EPA-650/4-75-010

STATISTICAL QUESTIONS RELATING TO THE VALIDATION OF AIR QUALITY SIMULATION MODELS

by

Glenn W. Brier

Consultant
1041 North Taft Hill Road
Fort Collins, Colorado 80521

Program Element No. 1AA009
ROAP No. 21 ADO

EPA Project Officer: Kenneth L. Calder

Meteorology Laboratory
National Environmental Research Center
Research Triangle Park, North Carolina 27711

Prepared for

U.S. ENVIRONMENTAL PROTECTION AGENCY
OFFICE OF RESEARCH AND DEVELOPMENT
NATIONAL ENVIRONMENTAL RESEARCH CENTER
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

March 1975

EPA REVIEW NOTICE

This report has been reviewed by the National Environmental Research Center - Research Triangle Park, Office of Research and Development, EPA, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into series. These broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and maximum interface in related fields. These series are:

1. ENVIRONMENTAL HEALTH EFFECTS RESEARCH
2. ENVIRONMENTAL PROTECTION TECHNOLOGY
3. ECOLOGICAL RESEARCH
4. ENVIRONMENTAL MONITORING
5. SOCIOECONOMIC ENVIRONMENTAL STUDIES
6. SCIENTIFIC AND TECHNICAL ASSESSMENT REPORTS
9. MISCELLANEOUS

This report has been assigned to the ENVIRONMENTAL MONITORING series. This series describes research conducted to develop new or improved methods and instrumentation for the identification and quantification of environmental pollutants at the lowest conceivably significant concentrations. It also includes studies to determine the ambient concentrations of pollutants in the environment and/or the variance of pollutants as a function of time or meteorological factors.

This document is available to the public for sale through the National Technical Information Service, Springfield, Virginia 22161.

Publication No. EPA-650/4-75-010

STATISTICAL QUESTIONS RELATING TO THE VALIDATION OF AIR QUALITY SIMULATION MODELS

I. Introduction. This study examines techniques that can be used in evaluating the predictive accuracy of air quality models and discusses some of the problems of comparing predicted versus measured values. It considers the statistical basis for some of these techniques and their associated figures of merit, scores or indices; and recommends a specific validation procedure to be followed. The study examines the effect of the inaccuracies in the input and output data used in the validation process, and offers some suggestions regarding the major problem of separating input-output data errors from those introduced by a poor mathematical representation of the physical and chemical processes.

II. Background. In the past few years the Environmental Protection Agency (EPA) has supported a major effort in the development of air quality models. The models we are concerned with are deterministic physically based relationships between emissions and ambient air quality, with a varying degree of formalism on the turbulent diffusion process and its resulting mathematical description. The model is a tool used by forecasters responsbile for short-term predictions as well as by control officials and planners to indicate the impact of proposed changes in such things as emission quantity, patterns and the like. A major problem has been the lack of suitable data for performing model tests, especially for the more complex models which appear to be promising but need improved emissions inventory and meteorological information if they are to be more useful. To help remedy this situation, the EPA is currently sponsoring a comprehensive Regional Air Pollution Study (RAPS). The RAPS, described by Ruff and Fox [1974], concentrates mainly on providing vast amounts of data of high quality which along with a much improved emissions inventory will result in a large base of data to be used in the development and validation of improved air quality models. Thus, it now seems appropriate to consider some of the ways in which this data base can be used effectively for validating models and to explain the significance and implications of recommended validation procedures.

III. Evaluation and Validation. - The Problem. Often the verification and scoring of predictions is controversial. Among the reasons may be the failure to objectively and quantitatively define the quantities to be compared or to agree on a scale of "goodness" to measure the difference between the predictions and

observations. Another reason may be the failure to define clearly the purpose or purposes of evaluation. Generally speaking, these are:

(1) To determine the correspondence between predictions and observations -- constituting a scientific, empirical or inferential evaluation;

(2) To determine the value of the predictions to "decision makers" constituting economic operational or decision-theoretic evaluation.

This study is concerned with the first of these objectives, and the agreement of a model with observations is referred to as validity. If the agreement is good, the model (or theory) is considered to be true, although it is generally recognized that the word true may be misleading since any model is at best an approximate description of reality. Once an agreement is reached on a scale to be used in measuring the goodness of the prediction, an absolute (but usually arbitrary) score or figure of merit can be defined to characterize this agreement (or lack of it). This score may be used to compare the performance of different models, or the performance of the same model under different circumstances or for different locations. Such a score may be useful in helping to make a choice between models, but the relative ranking given to different models may depend upon the particular score used, and the criticism can be made that the model that verifies best according to some arbitrary scoring system may not be the most useful model. In this study we will attempt to show how some of these difficulties might be avoided, or at least alleviated.

There are a number of statistical quantities that can be used in determining the correspondence between predictions and observations. But it is important to recognize the stochastic nature of the predictions and that validation statistics computed from the sample of data are only estimates subject to considerable fluctuations. The statistically conscious investigator realizes that however an experiment or observational program actually turned out, it could have turned out somewhat differently. By means of an appropriate statistical analysis, he attempts to make a valid assessment of the uncertainty of the results in terms of a probability statement or by setting confidence limits. The estimates discussed in this report are consistent in the probability sense, i.e. as the sample size increases, the estimates converge in probability to the parameters they are estimating. In actual practice the sample sizes are likely to be quite small, and estimates of the sampling variances are necessary if valid conclusions are to be drawn.

A validation procedure does not need to be limited to a comparison of the predictions with observed values. A good statistical analysis should have diagnostic

value, yielding clues about which parts of the model (or observations) may be responsible for the errors uncovered. A complex Air Quality Simulation Model (AQSM) contains many modules (for emissions, transport and diffusion, transformations, and removal). The input to one of the modules might be the output of one of the other modules, or a submodel might be used to process the basic data to be used as an input. Final model validation must be based upon validation of model components, but unfortunately in many cases the data are inadequate to do this. However, when data are available to separately validate a component of a model, the general principles to be followed are the same as for validating the final output. The techniques discussed below are applicable in either case.

a. The mean square error. For the purpose of a statistical summary, the mean square error (MSE) or the root-mean-square-error (RMSE) is often used. For a series of n predictions it is defined as

$$MSE = 1/n \sum_i d_i^2 \quad , \tag{1}$$

where d_i is the difference between prediction X_i and the corresponding observation Y_i. If the d_i are normally distributed then all the information about the frequency distribution of errors is contained in the statistics $\bar{d}$ and s_d, where

$$\bar{d} = 1/n \sum_i d_i \tag{2}$$

is a sample estimate of the bias (the tendency to over-predict or under-predict) and s_d is a sample estimate of the population standard deviation σ_d and is defined by

$$s_d = [1/n \sum_i (d_i - \bar{d})^2]^{1/2} \quad . \tag{3}$$

The mean absolute difference

$$1/n \sum_i |d_i|$$

is commonly used and may have some advantages or disadvantages in comparison with the MSE. (This will be discussed later.)

If the data are not normally distributed, the use of the MSE or mean absolute difference can be quite misleading, especially when comparing predictions and observations of short-term concentrations of pollutants, since the distributions are

likely to be non-normal with long "tails." Thus certain precautions must be taken, and this will be discussed in more detail in a later section. However, even if the prediction errors are normally distributed, the MSE does not tell the whole story for additional information can be obtained by considering the component parts representing various sources contributing to its value. It is easy to show that

$$MSE = \bar{d}^2 + s_X^2 + s_Y^2 - 2rs_Xs_Y \quad , \tag{4}$$

where

$$s_X^2 = 1/n\sum_i (X_i - \bar{X})^2 \quad , \tag{5}$$

$$s_Y^2 = 1/n\sum_i (Y_i - \bar{Y})^2 \quad , \tag{6}$$

and

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{[\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2]^{1/2}} \quad , \tag{7}$$

where $\bar{X}$ and $\bar{Y}$ represent the means of the predictions and observations respectively. For perfect predictions we must have

$$\bar{d} = 0 \quad , \tag{8}$$
$$s_X^2 = s_Y^2 \quad ,$$

and

$$r = 1 \quad .$$

The statistics s_x, s_y, r and $\bar{d}$ may be useful quantities in diagnosing the errors in the predictions and might provide useful indices for comparing models.

For the correlation to be near unity and the MSE to be near zero, both X and Y must be essentially free of error. We can consider the model prediction X as made up or a perfect prediction X* and an error term. Contributions to the error can come from imperfections in the model as well as from errors in the input data.

Let

$$X = X^* + m + e \quad , \tag{9}$$

where X^* is the prediction from the perfect model, m is the output error produced by the model, and e is the contribution from errors in the input variables, which may be original data or estimates derived from the output of a submodel. If the model is applied in a number of different circumstances (e.g., different receptor locations, different days, etc.), the set of predictions will have a variance σ_X^2, and if m and e are uncorrelated

$$\sigma_X^2 = \sigma_{X^*}^2 + \sigma_m^2 + \sigma_e^2 \tag{10}$$

where σ_m^2 and σ_e^2 are the variances of m and e respectively. Since X^* represents a deterministic prediction rather than a random variable, $\sigma_{X^*}^2$ represents the variance of the output of the perfect model as it is applied to different circumstances.

Likewise, there is some <u>true</u> observation Y^*, so that for a perfect prediction $Y^* = X^*$. The observed Y is given by

$$Y = Y^* + \varepsilon \tag{11}$$

where ε is the observational error. If Y^* and ε are uncorrelated then

$$\sigma_Y^2 = \sigma_{Y^*}^2 + \sigma_\varepsilon^2 \quad , \tag{12}$$

where σ_ε^2 is the error variance of the observations and $\sigma_{Y^*}^2$ represents the variation in the set of Y^* over time or space. The correlation coefficient between X and Y can be expressed as

$$\rho_{XY} = [1 + (\sigma_m^2 + \sigma_e^2)/\sigma_{X^*}^2]^{-1/2} \, [1 + \sigma_\varepsilon^2 / \sigma_{Y^*}^2]^{-1/2} \tag{13}$$

Thus the correlation cannot be unity unless $\sigma_m^2 = \sigma_e^2 = \sigma_\varepsilon^2 = 0$.

If the random errors m, e and ε in (9) and (11) have zero means the bias $\bar{d}$ will be zero. However, if $\bar{d} \neq 0$, it is not possible to determine the source of this bias by examining X and Y only, and additional independent information must be obtained. This problem is discussed in more detail in Section V.

If it is possible to get estimates of σ_e^2 and σ_ε^2 (as discussed later), then it is of interest to consider the index

$$I = (s_X^2 - \hat{\sigma}_e^2)/(s_Y^2 - \hat{\sigma}_\varepsilon^2) \tag{14}$$

where $\hat{\sigma}_\varepsilon^2$ is the estimated variance of the errors in the observations Y and $\hat{\sigma}_e^2$ is the estimated variance of errors in prediction due to data input errors. The index I is an estimate of $(\sigma_{X*}^2 + \sigma_m^2)/\sigma_{Y*}^2$, and for large n should be close to unity for a very good model, since by definition $\sigma_m^2 = 0$. This is a necessary, but not sufficient condition. Thus, if $I = 1$, the model might be good, since the variances of X and Y are the same. If $I \neq 1$, the model cannot be perfect since $\sigma_m^2 \neq 0$ and $r<1$. Thus, if approximate estimates of σ_e^2 and σ_ε^2 can be obtained, then the index I might be helpful in separating input-output data errors from those introduced by a poor physical model.

b. Regression analysis. In model validation a graphical representation by means of a scatter diagram can be made showing the relationship between the model predictions X and observations Y, as illustrated in Figure 1.

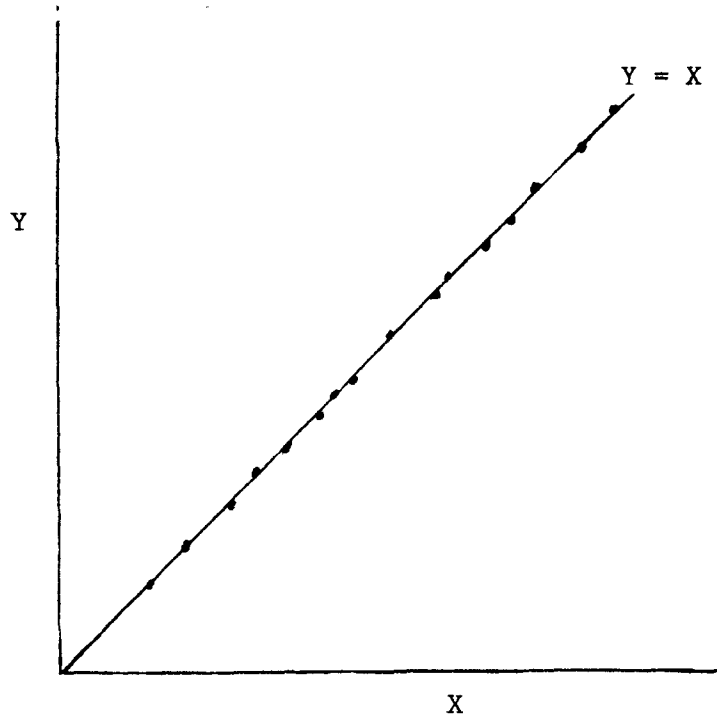

Figure 1. Graph showing relation between predictions X and observations Y.

If both X and Y are essentially free from error, the points will lie along the line $Y = X$ (or very close to it). If there are appreciable errors in X and Y, the line $Y = X$ will no longer give a good fit, but there will be another line

$$\hat{Y} = A + BX \tag{15}$$

(called the regression line) which will give a better fit. This line, with slope B and intercept A, is determined by the method of least squares, which minimizes the sum of squares of the deviations from the regression. This line has been used for model "calibration" or tuning, and although this procedure has been examined and criticized by Brier [1973], there is considerable merit in giving attention to the coefficient B in validation studies. The formula for the slope is

$$B = [\sum_i (X_i - \bar{X})\ (Y_i - \bar{Y})]\ [\sum_i (X_i - \bar{X})^2]^{-1} \tag{16}$$

In relation to data input and model error, it can be shown to be represented by

$$B = [1 + (\sigma_m^2 + \sigma_e^2)/\sigma_{X*}^2]^{-1} = rs_Y/s_X \tag{17}$$

Thus, if B is close to unity, it means that $\sigma_m^2 + \sigma_e^2$ is small relative to σ_{X*}^2 and is suggestive of a good model if the sample size is sufficiently large. Errors in the observation Y do not produce a bias in B but will affect its sampling distribution. Therefore, the slope B becomes a very meaningful statistic in a validation procedure, especially when σ_ε^2 becomes small. A good model must have B close to unity.

IV. The use of robust techniques. The standard correlation and linear regression procedures discussed above are based on a mathematical model where a number of assumptions are made. Some of the important ones are as follows:

(i) The regression line is linear.

(ii) The distribution of Y for a given X is Gaussian (or at least approximately).

(iii) The variance of the departures from the regression line is constant.

In many cases it is likely that these assumptions will not be valid for comparing predicted and observed value of pollutant concentrations. Inferences about means and variances will be sensitive to departures from assumptions such as error normality, especially in the case of short-term concentrations where observed values may vary by an order of magnitude or more. The RAPS modelling effort will be more concerned with short-term concentrations, such as predicting hourly averages, but even in long-term models where input and output errors have a chance to balance out, it is still important to determine the effects of departures from the basic assumptions. When these effects are appreciable it is desirable to use robust (resistant) statistical procedures. For example, instead of the MSE the median error can be used, or the value that is exceeded (say) only 10% of the time. Although it may be desirable to present the frequency distribution of error, a summary statistic is usually needed for making comparisons between different models, meteorological conditions or receptor locations. Transformation such as the logarithmic may be useful since percentage errors may be relevant. However, a 50% error at high concentrations may be more important than a 50% error at low concentrations where the measurements may be close to the background or noise level. Other transformations of the type X^{β} ($\beta<1$), for example, may be more useful in stabilizing the variance or reducing the influence of a few extremes or outliers -- that may not be representative.

In the case of correlation and regression analysis, a few outliers can sometimes have a dominant influence, perhaps even reversing the sign of the correlation. Graphical methods can be very useful in detecting outliers. For example, Figure 2 shows the plot of predicted and observed monthly-mean concentrations for eight Chicago stations in January 1967 as reported by GEOMET [1972]. The point number 4 has the largest prediction error, and if it is removed the calculated regression line will be very close to the line $Y = X$. New methods of robust regression estimation have been developed (see Hogg [1974]) and extended to the multivariate case where graphical methods may not be so effective in identifying outliers or determining their influence. A recent analysis of air pollution data from New Jersey and New York by Cleveland and Kleiner [1974] illustrate the usefulness of robust statistical procedures and graphical methods. A recent article by Hawkins [1974] discusses the use of principal components in detecting errors and outliers in multivariate data. This is not to say that all outliers should arbitrarily be removed or ignored, but the fact that they can be flagged makes it possible to give them further study and examine their influence on the estimates of summary

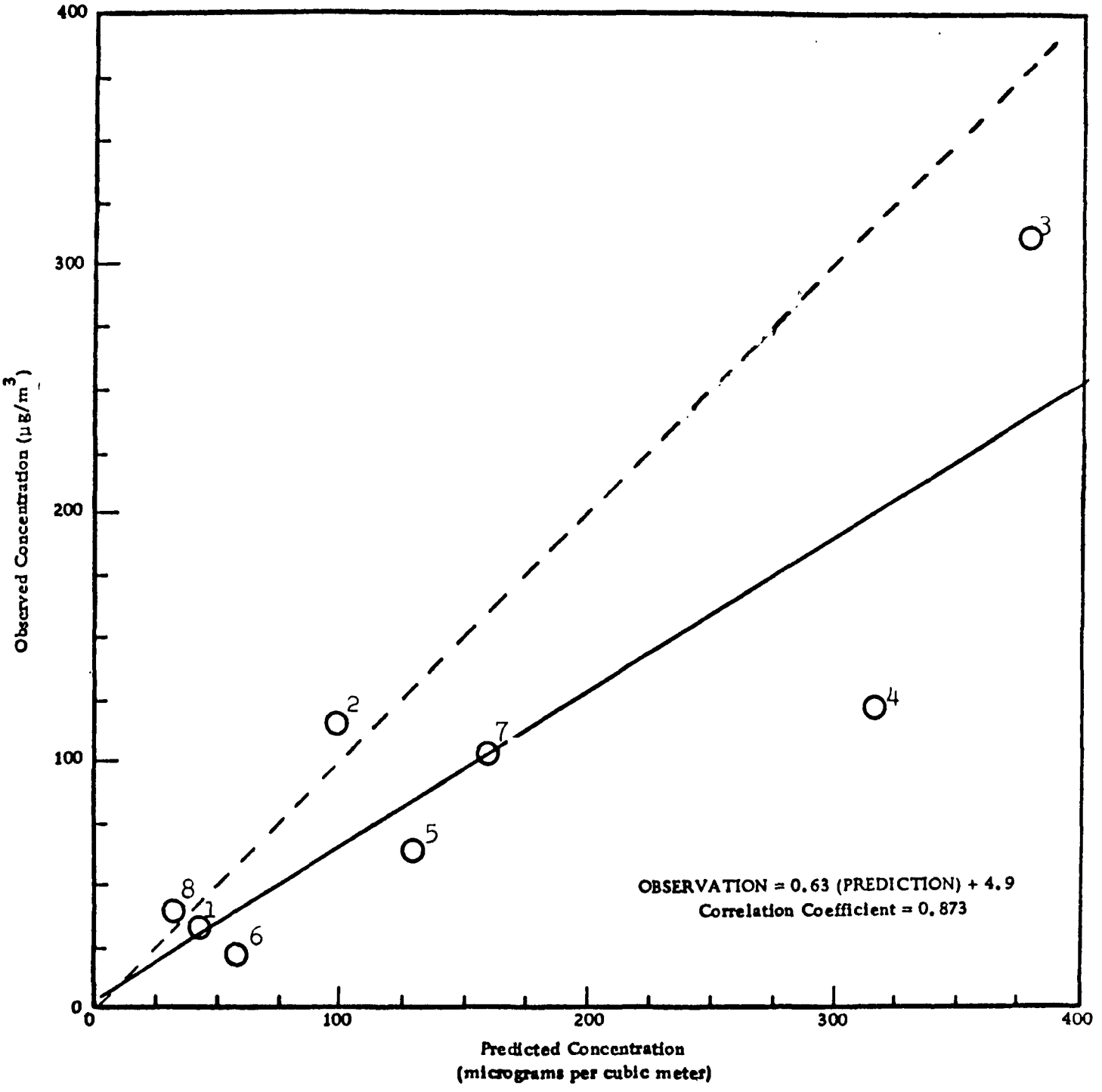

Figure 2. Regression Analysis of Monthly Mean Concentrations for Eight Chicago Stations (January 1967)

(Adapted from GEOMET, 1972)

statistics such as MSE, s_X^2 , s_Y^2 , B and r .

V. Input-output errors and the validation process.

One of the problems in the use of validation techniques has been the relative inattention given to the effect of input-output data errors on the validation process and to the related problem of separating these errors from those introduced by deficiencies in the model. Some mention of this has been made earlier in Section III with a brief discussion of some of the sources of error. The problem of erroneous output data is relatively simple since contributions from this source do not bias the regression coefficient B and for a sufficiently large sample it could be possible to validate a good model even though there might be considerable error on individual observations of concentration. Furthermore, the RAPS should provide a large amount of high quality data to minimize this problem. However, the problem of erroneous input is a much more serious one, and for the RAPS to be helpful here it must provide information on individual input errors as well as on the complete structure of the errors, involving not only their relationship with each other but with the model inputs. This section attempts to provide a general framework showing how this error information along with sensitivity analysis and model simulations might help to provide a solution to some of the problems.

To deal with the question of separating the effects of input errors from model errors is essentially the problem of distinguishing the relative contributions of m and e in (9). Since this is a fairly technical section we shall begin by establishing some useful notation.

The type of model we are dealing with can be thought of as a function f (say) which maps a point $\underline{Z}$ called the input vector, in Euclidean n -space, onto a real number X , called the output or prediction. We will deal in this section only with univariate output, so that we may write

$$X = f(\underline{Z}) \ , \ \underline{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_n \end{pmatrix} .$$

In our case $\underline{Z}$ contains such things as meteorological variables, locations and strengths of pollution sources and sinks, and coordinates of recording station (or stations). X is the predicted concentration.

The input vector $\underline{Z}$ has error $\underline{z}$ so that $\underline{Z} = \underline{Z}^* + \underline{z}$ where $\underline{Z}^*$ represents true input values. With this notation we may write expressions for m and e of (9) as

$$m = f(\underline{Z}^*) - X^* = f(\underline{Z}^*) - Y^* \quad ,$$

$$e = f(\underline{Z}) - f(\underline{Z}^*) \quad . \tag{18}$$

(These expressions cannot be used to compute $m + e$ unless $\underline{Z}^*$ and Y^* are known, in which case we have no problem. Most of our troubles arise from the unknowability of $\underline{Z}^*$).

Since m and e are random variables, the maximum information available is their joint distribution. We are going to have to settle for considerably less than this. The minimum information needed to make useful inferences on any random variable seems to be some estimate of its first two moments, mean and variance. This is what we will try for. Population means will be denoted by μ's and variances by σ^2's with subscripts denoting which random variable is being considered. Thus,

$$\mu_e = E(e),$$
$$\sigma_e^2 = E(e - \mu_e)^2$$

where E is the expectation operator. Likewise, we have

$$\mu_m = E(m),$$
$$\sigma_m^2 = E(m - \mu_m)^2 \quad .$$

Estimates of these quantities will be denoted by $\hat{}$'s . Thus $\hat{\mu}_e$ denotes an estimate for the long-term average value of the error in the output due to input error propagated through the model f .

As mentioned before, we cannot get separate estimates for parameters involving m or e separately by using only the observations Y and the predictions X for given inputs $\underline{Z}$. We require some additional outside information, namely some specific knowledge of the multivariate structure of input errors. We can, however, get some information on combinations of both m and e using only X and Y. To see this, it is easiest to take the case of zero (or negligible) observation errors, i.e. $\varepsilon = 0$ so that $X^* = Y^* = Y$. When this holds we have from (9)

$$m + e = X - Y.$$

Taking expectations:

$$\mu_m + \mu_e = \mu_X - \mu_Y \quad ,$$

and if m and e are assumed uncorrelated (as before)

$$\sigma_m^2 + \sigma_e^2 = \sigma_{X-Y}^2 \quad .$$

We wish our estimates to satisfy these same equations with $\hat{}$'s. We have available good estimates of the right hand sides of the above, namely the difference in sample means $\bar{X} - \bar{Y}$ and the sample variance s_{X-Y}^2 so that

$$\hat{\mu}_m + \hat{\mu}_e = \bar{X} - \bar{Y} \quad , \tag{19}$$

$$\hat{\sigma}_m^2 + \hat{\sigma}_e^2 = s_{X-Y}^2 \quad . \tag{20}$$

(Note: We may wish to replace $\bar{X} - \bar{Y}$ by 0 in (19) if it is known beforehand that the model f is definitely unbiased, i.e., $\mu_m + \mu_e = 0$.) These equations will enable us to estimate μ_m and μ_e separately once we have an estimate of either one singly and similarly for σ_m^2 and σ_e^2.

As a first step in attempting to separate m from e, we need first to look at the structure of the input error. Ideally, we should know the joint distribution $F(z_1, z_2, \ldots z_n)$ of the input errors. As a very minimum we require the mean vector $\underline{\mu}_z = E(\underline{z})$ and covariance matrix $\Sigma_{\underline{z}} = \mathrm{Cov}(\underline{z})$ of the input error vector.

First suppose we know (or have a reasonably good estimate of) $F(z_1, z_2, \ldots z_n)$. The numerical procedure is as follows:

(1) Break the n-dimensional input space into a number (N say) of subregions $R_1, R_2, \ldots R_n$, in such a way that the model $f(\underline{Z})$ is deemed to be reasonably constant within each subregion (i.e. the response of the model to the input variable is essentially uniform over the range covered by the subregion). This requires the knowledge obtained from a good sensitivity analysis of the model (such as the GEOMET [1972] study).

Note: If the sensitivity study shows that the model is essentially constant over the whole range of some particular input, considered singly and in combination with others, it should probably be eliminated as an input, and its (constant) effect be included as a parameter in the model.

(2) Determine (by numerical integration if necessary) the probability content of p_j of each R_j :

$$p_j = \int\int_{R_j} \ldots \int dF(z_1, z_2, \ldots z_n) \quad .$$

In this respect, it is highly recommended that the regions $R_1, \ldots R_n$, be N-dimensional rectangles if possible. If this is the case, then p_j's can be obtained by properly combining the values of $F(\underline{z})$ at the corners of the rectangles. This also provides each region with an easily computed centroid or "representative point," $\underline{\zeta}_j$ say. In any case we require some "representative point" for R_j.

What we have achieved by the above process is actually a discrete (N point) approximation of the probability density of the "input error" term e defined by (9). At value $f(\underline{\zeta}_j)$ we have probability of occurrence p_j . Whatever departures from this occur in the distribution of Y must be attributed (in the absence of observational errors) to the model error m .

We can now write down the estimates:

$$\hat{\mu}_e = \sum_{j=1}^{N} f(\underline{\zeta}_j)\, p_j \quad ,$$

$$\hat{\sigma}_e^2 = \sum^{N} [f(\underline{\zeta}_j) - \hat{\mu}_e]^2\, p_j \quad .$$

We also have

$$\hat{\mu}_m = \begin{cases} \bar{Y} - \bar{X} - \hat{\mu}_e & \text{(in general)} \\ -\hat{\mu}_e & \text{(if model is known to be unbiased),} \end{cases}$$

and

$$\sigma_m^2 = s_{X-Y}^2 - \hat{\sigma}_e^2 \quad .$$

This last equation assumes:

(1) No observation error $(\sigma_\varepsilon^2 = 0)$,

(2) No correlation between m and e $(\rho_{me} = 0)$.

If we do not know the distribution function F , or cannot get a reasonable estimate for it, but have the other information mentioned above, $\mu_{\underline{z}}$, $\Sigma_{\underline{z}}$, then one could make a kind of zero$^{\underline{th}}$ order first approximation to obtaining the needed results, assuming n-variate multi-normality, $\underline{z} \sim MVN\ (\mu_j, \Sigma_{\underline{z}})$ and proceeding with the above program on that basis. Previous experience with multivariate normal distributions indicates that whatever inferences are made with respect to the first two moments of the resulting distribution they are likely to be conservative but probably not too wildly bad, even if the real distribution departs rather markedly from multivariate normality.

The question of observation errors when they are present is fairly routine. We must obtain by separate means (e.g. by duplicate measuring instruments at some sites) estimates of μ_ε and σ_ε^2 . Then, under the assumption that ε is uncorrelated with both m and e (probably a good assumption), we replace $\bar{X} - \bar{Y}$ with $\bar{X} - \bar{Y} - \hat{\mu}_e$ and s_{X-Y}^2 with $s_{X-Y}^2 - \sigma_\varepsilon^2$ in the above scheme.

All of the previous discussion has been based on the premise that the errors m and e are uncorrelated. This seems to be a good place to start if one expects to make progress on the problem of separating input-output data errors from those introduced by the model. However, if m and e are correlated, a more complex and troubling question arises that leads to unsolved and perhaps unsolvable problems. In light of (18) and on both physical and mathematical grounds, it may be an unwarranted assumption to state that the correlation between m and e is zero. The consequences of making such an error could be costly in the estimation procedure, and could well lead to a <u>negative</u> estimate for σ_m^2 . That is, it could happen that if we assume no correlation between m and e , when in fact there is some, a situation in which $\hat{\sigma}_e^2 > s_{X-Y}^2$ could lead to the difficulty. (The presence of observation errors in Y could only make the situation worse.)

A closer look at the estimation procedure can show the possible effect of a poor estimate of ρ_{me} .

To avoid tedious notation, let $u = \dfrac{\hat{\sigma}_m}{s_{X-Y}}$, $v = \dfrac{\hat{\sigma}_e}{s_{X-Y}}$, $\alpha = \hat{\rho}_{me}$.

Then the general relation between u , v , α is:

$$u^2 + v^2 + 2\alpha uv = 1 \qquad (21)$$

which reduces to (20) when the estimated correlation, α is zero.

For a fixed α ($-1<\alpha<1$) , (21) represents an ellipse in u,v space. The problem can be illustrated graphically. Let $x = (u+v)/\sqrt{2}$, $y = (u-v)/\sqrt{2}$, representing a 45^o rotation of the axes. (21) then becomes

$$(1+\alpha)\, x^2 + (1-\alpha) y^2 = 1 .$$

If we sketch the various ellipses obtained for different values of α ($\alpha = 0$ gives a circle, $\alpha = \pm 1$ degenerates to a pair of straight lines) we get something like Figure 3. We have shown the situation for $\alpha \geq 0$. For negative α's the ellipses are simply rotated 90^o so that the major axis is along the x-axis, i.e. simply interchange x,y in Figure 3. With this to guide us, we see that a situation in which one of u or v is greater than 1 is not hard to evaluate; it just means we ought to have some negative α . Our estimate u should be placed on an ellipse instead of a circle. A perusal of Figure 3 and some contemplation of what it means quickly convinces one that a good estimate of the correlation between m and e ought to be a prerequisite to estimating σ_m^2 by means of (21). If an independent way could be found to estimate σ_m^2 then (21) can be used to estimate ρ_{me} . At this point it is not clear how this can be done without essentially knowing $\underline{Z}^*$, the true input. Further study will be needed to determine the importance of these questions and whether a solution is available.

VI. A validation procedure. Since model validations are likely to be carried out under a variety of circumstances, with variations in the quality and quantity of data, it does not seem desirable to specify a fixed set of rules to be followed blindly under all conditions. However, certain general guidelines and suggestions can be provided that should be applicable in most cases to give assistance in planning and executing a validation study. As experience is gained in the use of different types of models under different geographical or meteorological conditions, it will become clear that each case has its own particular problems, both physical and statistical, and that modifications are likely to be needed in the usual procedure so that the emphasis can be properly placed on the particular problem or problems at hand. The following discussion will emphasize those areas that need serious attention in any complete and thorough validation analysis.

One of the requirements would appear to be a good sensitivity analysis. Such an analysis should show up any internal inconsistencies in the model and help to

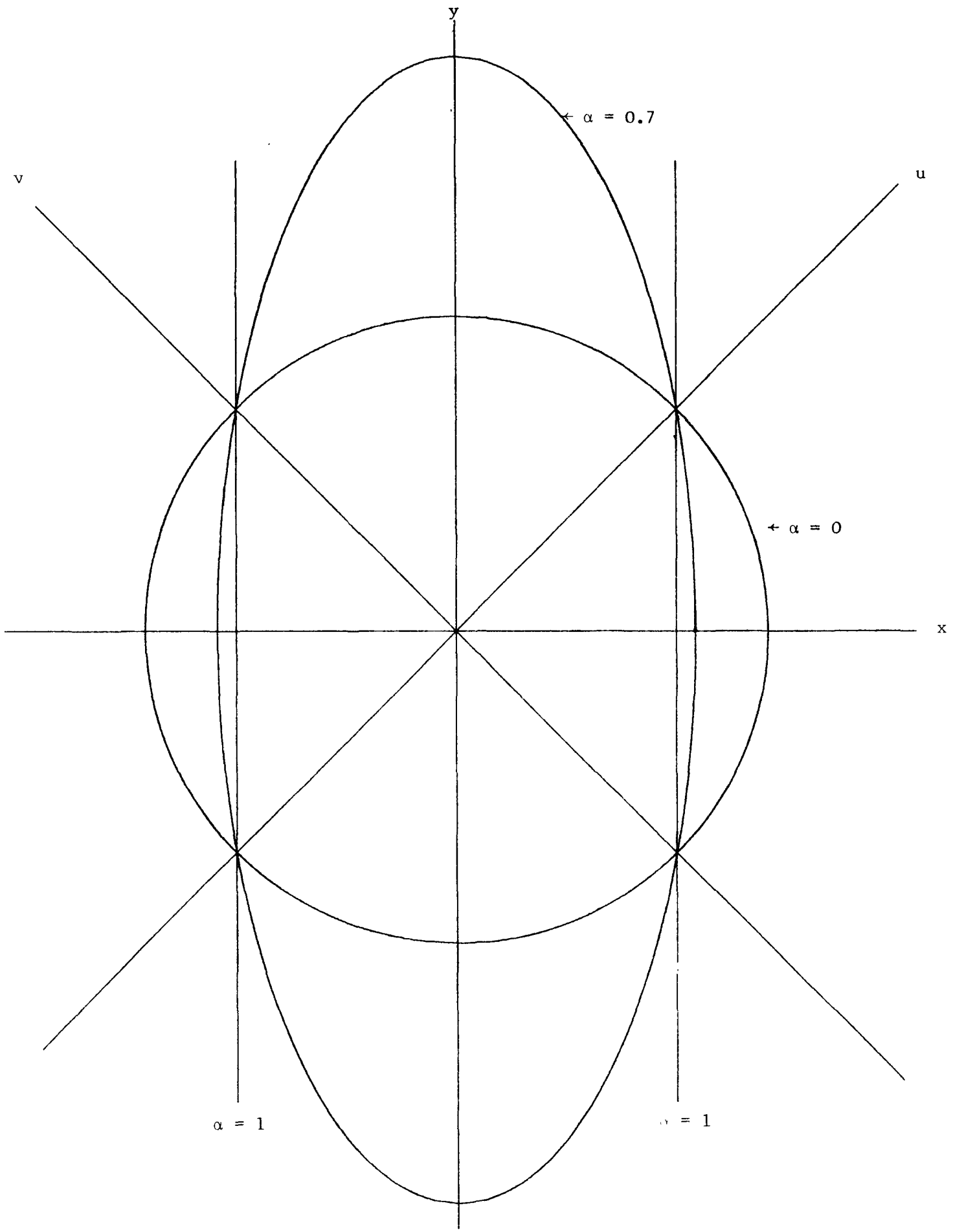

Figure 3

define and understand the real-world parameters which dominate the process. The results should indicate the main areas of interest, especially as related to data collection efforts. The analysis provides information which, along with knowledge of the input error structure, makes it possible to see how input error is propagated through the model. Without this it would not be possible to attack the problem of separating input-output data errors from those introduced by the model, although as discussed in Section V some problems may remain.

After the sensitivity study the MSE and regression analysis would logically follow. The comparison between the predictions X and observed concentrations Y involve various statistics such as estimated correlation coefficients, standard deviations, etc., but none of these is necessarily more important than the others and they can all provide useful information. The standard deviations of the predictions and the observations must be the same for perfect forecasts, but because there are likely to be departures from normality it is desirable to look at the overall distributions of the calculated and observed concentrations. For good prediction the correlation r and coefficient B should be close to unity. The value of r will be lowered if there are errors either in the predictions X or observations Y. It is desirable to have an independent estimate of the error variance of the observed Y's , and since they don't tend to bias the regression coefficient, a comparison of r and B with respect to the output errors might provide information on the influence of model and data input errors. However, as discussed in Section V, the separation of data input errors from model errors is a more complex problem.

In addition to obtaining an estimate of the variance of the observed Y's , one should have an estimate of the mean error or bias in the observation. This must be known if the tendency to over-predict or under-predict ($\bar{X}-\bar{Y} \neq 0$) is to be understood, i.e., whether to attribute a prediction bias to the bias in the observed concentration data or to error in the prediction due to input data or model failure.

As mentioned earlier, the statistics discussed above are related to the mean-square-error, or its square root (RMSE). If the error distributions are normal, it is quite appropriate to use the RMSE. Since in many cases the distributions will not be normal, it is desirable to use additional methods for summarizing the data. The mean absolute difference $|\bar{d}|$ can be computed since it is less affected by departures from normality. More information is provided by presenting a histogram showing the entire error distribution, from which it should be possible to determine the median difference $\tilde{d}$, which will not be affected by a few extreme

values. In addition, other percentile points can be stated which may be meaningful for particular applications.

In the data analysis careful attention should be given to effects of departure from normality on the statistics computed. A few outliers or extremes could have undue influence and invalidate some of the conclusions. Graphical procedures should be used to help in detecting errors, inconsistencies and unusual or unexpected situations. The use of robust techniques and data transformations should be considered when there appear to be appreciable departures from the assumptions usually made in standard statistical analysis.

Input-output data errors have been discussed in Section V where it was pointed out that observational data on the structure of the input errors are needed if there is to be a serious attempt to tackle the difficult problem of separating input-output data errors from those introduced by a poor mathematical representation of the physical and chemical processes. If data on input errors indicate that there are no interactions with each other or with the input variables and that the errors have constant variance over the range of input, then there is no problem. The sensitivity analysis, where one factor is varied at a time, and the numerical simulations using combinations over a reasonable range, should suffice. However, it is nearly certain that the situation will be more complex, and once some data or reasonable information is available on the structure of the input errors one can go ahead along the lines suggested in Section V.

VII. Recommendations and conclusions.

When Air Quality Simulations Models (AQSM) are applied to practical air pollution control or planning problems the user should be provided with an indication of the limitations and accuracy of a particular model in terms that he can understand. The tests used for the validation of a model usually consist of comparisons of the model calculations (X) with observed air pollutant concentrations (Y), from which the distribution of the errors (X - Y) can be obtained and various summary statistics computed which may be used to compare different models or to determine whether a particular complex model is an improvement over a simpler model. A validation process which includes a good statistical analysis can also yield clues about which parts of the model (or observations) may be responsible for the errors uncovered and, if the right data are available, help one to

separate input-output data errors from those introduced by a poor mathematical representation of the physical and chemical processes. Furthermore, it is recognized that the final model validation must be based upon validation of model components, since any AQSM contains main modules (for emissions, transport and diffusion, transformations, and removal) as well as submodels to estimate the required model inputs. Unfortunately the data are inadequate in many cases to separately validate the component parts, but in principle the technique is the same as validating the complete model, i.e., comparing the output (prediction) of the submodel with the measured value. Basically, we have a model (or submodel) which calculates an output X which is in error because of input errors e and model deficiencies that produce errors m. The output X is compared with the directly measured values Y that have errors ε. The error of prediction (X - Y) will be affected by these various sources of error and a thorough validation analysis attempts to identify the nature and source of these errors so that more meaningful comparisons can be made. It is important to recognize that these errors and the stochastic nature of the prediction produces sampling fluctuations in the validation statistics that must be considered before drawing inferences about models or comparisons between models Recommendations for a validation procedure follow.

For a measure of prediction accuracy, it is recommended that the mean square error (MSE) given by (1) be used along with $\bar{d}$, the mean bias of the prediction. If the prediction errors are normally distributed, then the MSE and $\bar{d}$ give all the necessary information about the distribution of errors. However, since the distributions are not likely to be normal, especially where pollutant concentrations are concerned, it is also desirable to obtain the frequency histogram of the forecast errors. From the examination of this distribution one can determine the median error $\tilde{d}$ or the value that is exceeded (say) only 10% of the time. Averages based on percentage errors might be more meaningful in some cases.

Next would come a regression analysis with the computation of the correlation coefficient r, the regression slope B and the intercept A. Each of these statistics gives useful information for making model comparisons and for diagnosing possible sources of error. Also, a comparison between the variance of the predictions (s_X^2) and of the observations (s_Y^2) is essential, but since there are to be departures from normal it is important to look at the overall distributions of calculated and observed concentrations. Since the standard correlation and regression procedures are based on a mathematical model with certain assumptions,

care should be taken to see that there are not sufficient departures from these assumptions to invalidate the conclusions. The use of robust techniques is recommended where such departures are indicated. Graphical and other techniques, including transformations, are suggested for detecting whether a few extreme values or outliers have undue influence on the results.

Attempts should be made to get estimates of the error variance (σ_e^2) of input data and the error variance (σ_ε^2) of the observations (Y). These estimates, $\hat{\sigma}_e^2$ and $\hat{\sigma}_\varepsilon^2$ respectively, can then be used to compute the index I, given by (14), which might indicate how good the model is and be helpful in separating input-output data errors from those introduced by the model. A more complete solution of this problem requires information on the detailed structure of the input errors along with a good sensitivity analysis. The sensitivity analysis is needed to show up any internal inconsistencies in the model, and together with the knowledge of the input error structure makes it possible to attack the problem of separating input-output data errors from those introduced by the model, as discussed in Section V. Proceeding along these lines (under the assumption that the errors m and e are independent) should help to delineate some of the effects of using good models with poor data or vice versa. It is anticipated that RAPS and other programs will soon provide the necessary data.

References

Brier, G. W., 1973: "Validity of the Air Quality Display Model Calibration Procedure," Environmental Monitoring Series, EPA-R4-73-017, Office of Research and Monitoring, National Environmental Research Center, EPA, Research Triangle Park, N.C. 27711.

Cleveland, W.S. and Kleiner, B., 1974: "The Analysis of Air Pollution Data from New Jersey and New York," Paper presented at the Annual Meeting of the American Statistical Association, St. Louis, Missouri, August 26-29, 1974.

GEOMET, 1972: "Validation and Sensitivity Analysis of the Gaussian Plume Multiple-Source Urban Diffusion Model." Final Report prepared under Contract Number CPA 70-94 for Division of Meteorology, Environmental Protection Agency, National Environmental Research Center, Research Triangle Park, N.C.

Hawkins, D.M., 1974: "The Detection of Errors in Multivariate Data Using Principal Components," Journal of the American Statistical Association, 69 (June 1974), 340-344.

Hogg, Robert V., 1974: "Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory," Journal of the American Statistical Association, 69 (December 1974), 909-923.

Ruff, R.E. and Fox, D.G., 1974: "Evolution of Air Quality Models Through the Use of the RAPS Data Base," Paper No. 74-124, National Environmental Research Center, EPA, Research Triangle Park, N.C. 27711.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA-650/4-75-010		

4. TITLE AND SUBTITLE	5. REPORT DATE
Statistical Questions Relating to the Validation of Air Quality Simulation Models	March 1975
	6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Glenn W. Brier	

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
Glenn W. Brier, Consultant 1041 N. Taft Hill Road Fort Collins, CO 80521	1AA009
	11. CONTRACT/GRANT NO. Special Study

12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Environmental Protection Agency National Environmental Research Center Meteorology Laboratory Research Triangle Park, North Carolina 27711	Final
	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This study examines some of the statistical problems that arise in the validation and evaluation of air quality models. It considers the various scores or indices that can be used in measuring the predictive accuracy of a model and shows how the verification statistics are affected by errors in the input and output data and imperfections in the model. Suggestions are made regarding the major problem of separating input-output data errors from those introduced by a poor mathematical representation of the physical and chemical processes, and recommendations are made regarding validation procedures to be followed as the RAPS data base becomes available.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air quality model Evaluation Predictive accuracy Regression analysis Statistical theory Validation Verification	Air Pollution Model Development	

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
Unlimited	Unclassified	24
	20. SECURITY CLASS *(This page)*	22. PRICE
	Unclassified	

EPA Form 2220-1 (9-73)