( c'_o \) and \( c'_o \) are the fluctuations about the mean \( C_o \) due to measurement error and natural variability, respectively. Here, \( \mu_m \) represents those variables that are not included in the model inputs and that vary from realization to realization; Lumley and Panofsky (1964) called \( \mu_m \) the realization index. The variation in quantities represented by \( \mu_m \) in different realizations leads to the random fluctuation \( c'_o \) (Venkatram 1982). In Eqs. (4.1) and (4.2), all of the fluctuation (primed) terms are assumed to have a zero mean.

With the given decomposition of \( c_p \) and \( c_o \), the residual \( d \) can be expressed as

\[
d = C_o + c'_o + c'_p - C_p - c'_p,
\]  \hspace{1cm} (4.3)

and its ensemble average, that is, the bias or model error, as

\[
\langle d \rangle = C_o - C_p.
\]  \hspace{1cm} (4.4)

If it is assumed that \( c'_o \), \( c'_p \), and \( c'_o \) are uncorrelated with one another or with \( \langle d \rangle \), then \( \langle d^2 \rangle \) is given by

\[
\langle d^2 \rangle = \langle d^2 \rangle + \langle c'_o^2 \rangle + \langle c'_p^2 \rangle + \langle c'_o^2 \rangle
\]  \hspace{1cm} (4.5)

as shown by Fox (1984). The first term on the right-hand side (rhs) of this expression is the square of the model error, while the second and third terms are the uncertainties in \( c_p \) and \( c_o \) due to uncertainties or errors in measurements. In principle, all of these terms can be minimized. The fourth term is the square of the natural variability, that is, \( \langle c'_o^2 \rangle = \sigma_o^2 \), which is essentially irreducible for a given model.

This decomposition of \( \langle d^2 \rangle \) is considered applicable to any dispersion problem. However, in most field experiments an ensemble defined in terms of the same model inputs does not repeat with sufficient frequency to determine \( \langle d \rangle \), \( \langle d^2 \rangle \), and \( \sigma_o^2 \). In some problems, however, there is the possibility that the dimensionless fluctuation \( c'_o/\sigma_o \) from different ensembles (different \( \xi' \)s) will be nearly universal, that is, will have the same probability distribution. This suggests that one should analyze the dimensionless residual \( d/\sigma_o \), which is advantageous because it leads to a new ensemble with many more samples.

We now consider a problem in which the stochastic fluctuation \( c'_o \) has a universal or self-similar probability distribution in the sense that it can be decomposed as

\[
c'_o(\xi, \mu_m) = \omega(\mu_m)\sigma_o(\xi)
\]  \hspace{1cm} (4.6)

(Venkatram 1982; Weil 1985), where \( \omega \) is a random variable that is dependent at most on \( \mu_m \), is uncorrelated with \( \sigma_o \), and has unit variance, \( \langle \omega^2 \rangle = 1 \). Laboratory data on in-plume concentration fluctuations show that an approximate self-similar distribution of \( c'_o/\sigma_o \) exists for a point-source plume in the CBL (section 5b). (As discussed in section 5b, this probably occurs because of the relatively narrow range of \( \sigma_o/C \) for this data, 0.6 < \( \sigma_o/C \) < 1.6.)

We now substitute Eq. (4.6) into Eq. (4.3), square...
both sides of that equation, divide the result by \( \sigma_c^2 \),
and take the average. The result is
\[
\left( \frac{d}{\sigma_c} \right)^2 = \left( \frac{C_o - C_p}{\sigma_c} \right)^2 + \left( \frac{c'}{\sigma_c} \right)^2 + \left( \frac{c''}{\sigma_c} \right)^2 + 1,
\]
where we have assumed that the correlations between \( \omega, c'/\sigma_c, c''/\sigma_c, \) and \( (C_o - C_p)/\sigma_c \) are zero. The average in Eq. (4.7) is over many \( d/\sigma_c \) values, each of which comes from a different ensemble as defined by \( \xi_n \). The common link between these ensembles is that \( c'/\sigma_c \) has the same probability distribution.

Further progress can be made if the first three terms on the rhs of Eq. (4.7) are all small compared to 1, that is, \( \sigma_c \) is the dominant contribution to \( \langle d^2 \rangle \) and the residuals. In this case, \( C_p \approx C_o \) and therefore
\[
c'/\sigma_c \approx (C_o - C_p)/\sigma_c \approx (C_o - C_p)/\sigma_c \approx d/\sigma_c.
\]
(4.8)

This means that the cdf of \( d/\sigma_c \) should approximate that of \( \omega \). This is an important result because the cdf of \( \omega \) is necessary to estimate the confidence limits on the model error, \( (C_o - C_p)/\sigma_c \). However, to use \( d/\sigma_c \) in this way, we must first demonstrate that the model error is indeed small, that is, \((C_o - C_p)/\sigma_c \ll 1\).

In the more general case, \( \langle c''^2 \rangle \) is of the same order as \( \sigma_c^2 \) and thus is a major contributor to \( \langle d^2 \rangle \), as found in the work of Lewellen and Sykes (1989). Their results are for hourly averaged GLCs at stationary monitors where \( \langle c''^2 \rangle \) is dominated by the horizontal wind variance, that is, the uncertainty in the transport direction. The \( \langle c''^2 \rangle \) term can be simulated by a Monte Carlo approach in which one assumes distributions of pfd's of the meteorological input uncertainties, the \( \xi_n \)'s. Lewellen and Sykes (1989) and Lewellen et al. (1988) used this approach to estimate confidence limits on the predicted concentration cdf's (see section 5c).

In both of the cases given, one needs to determine the model error and an effective way of doing this is with a residual plot, a graph of \( d/\sigma_c \) versus the inputs \( \xi_n \). Ideally, the points should be symmetrically distributed about \( d/\sigma_c = 0 \) and show no trend with \( \xi_n \) (e.g., Venkatram 1982). However, with real data, a trend or lack thereof is often difficult to detect due to the broad scatter in \( d/\sigma_c \). Therefore, it is useful to divide the points into a finite number of intervals along \( \xi_n \) and calculate the mean \( d/\sigma_c \) and its uncertainty in each; the mean \( d/\sigma_c \) is an estimate of \( (C_o - C_p)/\sigma_c \) in the interval. If the uncertainty limits of \( \langle d/\sigma_c \rangle \) in each interval include zero, then the model error is not statistically significant.

While the foregoing approach is preferred, it is difficult to apply in practice because of the requirement for \( \sigma_c \), which would have to be modeled (section 2b). General \( \sigma_c \) models that account for plume buoyancy, PBL inhomogeneities, etc. do not exist, with perhaps one exception (section 5c). Until they do and are accepted, other methods for analyzing residuals and determining model errors are needed. In the following, we discuss one simple method that is in line with the foregoing approach.

b. A practical method

Weil and Corio (1985) evaluated a Gaussian model for buoyant-plume dispersion in the CBL. The model was identical to PPSP (Weil and Brower 1984) except that the dispersion parameters \( (\sigma_x, \sigma_z) \) were given as a continuous function of the stability index \( U/w_a \) rather than in "stability classes," that is, finite ranges of \( U/w_a \). The model was evaluated with the same data used to test the PPSP model and included SO2 GLCs approximating hourly averaged values along the plume centerline.

Weil and Corio had no model or estimates of \( \sigma_c \) to nondimensionalize \( d \). However, they found that \( c_p/c_o \) approximated a lognormal distribution. In line with the 1980 Woods Hole workshop recommendations, they defined the residual as \( d = \ln(c_p) - \ln(c_o) \) and used its distribution to estimate confidence limits on \( \ln(c_p/c_o) \) in \( \xi_n \) intervals.

One of the most significant trends of \( c_p/c_o \) occurred with the dimensionless buoyancy flux \( F_* \) (Fig. 8). This is somewhat difficult to detect from a residual plot using the individual points \( (N = 145) \) due to the broad scatter in \( c_p/c_o \), see Fig. 8a. We show this type of residual plot (individual points) because it has been used often in

![Fig. 8. Ratio of predicted-to-observed ground-level SO2 concentration as a function of dimensionless buoyancy flux for a Gaussian plume model, which is a version of the PPSP model but with dispersion parameters \( (\sigma_x, \sigma_z) \) varying continuously with the stability parameter \( U/w_a \): (a) individual points, and (b) geometric mean and uncertainty in mean for data in \( F_* \) intervals. From Weil and Corio (1985).](image-url)
model evaluation (e.g., Venkatram 1982; Weil and Brower 1984; Hanna and Paine 1989). Figure 8b shows the geometric mean (GM) of \(c_p/c_0\) and its uncertainty, denoted by the vertical bars, for six \(F_s\) intervals each containing 20–30 points. The uncertainty was estimated as the 95% confidence limit of a lognormal distribution using the geometric standard deviation (GSD) and number of points in each interval; the GSD includes both the natural variability and the variability due to meteorological input uncertainties. As can be seen (Fig. 8b), the GM of \(c_p/c_0\) is not significantly different from zero for \(F_s < 0.07\), but for \(F_s > 0.07\), the model has a clear bias toward overprediction. (The result in Fig. 8b could also be shown by a trend analysis of the data in Fig. 8a.)

A similar trend of \(c_p/c_0\) with \(F_s\) was found for the PPSP model and explains its error toward overprediction in the EPA evaluations discussed in section 3a. The error occurs due to the oversimplified treatment of a highly buoyant plume trapped in the CBL during light winds and low CBL depths, that is, high \(F_s (\geq 0.1)\). If a plume is unable to penetrate the elevated inversion, the model assumes that the plume loses buoyancy and disperses as a passive tracer once its centerline reaches \(h\). In contrast, laboratory experiments (Willis and Deardorff 1987) show that highly buoyant plumes tend to remain near the top of the CBL and disperse downwards gradually over long distances; that is, the plume buoyancy inhibits vertical mixing due to the ambient convection. Weil and Corio (1988) modified the PPSP model to account for this behavior and found much less bias for the revised model.

In summary, one of the most important characteristics to demonstrate in evaluation is that the model error is zero over a broad range of input variables. In the simple method, this can be shown using residual plots of the form \(\ln(c_p/c_0)\) versus \(\xi_s\). The statistical significance of the error is estimated from a lognormal distribution of \(c_p/c_0\). The GSD of that distribution accounts for both the natural variability and the variability due to meteorological input uncertainty.

5. Evaluating model physics: The point-source problem

The first step in evaluating model physics should be an assessment of the scientific basis and formulation of the model. The second step should be an evaluation of model predictions using available observations from laboratory and field experiments. This should test the predicted concentration field as well as individual model components (e.g., \(\sigma_v, \sigma_z\)) with emphasis on good predictions for the “right” reasons.

In this section, we address the second step of the evaluation process. We apply the concepts discussed in section 4 to the evaluation of model-predicted GLCs downwind of elevated point sources; the GLCs are short-term averages (~1 h) and the distance range is typically \(x < 30\) km. Emphasis is on dispersion in the CBL and on data from intensive field experiments in which the GLC distribution is mapped in sufficient detail to locate the plume centerline and the concentration along it. Knowledge of the centerline removes much of the concentration variance that is obtained in experiments with only a few stationary monitors (Weil and Corio 1985; see also Appendix). Techniques suitable for this purpose include a large network of ground-level monitors and a mobile van (section 3).

In evaluating model physics, we also stress the use of laboratory data (Willis and Deardorff 1976, 1978, 1981, 1987) and numerical simulations (Lamb 1982) because they can be obtained under controlled conditions and can be repeated to generate well-defined ensemble-mean fields. The data and simulations have been used to evaluate dispersion models ranging from Lagrangian stochastic approaches (Baerentsen and Berkowitz 1984; Sawford and Guest 1987; Thomson 1987) to simple techniques based on the pdf of vertical velocity (Misra 1982; Venkatram 1983). In contrast, clear trends of \(C\) as a function of distance, meteorology, etc. are difficult to obtain from field experiments due to an insufficient number of observations under the same conditions.

In the following, we discuss the evaluation of models for 1) the mean concentration, which is the main model output currently used in regulatory applications, 2) the fluctuating concentration, which should be used more in the future, and 3) the probability distribution of \(c\), which is predicted currently by one advanced model. We then discuss the relevance of these evaluations, especially 3), to regulatory practice.

a. The mean concentration

Model evaluation using data from intensive field experiments should include a determination of the correlation between \(c_0\) and \(c_p\) and an analysis of the residuals (sections 3 and 4). The correlation is a semi-quantitative indication of model performance since it is a statistic based on a mixture of ensembles; that is, each point in the \(c_0, c_p\) plot comes from a different ensemble based on different input conditions. In addition, the correlation depends on the range of \(c_0\) and \(c_p\). Nevertheless, the fraction, \(r^2\), of the variance explained by the model is a useful measure of performance when evaluating two or more models with the same dataset; in the following, \(r\) is the correlation coefficient between \(\ln(c_0)\) and \(\ln(c_p)\).

As an example, we show the performance of two models for tall stack releases: the pdf model (Weil et al. 1986), which is based on recent understanding of the CBL, and CRSTER, a standard Gaussian plume model (EPA 1977).

In the pdf approach, plume elements are assumed to be emitted into a field of convective updrafts and downdrafts that travel with the mean wind; the convection elements have a random vertical velocity \(w\) that is prescribed by the pdf, \(p(w)\), which is non-
Gaussian. Vertical displacements of the plume due to source buoyancy and ambient convection are superposed to determine the “instantaneous” centerline height, \( z_c \). The ensemble-mean, crosswind-integrated concentration (CWIC), \( C^* \), is found by averaging the instantaneous CWIC distribution (a Gaussian) over all possible \( z_c \) values; the latter are determined from \( p(w) \). Simple expressions are obtained for the CWIC and the GLC.

An important step in developing the pdf model was the comparison of its predictions to data from convection-tank experiments. Figure 9 shows a comparison between the modeled CWIC near the surface and the laboratory data; as can be seen, the two are in good agreement both as a function of \( X \) and \( F_s \).

The pdf and CRSTER models were evaluated with SF\(_6\) GLCs downwind of the Kincaid power plant. Figure 10 compares \( c_o \) with \( c_p \) where the observations are maximum 1-h averages on crosswind sampling arcs. On a relative basis, the pdf model performs better than CRSTER. For example, the GM and GSD of \( c_p/c_o \) are 1.1 and 2.1, respectively, for pdf, and 1.0 and 4.1 for CRSTER; that is, the scatter is much greater for CRSTER as shown in Fig. 10. Likewise, the pdf model explains a greater fraction of the variance between \( c_o \) and \( c_p \) (\( r^2 = .34 \)) than does CRSTER (\( r^2 = .02 \)). Further work is necessary to explain the rather low magnitude of \( r^2 \) in terms of the natural variability, input data uncertainties, and potential model errors.

The zero predictions by CRSTER—points along the ordinate (Fig. 10b)—were excluded from the above statistics, \( r \) and the GM and GSD of \( c_p/c_o \). They resulted either from a predicted plume penetration of the elevated inversion, with no ground-level impact, or from vertical dispersion estimates much lower than inferred from \( c_o \); the underestimated dispersion is attributed to a model bias toward neutral stability when unstable conditions actually exist (Weil and Brower 1984).

The residuals \( d \) from Fig. 10 were analyzed to determine model biases and their probable causes (Hanna et al. 1986, appendix C). This was done following the approach of section 4b; that is, \( d \) was defined as \( \ln c_p - \ln c_o \), which approximated a normal distribution for the pdf model. The data were divided into \( \xi_s \) intervals with the GM and GSD of \( c_p/c_o \) and the uncertainty in the GM computed for each.

Figure 11 shows the GM of \( c_p/c_o \) as a function of \( U/w_s \). A perfect model would yield GM = 1 for all \( U/w_s \) intervals, and therefore, a departure of the GM from 1 indicates a bias or model error. The uncertainty in the GM is represented by the innermost horizontal bars on the vertical lines through the points, and the GSD is shown by the extremities of the lines.
Figure 11 shows that both models err toward underprediction in the highest $U/w_{*}$ category. This was expected for the pdf model since it was intended only for fully convective conditions that exist for $U/w_{*} < 6$ (Willis and Deardorff 1976); that is, there is no treatment of mechanically generated turbulence, which is important for $U/w_{*} > 6$. The model was evaluated for the entire $U/w_{*}$ range to determine when the error would become significant; this occurs for $U/w_{*} > 7$. For CRSTER, the underprediction for $U/w_{*} > 5$ probably was caused by its bias toward neutral stability when convective turbulence or a mixture of convective and mechanical turbulence existed as inferred from $U/w_{*}$. In the smallest $U/w_{*}$ category ($<1.2$), the overprediction of $c_{p}$ by the pdf model was probably due to the neglect of axial dispersion (i.e., in the mean wind direction), which is important in light winds (see Willis and Deardorff 1976).

Figure 12 shows $c_{p}/c_{o}$ as a function of $X$. As can be seen, the pdf model err toward overprediction for $X < 1$, a result probably caused by light winds and the neglect of axial dispersion. The decrease in the GM of $c_{p}/c_{o}$ with $X$ for $X < 4$ may be due to the increase of $w$ with $z$ in the region $z \leq 0.2h$, a feature neglected in the model. The $w$ variation would result in a longer time and distance for a plume element to reach the surface (Hanna et al. 1986). For CRSTER, the most significant bias occurs for $X < 1$, where it underestimates $c_{o}$ by about a factor of 2 or more.

From Figs. 11 and 12, we observe that the GSD and the uncertainty in the GM are noticeably smaller for the pdf model than for CRSTER. This is also true for other residual plots, for example, $c_{p}/c_{o}$ versus $F_{*}$, and is consistent with the relative scatter for the two models shown in Fig. 10.

In summary, we have shown the importance of correlation and residual plots in distinguishing model performance. We highly recommend their use in judging model physics and in diagnosing model strengths and weaknesses. The best models should be based on an overall impression from the correlation and residual analyses and on the scientific content of the models. We reiterate the importance of the plume-direction information in evaluating models (see Appendix). At this point, we hesitate to recommend a single performance measure. Clearly, the best model should be the one with the lowest bias or error (e.g., GM - 1) and

![Diagram](image-url)
the greatest confidence that the error is indeed small; this might suggest the GM of \( c'/\sigma_c \) over all comparisons and its uncertainty limits. Finally, model evaluation should be an iterative process in which the deficiencies found in a residual analysis are corrected, and a modified and hopefully improved model is then evaluated.

b. The fluctuating concentration

The preferred residual analysis requires \( \sigma_c \) models and the pdf of \( c'/\sigma_c \). In anticipation of further development and use of such models, we suggest ways of evaluating them and obtaining the \( c'/\sigma_c \) distribution.

1) Models for \( \sigma_c \)

Laboratory experiments probably offer the best opportunity for testing \( \sigma_c \) models. To date, such models have been evaluated using wind-tunnel experiments of dispersion in neutral boundary layers; they have been conducted for both elevated sources (Fackrell and Robins 1982a; Netterville 1979; Sykes et al. 1986) and surface releases (Wilson et al. 1982). For an elevated source, the results support the meandering-plume model (Fig. 7). With the exception of Hanna (1985), there have been no attempts to evaluate models using the convection-tank experiments of Deardorff and Willis (1984, 1988). To our knowledge, these are the only experiments simulating concentration fluctuations in the CBL.

Further model development and evaluation require more experiments, as has been recommended previously (Weil 1985; Wyngaard 1984). In particular, wind-tunnel experiments should be conducted to simulate time-averaged (~1 h) concentration fields including sufficient repetitions to determine \( \sigma_c \). In addition to providing direct measurements of \( \sigma_c \), for model evaluation, the experiments would improve our understanding of the \( \sigma_c \) dependence on averaging time. These experiments can and should be done for buoyant as well as nonbuoyant effluents. Similar experiments also should be conducted in a convection tank. One detail that merits careful attention is the effect of instrument time response on the measured \( \sigma_c \) (Kristensen et al. 1989).

Some field experiments have been conducted to determine \( \sigma_c \) but they generally have been limited to passive scalar releases from near-surface sources and to short downwind distances, \( x \ll 1 \) km (Dinar et al. 1988; Hanna and Insley 1987; Mylne and Mason 1991; Safford 1987). As discussed earlier, the main difficulty with field experiments is generating a large number of repetitions under the same conditions to determine representative ensemble properties, \( C \) and \( \sigma_c \).

2) Probability distribution of \( c'/\sigma_c \)

A residual analysis can be simplified if the probability distribution of \( c'/\sigma_c \) is self-similar, that is, if it is independent of the model inputs (section 4A). In the following, we first test for self-similarity of the cdf, \( \Omega(c'/\sigma_c) \), using the Deardorff and Willis (1988) laboratory data and then compare the cdf of residuals from two field experiments with these data.

Figure 13 shows the cdf of \( (c - C)/\sigma_c \) from the Deardorff and Willis data; again, we note that the cdf’s are for in-plume concentrations with \( c > 0.01Q/\hbar^2 \) (Fig. 6). We computed \( C, \sigma_c \), and the cdf’s from histograms of \( c \) given by these authors; their histogram for \( F_c = 0 \) is given in Fig. 6 of this paper. We assumed \( c'/\sigma_c \ll 1 \), which means that the fluctuation \( c' = c - C \approx c' \). The measurements were obtained photographically using a fluorescent dye, a fast detection method. The concentration averaging time was assumed to be less than 5 s, the minimum time interval between photographs (Deardorff and Willis 1988), which was less than the convective time scale in the experiments, \( h/w_* = 22 \) s. Thus, the maximum av-

![Fig. 13. Cumulative distribution function (cdf) of concentration fluctuation nondimensionalized by root-mean-square concentration for an elevated source (\( z_* = h = 0.1 \)) in a laboratory convectively mixed layer. Measurements obtained for various dimensionless distances \( x \) in the region \( |y| < \sigma_c \) and conditioned by \( c > 0.01Q/\hbar^2 \); cdf’s computed from histograms in Deardorff and Willis (1988). (a) nonbuoyant source, (b) buoyant source.](https://example.com/figure13.png)
eraging time was 0.23h/w, which corresponds to 2.3 min for a prototype value of h/w = 10 min. The concentrations were near-centerline values being obtained within the crosswind distance interval 0 < |y|/σ < 1 and spanning the downwind distance range 0.2 < X < 5.

Figure 13a for a nonbuoyant plume suggests that the cdf’s from the different distances collapse to a nearly single curve for c'/σc > −0.5, while for smaller c'/σc values, they exhibit a weak X dependence. The scatter in the cdf’s for Ω ≥ 98% precludes a confirmation of the collapse at the very highest percentiles. Figure 13b shows the cdf’s for a highly buoyant plume (F* = 0.26) and the same distance range. Again, the cdf’s collapse to a nearly single curve for c'/σc > −0.5 and show an X dependence for smaller c'/σc. Aside from enhanced scatter at Ω ≥ 98% in Fig. 13a, the cdf’s for the two cases collapse to essentially the same curve for c'/σc > −0.5. Thus, in this c’ regime, Ω is independent of plume buoyancy as well as distance; that is, it is self-similar.

Figure 14 shows the cdf of (c – C)/C for the same data. For the nonbuoyant plume (Fig. 14a) and Ω > 50%, there is more scatter among the cdf’s at different distances than in Fig. 13a. This suggests that σc is the better variable for scaling c’, at least in the upper part of the cdf. For the highly buoyant plume, there is little difference in the cdf’s for the two scalings; see Figs. 13b and 14b. Figure 14 also shows that there is a good overlap of the cdf’s for the buoyant and nonbuoyant plumes although there is clearly less scatter in the buoyant case. This could be due to the different measurement heights in the two experiments: the measurements were taken in the interval 0 < z/h < 0.2 for F*= 0 and in 0.4 < z/h < 0.6 for F*= 0.26.

The suggested collapse of the c'/σc cdf at the high end of the distribution in Fig. 13 also occurs for the cdf of the gamma pdf [Eq. (2.5a)] over the same σc/C range, 0.6 < σc/C < 1.6. Figure 15 presents gamma cdf’s for three σc/C values in this range and shows that all reduce to approximately the same result for c'/σc > −0.5. Thus, for in-plume fluctuations and over limited σc/C values, one can expect a collapse in the cdf of c'/σc. (For c'/σc > −0.5 in Fig. 15, the greatest differences between the cdf’s occur at the high percentiles, but even at Ω = 99%, the c'/σc value for the three curves only varies from 3 to 4, that is, a ±14% deviation about the mean of approximately 3.5.)

The foregoing results motivated us to compute the cdf’s of the scaled residuals d/C from field observations (Fig. 10) and to compare/contrast them with the cdf of c'/C from the laboratory data (Fig. 14). Ideally, we should scale d and c' by σc in comparing the two datasets, but we had no simple σc model for the buoyant-plume data in Fig. 10. For these data, d (=c_o – c_p) was obtained from comparisons of the pdf and CRSTER models with the observed GLCs at the Kincaid and Maryland power plants; we assumed C = C_o and c'_o = 0. The field and laboratory data are marked by a large difference in the concentration averaging time with the field value being about 1 h and the prototype value for the laboratory case being about 2 min. As a result, we did not expect general agreement between the c'/C and d/C cdf’s.

Figure 16 shows the cdf’s for the two quantities. For the pdf model, the d/C cdf is approximately the same at the two field sites (except for Ω > 95%) and overlaps with the laboratory data (c'/C) in the upper part of the distribution, Ω > 80%. For this model, we attribute the differences between the d/C and c'/C cdf’s to the differences in the averaging time. The cdf should be narrower for the longer averaging time (i.e., the field data; see Mylne and Mason 1991), which appears to
Fig. 15. Cumulative distribution function of concentration fluctuation nondimensionalized by root-mean-square concentration for the gamma distribution; gamma pdf given by Eq. (2.5a).

be the case (Fig. 16) at least for the pdf model–Kincaid dataset. Another difference between the $d/C$ and $c'/C$ cdf’s is that the field data include concentration uncertainties ($c'_i$) due to model input uncertainties. The $c'_i$ should broaden the $d/C$ cdf but further analysis is necessary to quantify this effect.

A comparison between the model and laboratory cdf’s for more similar concentration averaging times could be made using the original Maryland data, which were obtained with a mobile van. The approximate 1-h concentration is an average of approximately six “short-duration” measurements of $c$ along the mean plume centerline. The duration is comparable to the instrument averaging time, approximately 15–30 s, which is about 0.12–0.25 times the maximum prototype averaging time (2 min) for the laboratory data. The comparison of the cdfs for the short-duration field data and the laboratory measurements should be pursued.

Figure 16 also shows the cdf of $d/C$ for the CRSTER model, which is quite different from the laboratory data and the pdf model. As can be seen, the CRSTER cdf is quite broad with $d/C$ exceeding 4 in about 45% of the cases. This is caused by the model bias mentioned earlier—zero $c_p$ values due to predicted plume penetration of an elevated inversion or incorrect stability classification.

In summary, we showed that the distribution of $c'/\sigma_c$ from laboratory experiments approximates a self-similar curve in the upper part of the cdf and along the plume center plane; this result applies to in-plume concentrations with $0.6 \leq \sigma_c/C \leq 1.6$. The approximation is probably less accurate near the plume tails where $c'$ is more susceptible to measurement errors, $c'_o$. An exploratory investigation comparing the cdf of $c'/C$ and $d/C$ from the laboratory and field, respectively, showed an overlap in the upper part of the cdf’s; the differences in the cdf’s were attributed primarily to the differences in the concentration averaging times. The comparison encourages further work along these lines with the aim of establishing better relationships between these cdfs and explaining the large residuals in model–field–data comparisons. In addition, it motivates theoretical and experimental work to investigate the effects of averaging time and instrument response on the cdf of $c'/\sigma_c$.

c. The probability distribution of $c$

In a series of papers, Lewellen et al. (1985, 1988) and Lewellen and Sykes (1989) developed an advanced applied model that deals with the stochastic nature of dispersion; we refer to this as the ARAP model. In addition to the mean concentration, they predict $\sigma_c$ and the probability distribution of $c$, $\Omega(c)$, downwind of a source. Here $\Omega(c)$ is assumed to be a clipped normal distribution which is characterized by $C$ and $\sigma_c$. Both of these variables are predicted from a Gaussian plume model with the model for passive (nonbuoyant) contaminants given by Sykes et al. (1986) and the one for buoyant effluents by Lewellen et al. (1988). The modeled $\sigma_c$ for a 1-h–averaged concentration is obtained from the $\sigma_c$ for an instantaneous sample using a variance reduction factor based on an exponential auto-

Fig. 16. Cumulative distribution function of concentration fluctuation nondimensionalized by ensemble-mean concentration for an elevated source in a convective boundary layer. Shaded region represents laboratory data from Fig. 14. Points (field data) are differences (residuals) between observed ground-level concentrations along the plume centerline and predicted ensemble-mean concentrations $\bar{C}$ from the pdf and CRSTER models; residuals computed from data in Fig. 10.
correlation of concentration fluctuations (Sykes et al. 1989; Venkatram 1979). The autocorrelation time scale in this calculation is empirically estimated (see Sykes et al. 1989).

The ARAP model is evaluated by comparing predicted and observed probability distributions. However, the usual statistical comparison between these distributions cannot be made because every observation is a single sample from a different distribution; that is, each \( c_0 \), \( c_p \) pair from a different ensemble as defined by \( \xi_n \). Lewellen et al. (1985) proposed to deal directly with the finite-sample problem by assessing the confidence with which a model could be rejected as less than perfect. The evaluation criterion is one of rejection because any observation is theoretically possible if there is a continuous cdf extending to infinite concentrations. A model is rejected on the grounds that the observations are simply too unlikely if the model is assumed to be correct.

Confidence limits on the expected residual distribution are obtained by generating an ensemble of \( d \) distributions using the predicted cdf's of \( c \). For each sampler (or monitor) location and hour, a random selection of \( c \) is made from the predicted concentration cdf to give a "pseudo-observation" and a "pseudoresidual," which is the difference between this \( c \) and the predicted mean concentration \( C_p \) for that sampler and hour. By doing this at each sampler, one produces a set of pseudo-observations and pseudoresiduals for the entire sampling network, and from these sets one can generate a cdf of pseudo-observations and pseudoresiduals for the network. Typically, this random selection process is repeated 100 times to create 100 realizations of the cdf's for the network. The expected concentration and residual distributions from this process are taken as the average of these 100 realizations of the \( c \) and \( d \) cdf's. Confidence limits on the expected distributions are then obtained by choosing two curves such that 95% of the pseudo-\( c \) and pseudo-\( d \) distributions lie between them. If the observed \( c \) and \( d \) distributions lie outside these limits, then one can reject the model at the 95% confidence level. Further details concerning this approach are given by Sykes et al. (1989) and Lewellen and Sykes (1989).

An example is given here for a model of the mean and variance of the concentration based on SCIMP (second-order closure integrated model plume) (Lewellen and Sykes 1989). In addition to the fluctuations in \( c \) due to boundary-layer turbulence (\( c'_p \)), the model accounts for the uncertainty in \( c_p \) (i.e., \( c'_p \)) caused by uncertainties in the meteorological inputs. Figure 17a shows the cdf of residuals from the Kincaid dataset with about 20,000 1-h SF6 GLCs together with the estimated 95% confidence limits on the predicted distribution. With this large number of samples, the confidence limits are rather narrow in the middle 90% of the cdf, and therefore it is easy to detect model errors. As can be seen, the observed residual generally falls within the confidence limits. However, the limits broaden in the tails of the distribution, particularly on the positive side; this means that the large residuals are very uncertain.

For the 1-h GLCs at the Kincaid plant, more than 50% of the total variance \( \left( \langle d^2 \rangle \right) \) is due to \( c'_p \), that is, caused by meteorological input uncertainty. For the \( c'_p \) contribution, the greatest source is the uncertainty in the horizontal wind (Lewellen and Sykes 1989) or wind direction. This is discussed in the Appendix, where we find that the transport direction uncertainty is \( \sim 0.5 \nu_p / U \) and results in an uncertainty in the lateral plume position of the order of \( \sigma_p \) (for the hourly averaged plume). (Recall that the wind-direction uncertainty was removed in the analyses in section 5b.)

The pseudo-observation cdf along with its confidence intervals can be obtained for the cdf of \( c \) and evaluated using the observed distribution. This is a less demanding test since we do not compare predictions with observations paired in space and time. Figure 17b shows a comparison between the predicted and observed \( \Omega(c) \) for the same data and model as in Fig. 17a. Again, we see that the extreme values of the cdf are very uncertain quantities; for example, see the 99.99th-percentile value.
One limitation with this evaluation approach is that one can check only for consistency between observations and predictions, and thus, a model may appear consistent due to a large confidence interval. The size of the confidence interval depends on the quality and completeness of the model input data and on the accuracy of the modeled $\sigma_e$. A model that gives a good prediction with a narrow confidence interval is clearly better than an equally good prediction with a wide confidence interval (Sykes et al. 1989).

Before closing, we note two problems with the ARAP evaluation procedure. The first is the difficulty of accounting for correlations between the observations. The curves in Figs. 17 were generated by choosing the realizations independently at each sampler location, whereas in practice, nearby observations will be affected by the same large-scale eddies and are therefore not independent. The second is that the residuals in Fig. 17a should be made dimensionless since they include samples from different ensembles. Based on our previous discussion, $\sigma_e$ would be the logical choice. However, this requires careful thought for low-concentration regions, such as the plume tails, where the measurement error can be a significant contribution to the residual.

d. Relevance to regulatory practice

In this section, we discuss 1) the uncertainty in predicted second-highest concentrations, 2) use of the uncertainty in estimating the probability of attaining air-quality standards, and 3) the site variability in model performance.

1) UNCERTAINTY IN PREDICTED SECOND-HIGHEST CONCENTRATIONS

In regulatory practice, the focus is on model performance in predicting the annual second-highest concentrations over 1-, 3-, and 24-h averaging periods at stationary monitors. These concentrations correspond to the 99.99th, 99.97th, and 99.73d percentiles of the cdf. Due to their rarity, the predictions are highly uncertain, as shown by the confidence limits in Fig. 17. Knowledge of the uncertainty is necessary to evaluate model performance and to estimate the probability of attaining an air-quality standard. It is informative to compare confidence limits predicted from the ARAP approach with performance results obtained for the CRSTER model.

Turner and Irwin (1982) compared CRSTER model predictions of the second-highest SO$_2$ GLCs over one-year periods with measurements around seven power plants. The predictions and observations were reported for 3- and 24-h averaging periods; only the 3-h averages are discussed here. From the Turner and Irwin data, we computed the GM and GSD of $c_p/c_o$ for both the spatially paired and unpaired $c_p, c_o$ values. The results were restricted to cases with $\Delta E/z_h < 0.7$, where $\Delta E$ is the terrain elevation difference between the monitor and stack base. This restriction is consistent with the CRSTER limitations discussed by Turner and Irwin.

Sets A1–A3 in Table 1 present the results for the spatially paired data. They show that the model is essentially unbiased when averaging $c_p/c_o$ over all plants (see the “GM” column) since the confidence limit on the GM includes unity. The confidence limit was estimated from a lognormal distribution having the computed GSD and number of points $N$ shown. There is little variation in the results with $\Delta E/z_h$; this may be due to the sparse dataset and/or a sufficiently large GSD ($\sim 1.6$) to mask such variations. However, one can see that the 95% confidence limits on $R = c_p/c_o$ are large with a ratio of $R_{95}/R_{GM}$ of 2.6–2.8. This is partially due to the variation of the GM with plant; the GM ranged from 0.6 at one plant to 2.1 at another.

### Table 1. Ratio of predicted ($c_p$) to observed ($c_o$) ground-level concentration around power plants. For datasets A and B, $c_p$ and $c_o$ are the second-highest SO$_2$ concentrations from a year of monitoring data; for set C, $c_p$ and $c_o$ are the 99.99th-percentile values of SO$_2$ concentrations from 20,000 sample hours of data at the Kincaid Power Plant.

<table>
<thead>
<tr>
<th>Model data reference</th>
<th>Dataset</th>
<th>$c_p, c_o$ space paired</th>
<th>Averaging time (h)</th>
<th>$\Delta E/z_h$</th>
<th>$c_p/c_o (R)$</th>
<th>95% confidence limit$^b$ on GM</th>
<th>$R_{95}/R_{GM}^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRSTER; Turner and Irwin (1982)</td>
<td>A1</td>
<td>Yes</td>
<td>3</td>
<td>$&lt;0.7$</td>
<td>0.36–3.35</td>
<td>0.95</td>
<td>1.62</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>Yes</td>
<td>3</td>
<td>$&lt;0.6$</td>
<td>0.36–3.35</td>
<td>0.90</td>
<td>1.68</td>
</tr>
<tr>
<td></td>
<td>A3</td>
<td>Yes</td>
<td>3</td>
<td>$&lt;0.4$</td>
<td>0.47–3.35</td>
<td>0.95</td>
<td>1.68</td>
</tr>
<tr>
<td>CRSTER; Turner and Irwin (1982)</td>
<td>B</td>
<td>No</td>
<td>3</td>
<td>$&lt;0.7$</td>
<td>0.58–2.58</td>
<td>0.99</td>
<td>1.55</td>
</tr>
<tr>
<td>SCIMP; Lewellen and Sykes (1989)</td>
<td>C</td>
<td>No</td>
<td>1</td>
<td>0</td>
<td>0.1</td>
<td>1.3</td>
<td>nr</td>
</tr>
</tbody>
</table>

$^a$ $\Delta E =$ terrain elevation difference between monitor and stack base; $z_h =$ stack height.
$^b$ $\text{GM} =$ geometric mean of $c_p/c_o$; $\text{GSD} =$ geometric standard deviation of $c_p/c_o$.

$^c$ The confidence limit is calculated from an assumed lognormal distribution with the GSD shown and $N$ points; thus for the GM, $R_{95}/R_{GM} = \text{GSD}^{1.6}/100$, and for $c_p/c_o$, $R_{95}/R_{GM} = \text{GSD}^{1.6}/100$, and for $c_p/c_o$, $R_{95}/R_{GM} = \text{GSD}^{1.6}/100$.

$^d$ $R_{95} = c_p/c_o$ at upper 95% confidence limit; $R_{GM} =$ GM of $c_p/c_o$.

$^e$ nr = not reported.
The results for set B (Table 1) are for the spatially unpaired $c_p$ and $c_o$. They show that the model is effectively unbiased when averaging over all plants (see GM), but there is a large variation from plant to plant; see the column “Range.” The confidence interval on $c_p / c_o$ is reduced somewhat by comparison to the paired data, but it is still broad—$R_{95}/R_{GM} \sim 2.4$.

Set C shows the results from the Lewellen and Sykes (1989) analysis where we have estimated the ratio $R_{95}/R_{GM}$ from the 99.99th percentile in Fig. 17 assuming that the GM is 1; this percentile value approximates that of the second-highest value in the Kincaid dataset. The important point is that this ratio ($\sim 1.7$), while smaller than that found for the CRSTER model (set B) by comparing predictions with observations, is at least comparable to it. In addition, this ratio is essentially the same as found for individual plants in the Turner and Irwin dataset where multiple years of observations were available. This demonstrates that modeling of $\sigma_c$ and $Q(c)$ can provide credible estimates of the confidence limits of the extreme concentrations at fixed monitors.

In summary, we have shown that the ARAP model gives confidence intervals on second-highest concentrations that are comparable to those found for the CRSTER model by comparing $c_p$ with $c_o$. The main advantage of the ARAP model is that it does not require observations of $c$. This is important when applying models to new problems where monitoring data do not exist. Our analysis of the Turner and Irwin (1982) data shows that there is a slightly greater uncertainty in predicting high concentrations at a given monitoring station (set A) than in estimating the highest value over a network (set B). This should be considered in applications where one must predict the spatial distribution of high concentrations, for example, for the prevention of significant deterioration standards, which limit the increment in $c$ due to new sources.

2) USE OF UNCERTAINTY IN PREDICTING ATTAINMENT OF STANDARDS

The GSD of $c_p / c_o$ for the second-highest concentration is important in estimating the probability that an air-quality standard will be satisfied. The probability is given by

$$\Omega(c_2 < c_t) = \int_0^{c_t} p(c_2) dc_2$$

where $c_t$ denotes the second-highest concentration, $p(c_2)$ is its pdf, and $c_t$ is the standard. Assuming that $p(c_2)$ is lognormal, that is, Eq. (2.4) with $c_t$ equal to the GM of $c_2$ and $\sigma_t = \ln(GSD)$, we find

$$\Omega(c_2 < c_t) = \frac{1}{2}(1 + \text{erf} \zeta)$$

where

$$\zeta = \frac{\ln(c_t / c_2)}{\sqrt{2} \ln(\text{GSD})}$$

and erf is the error function (Carslaw and Jaeger 1947). Here, it is assumed that the model is unbiased—the GM of $c_p / c_o$ is 1. For a predicted $c_t = 0.8 c_t$, $\Omega$ is 80% and 66% for a GSD = 1.3 and 1.7, respectively; these GSD values correspond to the range shown in Table 1. If $c_t / c_2 = 0.7$, then $\Omega$ is 94% and 78% for a GSD of 1.3 and 1.7.

Clearly, the probability of attainment increases as $c_t / c_2$ decreases. In addition, the probability depends on the GSD of $c_p / c_o$—the model uncertainty at the percentile of the standard—the higher the GSD, the lower is the probability that the standard will be satisfied. Thus, uncertainty information is important for regulatory applications and should be useful to air-quality managers.

3) SITE VARIABILITY OF MODEL PERFORMANCE

The operational performance of models generally varies with the site, and this requires careful examination to determine systematic modeling errors that would be important in new source applications. For example, from the CRSTER evaluation by Turner and Irwin (1982), we found that the GM of $c_p / c_o$ for the second-highest 3-h GLC ranged from 0.6 to 2.1 at different power plants; these differences were statistically significant. Although the range could not be linked to any particular cause, we found that the largest GM occurred for the plant with the shortest stack (81 m) and possibly the lowest buoyancy flux; the GM range was 0.6–1.2 for the remaining plants where $z_t$ was typically 200 m.

A clearer demonstration of a model error was shown in Hanna and Paine's (1989) evaluation of the MPTER model (Pierce and Turner 1980), which is similar to CRSTER. Hanna and Paine evaluated MPTER and a new model HPDM (hybrid plume dispersion model) using hourly averaged $SF_4$ GLCs around the Kincaid and Bull Run (Tennessee) power plants. They compared predictions and observations of $c_t / Q$ averaged over the 25 highest values. At Kincaid, the models performed similarly with the average predicted $c_t / Q$ within 4% of the observed value. However, at Bull Run their performance was much different—for MPTER, the predicted $c_t / Q$ was less than one-half the observed value, whereas it was only 14% too low for HPDM. The variation in the MPTER performance was attributed to the much lighter winds typically found at Bull Run ($\sim 2$ m s$^{-1}$) than at Kincaid ($\sim 5$ m s$^{-1}$). This led to many more MPTER predictions of zero GLCs due to plume penetration of the elevated inversion, a situation discussed earlier for CRSTER (section 5a; Fig. 10). In contrast, HPDM was designed to deal with this situation and performed much better.

In summary, these examples show the range of results that can be found in operational performance evaluations. To narrow the range, we recommend that more evaluations with routine monitoring data be conducted for a wide variety of sites and dispersion.
climatologies, that is, sources with differing climatological records of winds, PBL depth, surface heat flux, etc. This activity should be closely coordinated with a model physics evaluation using intensive datasets.

6. Summary and recommendations

Over the past decade, substantial efforts have been made to evaluate air-quality models. Most, including those of the EPA, have focused on operational performance—the assessment of model accuracy in predicting the high concentrations at fixed monitors. This paper stemmed from our belief that this practice needs to be expanded to both evaluate model physics and deal with the large natural variability in concentration.

The variability is defined as the rms fluctuating concentration \( \sigma_c \) about the ensemble-mean concentration \( C \). For \( \sigma_c \approx C \), large residuals (\( d \)) exist between \( c_p \) and \( c_o \) which confuse model evaluations. Our main purpose was to review and recommend ways of evaluating model physics in light of the large \( \sigma_c \) and other sources of concentration variance (section 4). We focused on point-source models, concentration averaging times of \( \sim 1 \) h, and downwind distances less than about 30 km.

In evaluation, it is important to recognize that concentration is a random variable and should be described statistically by a probability distribution. The distribution can be parameterized by \( C \) and \( \sigma_c \), but dispersion models currently used in applications predict only \( C \). Laboratory experiments and models are the principal means of determining \( \sigma_c \), since this quantity is difficult to obtain in the field. However, \( \sigma_c \) models are in the developmental stage and more work is necessary before they can be used routinely in applications.

The current emphasis on operational performance evaluation is driven by regulatory applications and air-quality standards. The comparisons receiving most attention are those in which \( c_p \) and \( c_o \) are unpaired in space or time; that is, they are not drawn from the same ensemble (section 3). These evaluations are useful to air-quality managers, but they are not tests of model physics. Moreover, they can be misleading about model performance under different source conditions and dispersion climatologies (section 5d).

To evaluate model physics, we must separate the mean model error, \( C_o - C_p \), from the natural variability and other sources of concentration variance for a well-chosen ensemble of realizations. The ensemble is defined by the source and meteorological variables characterizing dispersion, which are given by the model inputs \( \xi_n \). We showed that a residual analysis is an effective way of isolating the model error. In the preferred approach, the dimensionless residual \( d/\sigma_c \), averaged over \( \xi_n \) intervals, should be zero and therefore show no trend with \( \xi_n \) (section 4a). Currently, the need for \( \sigma_c \) limits use of this approach. Therefore, we discussed an alternative—the residual is defined as \( d = \ln c_p - \ln c_o \), which has an approximate normal distribution for the point-source problem (section 4b).

Several examples were discussed to show the effectiveness of the residuals in identifying model errors, suggesting their probable causes, and distinguishing the performance of two models (sections 4b and 5a). The best model should be the one with the lowest bias or error and the greatest confidence that the error is small. In summary, we highly recommend the use of residual analyses as well as correlation plots and laboratory data in evaluating model physics.

Our review also suggested that evaluation of mean (\( C \)) models should not be viewed as testing model physics or operational performance to the exclusion of the other. Several examples showed the benefits of conducting these evaluations together (sections 3a, 4b, and 5d). Thus, we recommend that they be pursued simultaneously.

Further development and application of \( \sigma_c \) models is regarded as a future activity. We discussed ways of evaluating the models and determining the cdf of the random fluctuation, \( c_p/\sigma_c \), using laboratory data (section 5b). It was found that laboratory data from point-source plumes in the CBL approximated a self-similar cdf of \( c_p/\sigma_c \), a useful result in residual analyses. In addition, the cdf’s of \( d/C \) from model and field-data comparisons and \( c_p/C \) from the laboratory experiments exhibited some similarity as expected. This is encouraging and motivates further investigations along these lines.

At present, there is one advanced modeling approach that predicts \( C \), \( \sigma_c \), and the probability distribution of \( c \), \( \Omega(c) \), downwind of elevated sources—the ARAP model (e.g., Lewellen and Sykes 1989; section 5e). Here \( \Omega(c) \) is assumed to be a clipped normal distribution, and the \( C \) and \( \sigma_c \) necessary to parameterize it are predicted from a generalized Gaussian plume model. The ARAP model is evaluated by comparing predictions and observations of \( \Omega(c) \) and requires confidence limits of the predicted distribution, which are obtained by statistical resampling. The model is more computer intensive than typical dispersion models (for \( C \) only) and requires further work to demonstrate it as a practical tool. However, it generates an important quantity for applications—the uncertainty or confidence limits of the very high and rare concentrations.

The relevance of the uncertainty or confidence limits to regulatory applications was discussed in section 5d. The confidence limits predicted by the ARAP model for the second-highest concentrations, which are needed in applications, were comparable to those obtained for the CRSTER model by comparing \( c_p \) with \( c_o \). Predictions of these confidence limits are important for new sources where monitoring data do not exist. We also showed the use of the confidence limits in estimating the probability of attaining an air-quality standard. Probability estimates provide a rational way of assessing air-quality impact and are highly recommended for use in future assessments.

Acknowledgments. We wish to thank L. A. Corio for providing the concentration probability distribution of
the field data shown in Fig. 16; G. A. Briggs, J. F. Clark, and J. S. Irvin of the Environmental Protection Agency and B. L. Sawford for their thorough and constructive reviews of an earlier draft of this paper; W. H. Cox, J. A. Tikvart, and J. C. Wyngaard for helpful comments; L. Kristensen for providing the subroutine for evaluating the incomplete gamma function (for Fig. 15); and F. B. Huth for editorial assistance. We also thank anonymous reviewers for their constructive comments. This paper was prepared under a cooperative agreement between the Environmental Protection Agency and the American Meteorological Society. J. C. Weil is grateful for support from the U.S. Army Research Office under Contract DAAL03-90-G-0127 and from the Electric Power Research Institute under Contract RP1616-32 during the period of this work.

APPENDIX

The Uncertainty in Plume Transport Direction

In the evaluation discussed in section 5a, the uncertainty in the plume transport direction was removed by comparing the maximum $c_{\alpha}$ and $c_{\beta}$ on each sampling arc. Knowledge of the uncertainty is important when predicting concentrations at a few fixed monitors under the assumption that the plume direction $\alpha_p$ is equal to the hourly averaged wind direction $\alpha_w$. To determine the adequacy of this assumption, we calculated the rms deviation $\sigma_\alpha$ between $\alpha_w$ and $\alpha_p$:

$$\sigma_\alpha = \left[ \frac{1}{N-1} \sum (\alpha_p - \alpha_w - \Delta \alpha)^2 \right]^{1/2} \quad (A.1)$$

where

$$\Delta \alpha = \frac{1}{N} \sum (\alpha_p - \alpha_w), \quad (A.2)$$

and $\alpha_p$ was chosen as the angular position of the maximum concentration on an arc. Nine periods from the EPRI Kincaid experiments were analyzed; each was about 4 h long and contained about 16 arc-hours. All of the measurements were obtained in the CBL.

Figure A1 shows $\sigma_\alpha U/w_*$ as a function of $T_a U/h$ where $T_a$ is one hour. On average, $\sigma_\alpha \sim 0.5 w_*/U$, which is close to the ensemble-averaged rms wind direction, $\sigma_{\alpha_W} = 0.6 w_*/U$. That is, the fluctuation in the "instantaneous" wind direction. At short range where the plume width $\sigma_y$ grows linearly with $x$ (i.e., $\sigma_y = \sigma_{\alpha_W} x$), the uncertainty in the lateral plume displacement ($\sigma_{\alpha_W} x$) from the mean wind axis is of the order of $\sigma_y$. Farther downwind where $\sigma_y$ grows somewhat less rapidly with distance, the displacement should exceed $\sigma_y$. Thus, in models where one assumes $\alpha_p = \alpha_w$, the uncertainty in the plume position is large.

For reference, we estimated the variance in the time-averaged wind direction from Eq. (2.13) by replacing $\sigma_y^2$ by $\sigma_{\alpha_W}^2$ and assuming an exponential autocorrelation function for direction; that is, $\rho(t') = \exp(-t'/T_E)$, where $T_E$ is the Eulerian time scale. This estimate applies to stationary, homogeneous turbulence. With these substitutions, $\sigma_\alpha$ is given by

$$\frac{\sigma_\alpha U}{w_*} = 0.85 \left[ -\frac{T_a}{T_a} (e^{-T_a/T_E} - 1) \right. + \left. \left( \frac{T_a}{T_a} \right)^2 (e^{-T_a/T_E} \left( \frac{T_a}{T_E} + 1 \right) - 1) \right]^{1/2}. \quad (A.3)$$

By assuming that the eddy lateral length scale is proportional to $h$ and its translational speed is $U$, we can approximate $T_E$ as

$$T_E = \frac{b}{U} h \quad (A.4)$$

where $b$ is a dimensionless coefficient.

Equation (A.3) is shown in Fig. A1 for $b = 1$ and $10$, and in both cases $\sigma_\alpha$ tends to $0.6 w_*/U$ at small values of $T_a U/h$ as it should. A comparison of this equation with the data suggests that the lateral length scale is of the order of 10h. This is probably true even though the assumptions leading to Eq. (A.3) are probably not satisfied, that is, the real boundary layer is nonstationary and nonhomogeneous. However, the main point is that $\sigma_\alpha$ is large and should therefore result in large uncertainties in predicted GLCs at stationary monitors.

REFERENCES


—, and G. E. Willis, 1984: Groundlevel concentration fluctuations from a buoyant and a non-buoyant source within a laboratory convectively mixed layer. *Atmos. Environ.*, 18, 1297–1309.


Venkatram, A., 1979: The expected deviation of observed concentrations from predicted ensemble means. Atmos. Environ., 13, 1547–1549.
—, and B. W. Simms, 1985: Exposure time effects on concentration fluctuations in plumes. Dept. of Mechanical Engineering, University of Alberta, Edmonton, Canada, Report No. 47.