245

NATIONAL COOPERATIVE
HIGHWAY RESEARCH PROGRAM REPORT

245

METHODOLOGY FOR EVALUATING HIGHWAY AIR POLLUTION DISPERSION MODELS

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM
REPORT 245

METHODOLOGY FOR EVALUATING HIGHWAY AIR POLLUTION DISPERSION MODELS

J. R. MARTINEZ, H. S. JAVITZ, R. E. RUFF,
A. VALDES, K. C. NITZ, and W. F. DABBERDT
SRI International
Menlo Park, California

RESEARCH SPONSORED BY THE AMERICAN ASSOCIATION OF STATE HIGHWAY AND TRANSPORTATION OFFICIALS IN COOPERATION WITH THE FEDERAL HIGHWAY ADMINISTRATION

AREAS OF INTEREST:
ENERGY AND ENVIRONMENT
(HIGHWAY TRANSPORTATION)

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL
WASHINGTON, D.C.

DECEMBER 1981

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

Systematic, well-designed research provides the most effective approach to the solution of many problems facing highway administrators and engineers. Often, highway problems are of local interest and can best be studied by highway departments individually or in cooperation with their state universities and others. However, the accelerating growth of highway transportation develops increasingly complex problems of wide interest to highway authorities. These problems are best studied through a coordinated program of cooperative research.

In recognition of these needs, the highway administrators of the American Association of State Highway and Transportation Officials initiated in 1962 an objective national highway research program employing modern scientific techniques. This program is supported on a continuing basis by funds from participating member states of the Association and it receives the full cooperation and support of the Federal Highway Administration, United States Department of Transportation.

The Transportation Research Board of the National Research Council was requested by the Association to administer the research program because of the Board's recognized objectivity and understanding of modern research practices. The Board is uniquely suited for this purpose as: it maintains an extensive committee structure from which authorities on any highway transportation subject may be drawn; it possesses avenues of communications and cooperation with federal, state, and local governmental agencies, universities, and industry: its relationship to its parent organization, the National Academy of Sciences, a private, nonprofit institution, is an insurance of objectivity; it maintains a full-time research correlation staff of specialists in highway transportation matters to bring the findings of research directly to those who are in a position to use them.

The program is developed on the basis of research needs identified by chief administrators of the highway and transportation departments and by committees of AASHTO. Each year, specific areas of research needs to be included in the program are proposed to the Academy and the Board by the American Association of State Highway and Transportation Officials. Research projects to fulfill these needs are defined by the Board, and qualified research agencies are selected from those that have submitted proposals. Administration and surveillance of research contracts are the responsibilities of the Academy and its Transportation Research Board.

The needs for highway research are many, and the National Cooperative Highway Research Program can make significant contributions to the solution of highway transportation problems of mutual concern to many responsible groups. The program, however, is intended to complement rather than to substitute for or duplicate other highway research programs.

NCHRP REPORT 245

Project 20-18 FY '79
ISSN 0077-5614
ISBN 0-309-03410-8
L. C. Catalog Card No. 82-50421

Price: $8.40

NOTICE

Published reports of the

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

are available from:

Transportation Research Board
National Academy of Sciences
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

Printed in the United States of America.

FOREWORD

By Staff
Transportation
Research Board

This report will be of principal interest to air pollution technologists including meteorologists, statisticians, and highway engineers concerned with air quality measurements and preparation of environmental impact statements. Such technologists who use highway air pollution dispersion models to estimate levels of carbon monoxide associated with proposed projects will find the report, and accompanying computer software, to contain a helpful methodology and data base for evaluating model performance.

Federal and state regulations require that environmental impact statements be prepared for highway projects so that highway decisions are consistent with State Implementation Plans for achieving and maintaining air quality standards. The air quality portion of an EIS usually includes microscale modeling of current and future carbon monoxide concentrations. Several microscale models have been developed; they vary in approach, complexity, accuracy, and cost. The models include simple line-source-oriented Gaussian models as well as more elaborate numerical models. The reliability and predictive accuracy of the models are not well defined because experimental data suitable for model evaluation have only recently become available and also because there is the lack of a standardized methodology for testing and judging model performance. The research described in this report was designed to fill these gaps.

The report contains a comprehensive methodology for assessing the performance of dispersion models used to estimate highway-related air pollution. The methodology is also applicable to other types of models. Application of the methodology requires a substantial data base developed as part of the research. The data base provides carbon monoxide, meteorological, grade and alignment, and traffic data describing a wide range of environmental conditions. A preliminary evaluation of selected models was performed to demonstrate the use of the methodology and data base.

Appendixes E, F, and G of the report document computer software and the data base and have not been published herewith. Potential users of the methodology and data base should request copies of the appendixes and computer tapes from the NCHRP and supply a blank 12-in. diameter and an 8-in. diameter, 9-track tape or equivalent with a density of 1,600 BPI.

CONTENTS

ACKNOWLEDGMENTS

The research reported herein was performed under NCHRP Project 20-18 by the Atmospheric Science Center and the Statistical Analysis Department of SRI International. W. F. Dabberdt, Associate Director of the Atmospheric Science Center was the Principal Investigator responsible for overall project supervision and review.

J. R. Martinez, Senior Research Engineer, was the project leader, and led the data base assembly task and the task on applications of the methodology. H. S. Javitz, Director, Statistical Analysis Department, developed the statistical methodology. R. E. Ruff, Manager, Environmental Analysis and Applications Program, developed the sensitivity analysis methodology. A. Valdes, Statistical Analyst, wrote the computer program and user's guide for diagnostic analysis. K. C. Nitz, Scientific Programmer/Analyst, led the data processing tasks associated with data base assembly and model applications. H. Shigeishi and L. Jones contributed to the data processing tasks.

Sincerest thanks are given to the following researchers who contributed either the dispersion models or the data used in the study: Drs. Robert Lamb and William Petersen of the U.S. Environmental Protection Agency, Dr. D. Chock of the General Motors Corporation, Mr. Walter Frick of the Oregon Department of Highways, Messrs. Earl Shirley and Paul Benson of the California Department of Transportation, Dr. Jerry Bullin of Texas A&M University, and Dr. S. T. Rao of the New York State Department of Environmental Conservation. Special thanks also are extended to Dr. N. Sundararaman and Mr. Steve Albersheim of the Federal Aviation Administration who reviewed the report. Their comments were most appreciated.

METHODOLOGY FOR EVALUATING HIGHWAY AIR POLLUTION DISPERSION MODELS

SUMMARY

National, state, and local environmental regulations require that the air quality impacts of transportation-related projects be analyzed and quantified. To this end, the Federal Highway Administration (FHWA) has issued guidelines to ensure that air quality effects are considered during planning, siting, and construction of highway improvements, so that highway decisions are consistent with State Implementation Plans (SIPs) for achieving and maintaining air quality standards.

The level of carbon monoxide (CO) associated with a given project is a highway-related air quality impact that requires evaluation. In general, it must be determined whether the ambient standards for CO (35 ppm for 1 hour and 9 ppm for 8 hours, not to be exceeded more than once a year) will be satisfied or exceeded as a result of highway improvements. This requirement calls for estimating CO concentrations on both local and areawide scales.

A number of methods of varying sophistication and complexity are used to estimate CO levels. The techniques include simple line-source-oriented Gaussian models as well as more elaborate numerical models. It can be safely asserted that the number of highway dispersion models in use, especially the Gaussian types, is probably quite large, although some models are more popular than others. However, the reliability and predictive accuracy of the models are not well defined because experimental data suitable for model evaluation have only recently become available, and because there has been a lack of a standardized methodology for testing and judging model performance. The research described in this report was designed to fill these gaps.

This work describes a comprehensive methodology for assessing the performance of dispersion models used to estimate highway-related air pollution. The methods are also applicable to other types of models. The objectives of the study are to: • develop methods for evaluating the performance of highway air pollution dispersion models • compile and document a data base to be used to assess model performance • perform a preliminary evaluation of selected models.

Evaluation Methodology

The evaluation methodology encompasses three different types of analyses which examine different, but complementary, aspects of model performance, namely:

1. Accuracy analysis—measures the predictive performance of the model.
2. Diagnostic analysis—identifies conditions associated with inaccuracies in the model's predictions.
3. Sensitivity analysis—quantifies the model's response to uncertainty in the model input data.

Criteria for Model Evaluation. In evaluating the accuracy of a model, the researchers sought to answer the following questions: (1) How well does the model

predict maximum pollution levels? (2) How well does the model predict the number of exceedances of the relevant air quality standard? (3) How well do the fluctuations in predictions follow fluctuations of the observed pollutant levels in time and space? (4) How closely do the computed concentrations approximate the numerical value of the observations?

The first two questions are motivated by the fact that the air quality standards for carbon monoxide (CO) are framed in terms of maximum 1-hour and 8-hour averages not to be exceeded more than a specified number of times in a year. Hence it is important for the model to estimate accurately the maximum concentrations and the number of times the standard is exceeded.

Question (3) is tied to the need to know whether the computed values coincide in space and time with the observations. There is a need to ascertain whether local areas of high concentration are correctly identified and whether time fluctuations of ambient CO are reproduced properly.

The fourth question is concerned with the fidelity with which the predictions reproduce the numerical value of the observations. By analyzing the residuals (observed less predicted) it is possible to examine tendencies to overpredict or underpredict.

A goal of the model evaluation procedure is to obtain a figure of merit (FOM) that is a summary figure that quantifies the model's composite performance in the areas defined by the previous four questions. In general, different statistics are required to measure the model's performance in each area. After evaluating a number of candidate statistics, six were selected to describe the model's accuracy. The six statistics selected are the components of the FOM. Methods for deriving the FOM are described in Chapter Two.

Statistical Measures of Model Performance. The six statistics chosen to measure the four aspects of model performance are either well known or are easily interpreted:

Statistic 1—To measure a model's ability to estimate peak concentrations, the ratio of the average of the largest 5 percent of the predicted concentrations to the average of the largest 5 percent of the observed concentrations has been selected.

Statistic 2—To measure how well a model predicts the frequency of exceedance of the air quality standard, the difference between the predicted and observed proportion of exceedances of the standard has been selected.

Statistic 3, 4, and 5—To measure the coincidence of predicted and observed concentrations, the temporal, spatial, and Pearson's correlation coefficients have been selected.

Statistic 6—To measure the degree of agreement in concentration levels, the root-mean-square error of the observed and predicted concentrations has been selected.

The mathematical definitions of these statistics are provided in Appendix B of this report. Hereafter, the six statistics will be denoted by the symbols S_1 through S_6.

For each of the six statistics, a choice was made between untransformed or log-transformed concentrations. Untransformed concentrations are used in S_1, S_2, and S_6; and transformed concentrations, for S_3, S_4, and S_5.

Treatment of Data Errors. Traditionally, the comparison of prediction and observation has been performed under the assumption that the measurements are correct (if not exact) and that differences between computed and measured values

are the fault of the model. Since the measured concentrations are subject to error, such an assumption tends to place an unfairly heavy burden on the model. The authors have attempted to remedy this situation by explicitly considering the effect of data errors on the six statistics selected to evaluate the performance of the model. The approach is to modify the formulas for the six statistics of interest to include terms associated with data errors.

In general, three sources of error affect the comparison of observed and predicted concentrations: (1) measurement errors (in CO or tracer concentrations), (2) errors in the input to the model, and (3) modeling errors.

The first two sources of error can include both a random component and a constant, or slowly varying, bias. The model output combines the effects of input and modeling errors. Modeling errors are due to the inability of the model to simulate the physical phenomena of interest. Sensitivity analysis can help to estimate the magnitude of the errors associated with imperfect model input, and judicious application of other diagnostic methods can help to identify some of the sources of modeling errors, but the output of the model will nevertheless be afflicted by errors.

The discussion that follows is concerned with the treatment of random errors in the measurement of concentrations. The approach adjusts only for the random error. Thus, it is assumed that the measurements have been adjusted for bias. If the bias is unknown, it will appear as modeling error. The precision, rather than the accuracy, of the data is of concern.

For each of the six statistics previously defined, formulas have been derived that include the effect of random observational errors. Often, the error is specified as being proportional to the true concentration (e.g., the error is ±10 percent). Measurement precision is sometimes specified as a constant bound (e.g., ±3 ppm) rather than as a percentage. These two types of error specification require that the error be modeled in two different ways. A set of six modified performance statistics has been derived for each type of error specification.

Procedure for Computing a Figure of Merit. It is of interest to combine the six component statistics into a single numerical index or figure of merit that characterizes the overall predictive ability of the model.

To compute the FOM, the value of each statistic S_1 through S_6, was transformed to an arbitrary numerical scale ranging from 0 to 10. A value of zero denotes essentially no agreement between observation and prediction, and a value of 10 denotes the best, but not necessarily perfect, agreement. The FOMs for the individual statistics were then combined to obtain the overall FOM.

Two approaches were used to obtain the overall FOM. In the first approach, the overall FOM is a weighted average of the six individual FOMs; the second approach sets the overall FOM equal to the smallest individual FOM. Because comparison of the two approaches indicated that the second scheme is very conservative, the weighted average method is recommended for most uses.

A computer program was prepared that calculates the six statistics and associated FOM. The program also computes confidence intervals for each statistic, and calculates the value of the statistic adjusted for measurement error.

Sensitivity Analysis. The sensitivity analysis methodology defined a general approach to the problem that yields the variance matrix for the predicted concentrations. This matrix is a function of the errors in the input parameters of the model. Hence, the variance matrix of the predictions provides estimates of the uncertainty in the output caused by input errors. Recognizing the uncertain nature of the specification of input errors, a simplified sensitivity analysis procedure was derived from the general approach.

Model Evaluation Data Base

An extensive data base was assembled that provides the model input data and the observations with which model predictions are to be compared. The data base is composed of five data sets provided by:

- General Motors Corporation (GM)
- New York Department of Environmental Conservation (NYS)
- Texas A&M University
- California Department of Transportation (CALTRANS)
- SRI International

The five data sets were uniformly formatted and assembled into a data archive stored on magnetic tape. The archive was extensively documented in a user's guide that is contained in Appendix G of this report.

Applications of Methodology to Existing Models

A preliminary evaluation of six selected models was performed to demonstrate the application of the methodology; four of the models are Gaussian and two are numerical. The evaluation used data for one site in the Texas data base, and included calculating the performance statistics and associated FOM. The effect of observational error on the performance statistics was also investigated. Diagnostic analyses were performed to identify potential modeling problems. Methods of ranking the models in accordance with performance were investigated and illustrated. A simplified sensitivity analysis was also performed.

CHAPTER ONE

INTRODUCTION AND RESEARCH APPROACH

OBJECTIVES OF THE STUDY

This report describes the development and application of a comprehensive methodology for assessing the performance of dispersion models used to estimate highway-related air pollution. Specifically, the goals of the study were to:

- Develop methods for evaluating the performance of highway air-pollution dispersion models.
- Assemble and document a data base to be used to assess model performance.
- Demonstrate the application of the model evaluation procedure by performing a preliminary evaluation of selected models.

Although the focus of this work was on highway-related models, the methods developed are general and thus are also applicable to other types of models.

It is emphasized that this study was not intended to provide a complete, definitive evaluation of the models selected. Instead, the evaluations are only intended to demonstrate the use and interpretation of the methodology. Thus, the performance of the models is incidental to the focus of the study.

RESEARCH APPROACH

Model Evaluation Methodology

The evaluation methodology encompasses three different types of analyses that examine different, but complementary, aspects of model performance, namely:

1. Accuracy analysis—measures the predictive performance of the model.
2. Diagnostic analysis—identifies conditions associated with inaccuracies in the model's predictions.
3. Sensitivity analysis—quantifies the model's response to uncertainty in the model input data.

The accuracy analysis is the principal element of the model assessment process. Six statistics have been defined that describe quantitatively the predictive performance of dispersion models. To define the statistics, model performance was separated into three complementary categories as follows:

1. Ability to predict exceedances of concentration thres-

holds (which may be equal to the ambient air quality standard).

2. Ability to track pollutant levels in space and time.
3. Ability to replicate the numerical value of observed concentrations.

Two statistics measure performance in the first category, three in the second, and one in the third. Ultimately, the six statistics were combined into a single figure of merit (FOM), which is a numerical index that describes the composite performance of the model.

A novel feature of the method is that it takes into consideration the presence of error in the measured pollutant concentrations. This quantifies model performance more fairly because the predictions need not exactly match observations that are inevitably imprecise to some degree. Thus, the formulas for the six statistics have been modified to include the effect of observational error. Without this modification the statistics would tend to yield a pessimistic assessment of model performance. Two types of observational error were treated: errors defined as a percentage and errors specified as a constant bound of the measured pollutant concentration.

A computer program that calculates the six statistics has been prepared. A user's guide to the program is contained in Appendix E of this report.

Diagnostic analyses of model performance can take a variety of forms. The approach used in this report was to provide the analyst with a diverse set of diagnostic tools that can be molded to fit specific needs. To this end, a computer program was prepared for performing diagnostic analyses. The program includes graphical displays and several selected statistical procedures; it is extensively documented in Appendix F of this report.

Sensitivity analysis is used to quantify the variation of the model output due to estimated errors in the model inputs. The approach focused on estimating the error bounds associated with the model input data, calculating a sensitivity matrix, and combining the two to obtain the model output variation. Once the error-induced model output variation is calculated, this variation can be related to the statistics used in accuracy analysis.

Data Base Assembly

A comprehensive data base has been assembled that can be used to evaluate model performance for various highway configurations and meteorological and traffic conditions. The data base includes data from at-grade, above-grade (elevated), and below-grade (depressed) roadways, and is composed of data from measurements programs conducted by SRI International, Texas A & M University, New York State Department of Environmental Conservation, California Department of Transportation, and General Motors Corporation. In general, these data are unique and are distinguished by a fine level of detail in the measurement of meteorological, pollutant, and traffic conditions. The data base assembled for this study uses a uniform format to facilitate access to the data. A user's guide to the data base was prepared and is contained in Appendix G.

Applications

The use of the model evaluation methodology developed in this study was demonstrated by applying it to six selected models, of which four are Gaussian and two are numerical. The model evaluation was performed using a subset of the data base previously mentioned. It is stressed that the model evaluations are solely intended to illustrate the application of the methodology rather than to provide a definitive assessment of the merits of the selected models.

Report Organization

The report is organized into seven chapters and seven technical appendixes that include mathematical derivations and user's guides to the software and data base.

The findings of the study are discussed in Chapters Two through Four. Chapter Two describes the development of the statistical methodology. Sensitivity analysis is discussed in Chapter Three. Chapter Four contains a description of the data base. The application of the methodology is described in Chapter Five. Conclusions and recommendations are given in Chapter Six. Chapter Seven contains an extensive bibliography of model evaluation methods.

Appendixes A through D describe the mathematical details of the statistical methodology. Appendixes E and F, respectively, contain user's guides to the computer programs that calculate the evaluation statistics and that perform the diagnostic analysis. The user's guide to the data base is given in Appendix G.

CHAPTER TWO

FINDINGS—STATISTICAL METHODOLOGY

This chapter presents the findings relative to the statistical methodology used to evaluate air pollution prediction models. Previous literature on model evaluation methods, the methodology of accuracy analysis, and the methodology of diagnostic analysis are discussed.

Computer programs were developed to perform the calculations associated with the accuracy and diagnostic analyses. These programs are described in Appendixes E and F.

LITERATURE REVIEW

An extensive literature review was conducted to examine patterns and trends in model evaluation methodology. The publications reviewed fell into two categories: methodological papers that developed and analyzed evaluation techniques and applied papers that focused on the use of the methods.

Many of the early model-verification efforts used correla-

tion analysis exclusively. Later, the linear regression analysis became widespread and these two analyses became the dominant approaches. The current trend, however, is to use multiple methods to assess model performance.

The methods found in the literature could be categorized as either primarily diagnostic or accuracy evaluation oriented. Diagnostic methods served to identify sources and conditions associated with modeling errors, were qualitative or graphical, and entered into the model building process during the developmental stage. Accuracy evaluation methods summarized overall model performance by assigning a numerical value to the degree of agreement between computation and observation.

The accuracy evaluation methods were further subclassifiable as fidelity, exceedance, or correspondence measures. The fidelity measures quantified the algebraic or percentage differences between the predictions and observations using the natural pairing. The exceedance measures quantified, in the aggregate, the distributional differences between the predictions and observations (e.g., without reference to the natural pairing). The correspondence measures quantified the correlations between predictions and observations.

Very few authors considered the effects of observational error on the model evaluation process or addressed the issue of assigning different weights to various types of modeling errors. None of the sources reviewed considered combining the various performance measures into a single figure of merit.

ACCURACY ANALYSIS

Evaluation Statistics

In evaluating the accuracy of a model, answers were sought to questions regarding how well the model predicts maximum pollution levels and the number of exceedances of the relevant air quality standard, how closely the spatial and temporal fluctuation patterns of predictions and observations coincide, and how closely the predictions approximate the observations. The statistics used to address these questions are called exceedance, coincidence, and fidelity measures, respectively.

In deciding what exceedance, coincidence, and fidelity measures to select, statistics that could not easily be construed as estimates of physically meaningful quantities were excluded from consideration. In addition, it was found useful to distinguish between "macro" and "micro" statistics. The former generic class of statistics is calculated without the pairwise matching of model predictions and observed concentrations. The latter generic class of statistics requires pairwise matching. It was concluded that the most appropriate exceedance measures would be macrostatistics and that the most appropriate measures of coincidence and fidelity would be microstatistics.

After evaluating a number of candidates, six statistics were chosen to describe the model's accuracy. To measure a model's ability to estimate peak concentrations, the ratio of the largest 5 percent of the predicted concentrations, to the largest 5 percent of the observed concentrations was selected. To measure how well the model predicts the frequency of exceedance of the air quality standard, the difference between the predicted and observed proportion of exceedances of the standard was selected. To measure the coincidence of predicted and observed concentrations, the temporal, spatial, and Pearson's correlation coefficients were selected. Finally, to measure the degree of agreement in concentration levels, the root-mean-squared (RMS) error of the observed and predicted concentations was selected.

Confidence Intervals and Sample Size Determination

Confidence bounds on the values of the six evaluation statistics are computed because there is a degree of uncertainty associated with those values. Typically the data base is a sparse sample of all of the potentially available data and consists of a few months or years of data with temporal gaps, gathered at a handful of monitoring stations. Confidence intervals quantify the degree of uncertainty associated with the finiteness of the data base and are used both to gain an appreciation of the magnitude of the variability of the evaluation statistics and to estimate the sample size required for a given precision in estimation. Both parametric and nonparametric confidence intervals are derived for each statistic. The parametric confidence intervals depend for their validity on assumptions concerning the joint distribution of observed and predicted concentrations but do not depend on the data base. The nonparametric confidence intervals do not depend on assumptions concerning the joint distribution but are computed from the data. Based on experience, the parametric confidence intervals will tend to be too narrow, but are appropriate for the initial estimate of the sample size necessary to obtain a given level of precision. The nonparametric confidence intervals are used after the original sample has been drawn to refine the sample size determination and to assess more accurately the variability of the evaluation statistics.

Adjustment for Observational Error

Traditionally, the comparison of predictions and observations has been performed under the assumption that the differences between the computed and measured values are the fault of the model. Because the measured concentrations are subject to error, such an assumption places an unfairly heavy burden on the model. The authors of this report have attempted to remedy this situation by explicitly considering the effect of errors in the observed concentrations on the statistics selected to evaluate the performance of the model. Correction factors to be applied to the statistics to adjust for observational error are derived under two different assumptions concerning the nature of that error.

The Figure of Merit (FOM)

The figure of merit summarizes in a single statistic the correspondence between the observed and predicted concentrations as measured by the six component evaluation statistics. In deriving the FOM, a sequence of steps is undertaken, including the construction of a ten-point scale for each of the six component statistics, and the choice of a strategy (e.g., minimum liability, minimum average liability) for combining the scale-transformed and averaged statistic values.

THE DIAGNOSTIC ANALYSIS

Diagnostic statistics assist the researcher and highway planner to identify conditions associated with inaccuracies in

the model's predictions. Six categories of diagnostic techniques were found useful: time series, spatial, scatterplot, frequency distributions, multiple regression, and specialized displays. Within each category, the following were identified: diagnostic statistics that were appropriate to the examination of the agreement between all observations and predictions, the magnitude of the peak daily observations and predictions, and the time difference between the peak daily observations and predictions. An iterative procedure for using these diagnostic statistics is discussed.

DEVELOPMENT OF METHODOLOGY

Review of Evaluation Techniques

An extensive survey of the literature was performed to examine patterns and trends in model evaluation methodology. Table 1 gives the publications reviewed, along with the model evaluation methods used or proposed by their authors. (Full citations for literature references in Table 1 (and other publications cited in this section) are included in Chapter Seven.) The publications fall in two categories: methodological and applied. The methodological papers develop and analyze evaluation techniques; the focus of the applied literature is on the use of the methods. The methodological publications are those by Bornstein and Anderson (1979), Brier (1973), Goodin et al. (1976), Hayes (1979), Nappo (1974), Snee (1977), and Zannetti and Switzer (1979). The reports by Darling et al. (1977) and by Ruff and Javitz (1979) fall in both categories. The remaining publications are in the applied category. (See Bibliography for complete citations.)

It is evident from Table 1 that the correlation/linear regression method is the most popular approach used to characterize model performance. As the table shows, many of the early model verification efforts used correlation analysis exclusively. However, the current trend is to use multiple methods to assess model performance, as evidenced by the wide variety of techniques given in Table 1.

The methods in Table 1 can be classified as:

- Diagnostic methods—help to identify sources and conditions associated with modeling errors.
- Accuracy-evaluation methods—summarize overall model performance by assigning a numerical value to the degree of agreement between computation and observation.

In general, the diagnostic methods are qualitative or graphical, and enter the model-building process during the developmental stage when modifications of the model are contemplated. These methods also serve as informal assessment techniques that help the modeler to visualize the performance of the model. The accuracy-evaluation methods would be used when the emphasis is on a quantitative description of the predictive performance of a model rather than on model development.

Because some techniques can serve both purposes, the line between the two categories is sometimes blurred. Almost any statistic that quantifies model performance can also be used to assist in the process of subjective judgments that enters into model development. Moreover, some accuracy-evaluation and diagnostic methods are closely related. For example, the slope of the regression line is related to the scatterplot, but the former is a quantitative performance index and the latter a qualitative tool. Because of the dual purpose served by many accuracy-evaluation methods, ambiguity is avoided by restricting the use of the term "diagnostic" to nonquantitative techniques.

The various methods are assigned to the diagnostic or accuracy-evaluation category in Table 2. In this classification, the diagnostic methods include all the graphical techniques, which provide a visual indication of the model's behavior. By contrast, all the methods assigned to the accuracy-evaluation class describe the model's performance by a single number.

The accuracy-evaluation methods are generally complementary rather than mutually exclusive. Thus, some of the methods better quantify a particular aspect of model performance than do others. For example, the correlation coefficient measures the correspondence between prediction and observation in the aggregate, but gives no indication about tendencies to overpredict or underpredict. Supplementing the correlation coefficient with the slope and intercept of the regression line helps to remedy this inadequacy. To measure multiple attributes of a model's performance, more than one accuracy-evaluation statistic will be required. The literature review suggests that the use of multiple statistics is the current trend.

Most of the accuracy-evaluation methods in Table 2 can be classified into one of three classes. The first class consists of techniques that quantify, in the aggregate, the magnitude of the error between predictions and observations. For convenience, this class of statistics is called "exceedance measures." Examples include the difference between the 80th percentiles of the predicted and observed concentrations and the ratio of the average-predicted to the average-observed concentration. An essential feature of the exceedance measures is that they do not utilize the natural pairing of the predictions and observations. The second class consists of techniques that quantify the errors between the predictions and observations using the natural pairing. This class of statistics, which includes the root-mean-square error, is called "fidelity measures." The third class, which is termed "coincidence measures," contains methods that quantify the correlation between prediction and observation. The various accuracy-evaluation methods have been assigned to these three classes in Table 2. Two of the techniques do not fit either class and are listed in a "miscellaneous" class.

The literature review revealed only two reports that consider the effects of observational error in the model-evaluation process. Brier (1973) discusses the effect of normally distributed errors in the observations, and Maldonado and Bullin (1977) recognize the presence of observational errors by assigning a tolerance band to the observations.

With two exceptions, none of the reports studied addresses the issue of assigning different weights to various types of error, such as overprediction and underprediction. The biased log-difference-squared loss function defined by Darling et al. (1977) allows errors to be weighted (i.e., biased, in various ways), and so does the generalized accuracy score proposed by Ruff and Javitz (1979). Although it is difficult to decide the weights to be assigned to different error types, there is no *a priori* reason to believe that weighting errors

Table 1. Summary of literature review.

Reference*	Model	Source Type	Evaluation Method†	Other Methods
Bornstein and Anderson (1979)	CO highway models, Gaussian, puff, SO_2, statistical	Single site, elevated, point, surface, area	C	Superimposed time series plots (smoothed values; log or linear scales) Contingency tables Transaction plots Isopleth diagrams Normalized and unnormalized deviations about 45 degree line Frequency distribution plot Scattergram Cumulative frequency plots Percent of calculated values within given "tolerance" of observed Bivariate frequency distribution Ratio of averages Root-mean-square error Percent correct predictions versus forecast time Average fractional percent error
Brier (1973)	AQDM and GEOMET	Multiple	SA, C	Considers errors in computed and observed data Scattergram Mean square deviation from regression
Burr and Clymer (1976)	OCAM Model	Point, area	C, LR	
Calder (1970)				
Chock (1977, 1978a,b)	Advection-diffusion, Gaussian	Line	C	Correlations and variances Compares predicted to observed concentration ratios as a function of wind speed and wind Predicted-to-observed concentration ratios averaged for all runs
Christiansen (1976); Porter and Christiansen (1976)	TCM, TEM	Point, area	C, LR	
Clarke (1964)	Simple, non-computerized	Point, area	C, LR	Error frequency distributions
Darling et al. (1977)	Gaussian and numerical CO models for highway applications	Line, area	C	Compared output of two models against each other using: Correlation coefficient Average absolute difference 80th percentile difference Log-difference-squared loss function Biased log-difference-squared loss function Rank correlation Natural histogram
Dowell et al. (1979)	Gaussian model using Briggs plume rise		C, LR	Variance explained (coefficient of determination), intercept
Duewer et al. (1978)	Eulerian model			Ratios (predicted/measured) Comparisons of time histories Scattergrams of observed versus calculated mean concentrations and observed versus calculated maximum concentrations Root-mean-square error Spatial correlation coefficient Defines four statistical measures based on correlation coefficient (quantiles of the distribution of correlation coefficients) Ratio of mean computed concentration to mean observed concentration

Table 1 (Continued

Reference*	Model	Source Type	Evaluation Method†	Other Methods
Egan and Lavery (1973)	Numerical advection-diffusion model	Line, area		Ratio of observed to predicted concentrations
Elderkin (1974)	Diffusion-deposition model (Gaussian)	Elevated (particulate tracer over Hanford diffusion grid)	C	Observed versus predicted concentration as a function of stability Geometric means/standard deviation of observed/predicted ratio Autocorrelation function
Eskridge and Demerjian (1977)	Numerical advection-diffusion	Line	C, LR	Scattergram of observed/predicted ratio Linear least-squares fit Frequency distribution
Fortak (1969)	Statistical model for Bremen	Point, area		Cumulative frequency distributions
Fuggle (1977)	AQDM (Air Quality Display Model), CDM (Climatological Dispersion Model)			Contingency tables for concentration categories
Gifford (1973)	ATDL	Area	C, LR	
Goodin et al. (1976)	Various		SA	Residual error as function of weighting schemes
Groff (1973)	Econometric Time Series Models			Frequency of model ranking based on multiple tests with a variety of data sets
Habegger et al. (1974)	Gaussian steady-state plume model	Line	C, LR	Multiple or nonlinear regression for concentration dependence on model inputs Compares 17 models for at grade and depressed roadways
Hayes (1979)				Discussion of model evaluation approaches; recommended performance measures: Difference in peak station prediction and peak station measurement. Difference in time of occurrence of peak. Average and standard deviation of mean residual about the perfect correlation line normalized by average of predicted and observed concentration Average and standard deviation of absolute mean deviation about perfect correlation line normalized by average of predicted and observed concentrations Temporal correlation coefficients Spatial correlation coefficients
Koch et al. (1976); Koch and Thayer (1971)	SCIM	Point, area	C, LR	Concentration frequency distributions
Koogler et al. (1971)	Computerized model	Point, area	C	Chi-square test on two-way contingency table
Liu and Seinfeld (1974)	Eulerian and Lagrangian models			Trajectory (Lagrangian) model; ratios of predicted concentration to exact solution, plotted as function of time for: Horizontal diffusion Vertical winds Wind shear Grid (Eulerian) model; analytical solution versus computed percentage error

Table 1 (Continued)

Reference*	Model	Source Type	Evaluation Method†	Other Methods
Maldonado and Bullin (1977)	TRAPS, Gaussian dispersion, CALINE-2	Moving point, line	C, LR	Average error Average mean square error Percent within ± 1 ppm Percent within ± 2 ppm
	AIRPOL-4	All road points integrated		
	HIWAY, Gaussian plume	Line, point		
McCollister and Wilson (1975)	Linear stochastic models			Ratio of mean absolute value of forecasting error to mean observed concentrations Computer-animated films comparing measured and forecast temporal and spatial pollution patterns
McNider (1977)	AQDM	Point, area	C	Measured versus model average concentration categorized by wind direction, speed, and stability classes Scattergram comparing arithmetic annual averages
Miller et al. (1976)	Aquatic ecosystem		SA	Variation of output from changes of input parameters Maximum error estimates Relative sensitivity coefficients
Miller and Holzworth (1967)	Simple, non-computerized model	Point, area	C, LR	Chi-square test on two-way contingency table
Mukherji et al. (1976)	PTMTP	Point	SA, C, LR	
Mullen et al. (1977)	AVMSTM (Steady-state, simple Gaussian plume)		C	
Nappo (1974)	General models	Area, point	C	Spatial and temporal correlation
Newman and Spiegler (1974)	ERTAQ	Area	SA, C, LR	
Noll et al. (1978)	HIWAY, CALINE-2	Line	SA, C, LR	Ratio of average of predicted concentration to average measured concentration
Okamoto and Shiozawa (1978)	Gaussian	Area	C	Root-mean-square error Plot of percentile concentration curves Plot of average concentrations classified by wind direction
Prahm and Christensen (1977)	Gaussian	Multiple	C, LR	Spatial correlation
Prasad (1976)	AQDM	Area	SA, C, LR	
Rao et al. (1979)	Gaussian, numerical	Line	C, LR	Comparison of frequency distributions, scatterplots, ratio of predicted to observed
Rao and Visalli (1981)	Gaussian	Line	C, LR	Extreme value statistics, analysis of residuals, scatterplots Average percent error, mean fractional error Frequency distributions, contingency tables
Reynolds et al. (1974)	General model for 3-D time dependent photochemical air pollutants		SA, C	Analysis of residuals

Table 1 (Continued)

Reference*	Model	Source Type	Evaluation Method†	Other Methods
Ruff and Javitz (1979)	Gaussian	Point, area	C, LR	Scatterplot Linear regression with confidence and prediction bands Correlation coefficient Chi-square test Interstation error correlation Multiple regression of residuals Residual time series Generalized accuracy score
Ruff and Simmon (1977); Guldberg et al. (1976)	CDM	Point, area	SA, C, LR	
Shieh and Shir (1976)	K-theory grid type model	Point, area	SA, C, LR	
Smith and Ruch (1979)	Valley MODCDM		LR	Frequency distribution Scattergram of measured/computed concentrations
Snee (1977)	Statistical models		C, LR	Residual standard deviation
Tesche et al. (1979)	Urban airshed model	Area	C	Percent difference between predicted and observed maximum Concentration isopleths Distance required to bracket an observation Scattergram Fractional deviation computed along perfect correlation line Comparison of predicted and observed frequency distributions Relative error
Thayer (1974)	Gaussian	Airport	C, LR	Percentile and mean of observed versus predicted Regression analysis on medians
Tikvart and Mears (1976)	Single source model	Point		Comparison of highest and second-highest concentrations Concentration frequency distributions
Turner (1964)	Computerized model	Area		Error frequency distributions Average absolute and RMS error
Turner et al. (1972)	AQDM, CDM, Gifford 1972 (simple dispersion), Modified Hanna Model		C, LR	Root-mean-square error Mean absolute error Largest negative error Largest positive error Range of errors Error at location withh highest measured concentration
Wang (1974)	AVAP (airport pollution)	Airport	C, LR	Scatterplot of observed versus predicted concentrations Histogram of calculated and observed
Zannetti and Switzer (1978)	Numerical models		C	Residual time series Autocorrelation Plot performance of model versus length of learning period Plot of prediction lag versus root-mean-square error Plot of prediction lag versus correlation coefficient Model output plotted versus input parameters Plot of percentage of "correct" prediction versus forecasting lag

*Full citations for these references are given in the bibliography.

†SA = Sensitivity Analysis, C = Correlation, LR = Linear Regression.

Table 2. Classification of model evaluation methods.

Diagnostic Methods	Accuracy Evaluation Methods
Superimposed time series plots	Fidelity Measures
Transection plots	Percent of calculated values within specified tolerance of observation
Isopleth diagrams	Mean-square and root-mean-square error
Normalized and unnormalized deviations about 45 degree line	Average ratio of predicated to observed concentration
Scatter diagram of observations as a function of computations	Geometric mean and standard deviation of observation/prediction ratio
Cumulative frequency plots	Difference in peak-station prediction and measurement
Plot of bivariate frequency distribution	Average error (may be normalized)
Plot of ratio of predicted to observed concentration as a function of wind speed and direction	Mean of absolute error (may be divided by mean of observed concentrations)
Autocorrelation function	Largest negative and positive error
Contingency tables	Range of errors
Residual error with different weighting schemes	Generalized accuracy score
Sensitivity analysis	Average fractional percent error
Multiple regression of concentration dependence on model inputs	Biased or unbiased log-difference squared loss function
Plot of average predicted and observed concentrations classified by wind direction	Exceedance Measures
Distribution of residuals	Ratio of average predicted to average observed concentration
Residual time series	80th-percentile difference
Plot of prediction lag as a function of root-mean-square error and correlation coefficient	Correspondence Measures
Model output plotted as a function of input parameters	Correlation coefficient
Interstation error correlation	Slope and intercept of regression line
Multiple regression of residuals	Spatial correlation coefficient
Natural histogram	Quantiles of the spatial distribution of the correlation coefficient
Percent correct predictions as a function of forecast time	Temporal correlation coefficient
	Difference in time of occurrence of predicted and observed peak concentration
	Rank correlation
	Miscellaneous
	Frequency of model ranking
	Mean-square deviation from regression line

equally (the usual procedure) is especially advantageous under all circumstances.

Finally, it is apparent from the literature review that current practice is to use multiple measures to evaluate model performance. However, none of the sources reviewed considers combining the various performance measures into a single figure of merit.

The Accuracy Analysis

Types of Statistical Measures in an Accuracy Analysis

In evaluating the accuracy of a model, answers are sought to the following questions: (1) How well does the model predict maximum pollution levels? (2) How well does the model predict the number of exceedances of the relevant air quality standard? (3) How well do the fluctuations in predictions follow fluctuations of the observed pollutant levels in time and space? (4) How closely do the computed concentrations approximate the observations?

The first two questions are motivated by the fact that the air quality standards for carbon monoxide (CO) are framed in terms of maximum hourly and eight hourly averages not to be exceeded more than a specified number of times in a year. Hence it is important for the model to estimate accurately maximum concentrations and the number of times the standard is exceeded. Accordingly, it is desired to characterize numerically the model's performance in these categories. The statistics used for these purposes have been termed *exceedance measures*.

The third question is tied to the need to know whether the computed values coincide in space and time with the observations. It is of interest to ascertain whether known spots with high CO levels are correctly identified, and whether time fluctuations are reproduced properly. The statistics used to quantify this aspect of the model's behavior have been called *coincidence measures*.

The last question is concerned with the fidelity with which the predictions reproduce the numerical value of the observations. By analyzing the residuals (observed minus predicted), it is possible to examine tendencies to overpredict or underpredict. The statistics used to quantify the difference between predicted and observed concentrations have been termed *fidelity measures*.

Criteria for Selecting the Component Statistics of the Figure of Merit

The components of the FOM are statistics selected or designed to measure the four aspects of model performance previously defined. In the evaluation of various statistics for this purpose, the analysis was restricted to those that can be construed as estimates of a physically meaningful quantity. Examples of statistics of this nature include the root-mean-

square error and the correlation coefficient. These statistics have the property that as the sample size becomes very large, the estimate gravitates toward the "true" value of the parameter being estimated, and the confidence interval around the estimate shortens.

The use of statistics that could *not* be easily construed as estimates of physically meaningful quantities was rejected. Statistics of this nature are basically used to test hypotheses, diverge to infinity as the sample size becomes large, and must be referenced to statistical tables in order to determine their significance. Examples of these "test statistics" include rank-sum test statistics and chi-square test statistics. It is very difficult to derive, or even sometimes define, the concept of a "confidence interval" for this type of statistic.

In evaluating statistical estimates as candidates for the components of the FOM, it was found useful to distinguish between two generic types of statistics, which have been labeled "macro" and "micro" statistics, respectively. Macrostatistics are calculated without the pairwise matching of model predictions and observed concentrations. The ratio of the highest predicted concentration to the highest observed concentration (where these concentrations may occur at different locations and times) is an example of a macrostatistic. Microstatistics require the pairwise matching of model predictions and observed concentrations. The correlation coefficient is an example of a microstatistic. In a statistical sense, the macrostatistics may be computed from the marginal distributions of the model predictions and observed concentrations, whereas the microstatistics require the joint distribution of the model predictions and observed concentrations. In a practical sense, the microstatistics are oriented toward establishing the model's ability to reproduce short-term fluctuations of the pollutant; thus, they may be said to be process-oriented. By contrast, the macrostatistics are oriented toward the determination of whether the model can be used to predict the violation of, or conformity to, the air quality standards on a long-term basis. In this sense, macrostatistics may be said to be exceedance-oriented.

After considering the orientation of these two generic classes of statistics, it was concluded that the most appropriate exceedance measures would be macrostatistics and that the most appropriate measures of coincidence and fidelity would be microstatistics. The exceedance measures are concerned with the peak concentrations and the concentrations above the air quality standards. From the form of the air quality standards, it appeared more important to gauge the magnitude of the peak concentration and the frequency of the violations of the standards than to measure their particular time or location of occurrence. (Insight is gained into the agreement of the timing and spatial location of these higher concentrations from the coincidence measures.) If a model can accurately estimate these quantities, it may be confidently used to predict whether the standards will or will not be met in a region, regardless of where and when the concentrations actually occur. Certain types of models may perform perfectly well in this regard, and yet be very poor predictors of the exact time or location of these higher concentrations. Thus, macrostatistics are oriented towards the measurement of the exceedances without regard to the exact time or location of the higher concentrations.

On the other hand, in evaluating the air quality impact of a new road it is also necessary to perform a detailed examination of the area, timing, and probable cause of the pollutant levels, and this requires that the model be able to reproduce the air pollution process. As noted earlier, one aspect of such an examination concerns the coincidence of the predicted and observed levels over time and space. That is, the model predictions should be large when and where the observed concentrations are large, and the predictions should be small when and where the observations are small. It is also desired to describe the fidelity with which a model reproduces observations. A model that performs well in terms of maximizing coincidence and minimizing the error magnitude (i.e., maximizing fidelity) will prove useful in examining the air quality in detail, because it is capable of reproducing the air pollution process. The natural statistical estimates of coincidence and fidelity are microstatistics.

Definition of the Component Statistics

After considering a number of statistics as potential candidates to measure the four aspects of model performance, six statistics were chosen for this purpose; hereafter they will be denoted by the symbols S_1 through S_6. These six statistics are either well known or are easily interpreted:

- The ratio of the largest 5 percent of the observed concentrations to the largest 5 percent of the predicted concentrations (S_1) has been selected to measure a model's ability to estimate peak concentrations.
- The difference between the predicted and observed proportion of exceedances of the standard (S_2) has been selected to measure how well a model predicts the frequency of exceedance of the air quality standard.
- Pearson's correlation coefficient (S_3), temporal correlation coefficient (S_4), and spatial correlation coefficient (S_5) have been selected to measure the coincidence of predicted and observed concentrations.
- The root-mean-square error of the observed and predicted concentrations (S_6) has been selected to measure the degree of agreement in concentration levels.

The mathematical formulas for the six component statistics are given in Appendix B.

A few comments on these statistics are in order. The statistic chosen to measure agreement in peak concentrations actually measures the agreement in the average magnitude of the largest 5 percent of the concentrations. This was done in order to assist in stabilizing the variance of the statistic.

With respect to the statistic chosen to measure the agreement in the exceedance frequency, the difference in the frequencies, rather than a relative error measure, has been used because the air quality standard is defined in terms of frequencies and because relative errors are very sensitive to low frequencies.

For each of the six statistics, a choice was made between untransformed or log-transformed concentrations. Untransformed concentrations were used in the first, second, and sixth statistics. A principal reason for using log-transformed concentrations is to convert absolute differences to ratios. In the case of S_1, this is unnecessary because S_1 is itself a ratio. Furthermore, there is little or no advantage in compressing the scale logarithmically to compute S_1 because, barring gross errors, it is likely that the highest predicted and ob-

served levels will be of comparable magnitude. The second statistic is invariant to the transformation; hence, there is no need to convert the data to log concentrations. Untransformed concentrations were used for the sixth statistic because it is desired to estimate the absolute error instead of the relative error.

Log-transformed concentrations were used in the computation of the three correlation coefficient statistics. Without this transformation the correlation coefficients would be dominated by the largest predicted and observed concentrations. Also, the scatterplot of the predicted and observed concentrations often tends to resemble more closely the bivariate normal distribution when the axes are logarithmic. In addition, the residual variance around the regression line is usually close to being constant using logarithmic axes, which allows the interpretation of $1 - \rho^2$ as the proportional reduction in variance.

The statistic chosen to measure the agreement in absolute concentration levels may be expressed as a sum of two statistics—one measuring the bias (or systematic error) in prediction and the other measuring the magnitude of the "random" error corrected for bias. For evaluation purposes it is not necessary to distinguish between error caused by bias and random error, although it is essential for diagnostic purposes. Consequently, statistics for measuring bias and random error have been included in the set of diagnostic statistics. The root-mean-square error was chosen to be used in the linear scale rather than in a logarithmic scale because the errors associated with larger concentrations have a more serious implication for model accuracy than do errors associated with smaller concentrations.

Confidence Intervals and Sample Size Determination

Confidence intervals quantify the uncertainty in the estimated value of the evaluation statistics associated with the finiteness of the data base. The data base typically consists of a few months or years of data with temporal gaps, gathered at a handful of monitoring stations. Consequently, the data base may be considered to be a sparse sample of all of the potentially available data. The predictive model may have performed somewhat better or worse on this data sample than it would have performed on a larger and more complete data base. The values of the component statistics used in the evaluation, therefore, might overstate or understate the predictive ability of the model.

Two types of confidence intervals are considered: parametric and nonparametric. The parametric confidence interval is based on assumptions concerning the form of the joint distribution of observed and predicted concentrations. The nonparametric confidence intervals do not depend on these types of assumptions. Based on experience, the assumptions underlying the parametric confidence intervals will be violated and the intervals will be too narrow. However, the parametric confidence intervals may be accurate enough to estimate the sample size necessary for a given precision, and it is intended that they be used for that purpose. After the sample size has been estimated and the appropriate sample has been drawn, the more accurate nonparametric confidence interval can be calculated. On the basis of the nonparametric confidence interval, the investigator may decide to revise his estimate (obtained from the parametric confidence intervals) of the necessary sample size, and draw a larger sample. Unfortunately, the nonparametric confidence intervals cannot be used to obtain the initial estimate of the sample size because they cannot be calculated without data.

The mathematical derivation of the parametric and nonparametric confidence intervals for each of the six component statistics is presented in Appendix C.

The Adjustment for Observational Error

Traditionally, the comparison of prediction and observation has been performed under the assumption that the measurements are correct (if not exact) and that differences between computed and measured values are the fault of the model. This assumption permeates, for example, the classical bivariate correlation/linear regression analysis, which considers the explanatory variable to be known without error. Because the measured concentrations are subject to error, such an assumption tends to place an unfairly heavy burden on the model. An attempt has been made in this study to remedy this situation by explicitly considering the effect of data errors on the six statistics selected to evaluate the performance of the model.

Various approaches can be used to consider the effects of data errors. Two different methods for dealing with imperfect data were briefly mentioned in the context of the literature review (Brier, 1973; Maldonado and Bullin, 1979). The approach in this report (Project 20-18) is to modify the formulas for the six statistics of interest to include terms associated with data errors.

In general, three sources of error affect the comparison of observed and predicted concentrations:

1. Errors in the inputs of the model.
2. Modeling errors.
3. Measurement errors in observed concentrations.

The output of the model combines the effects of input and modeling errors. Sensitivity analysis can help estimate the magnitude of the errors associated with imperfect model input. Judicious application of other diagnostic methods can help to identify some of the sources of modeling errors, but the output of the model will nevertheless be affected by errors. Modeling errors are caused by the inability of the model to simulate the physical phenomena of interest.

The third source of error—error in the measurement of the observed concentrations—affects the extent of agreement that can be obtained between observed and predicted concentrations. This source of error can include both a random component and a constant, or slowly varying, bias. The approach adjusts only for the random error in the observations. Thus, it is assumed that the measurements have been adjusted for bias; by implication, the mean of the error distribution is zero. It is emphasized that if the bias is known (for example, by analyzing the calibration data for the instruments), the data should be corrected. If the bias is unknown, it will appear as modeling error. Of concern here is the precision, rather than the accuracy, of the data. (It should be noted that federal regulations require that both accuracy and precision data be reported on a regular basis (*1*).)

The question naturally arises as to where one obtains information about measurement errors. One source of data is the

manufacturer's specification. This should be considered as a lower bound on the measurement error: the measurement error cannot be smaller, and will probably be higher. Calibration records for each instrument are the best source of information about measurement accuracy and precision. The Texas data base (see App. G) used calibration data to obtain estimates of measurement error. Whether calibration data are available depends on the design of the experimental program. Of the five data bases used in this study, only the Texas data base explicitly defined the measurement errors for the various instruments (see Table G-58 in App. G).

For each of the six component statistics formulas have been derived that include the effect of random observational errors under two models. In the first model (called the multiplicative model), the observational error is assumed to be proportional to the true concentration (e.g., the error is ±10 percent). In the second model (called the additive model), the standard deviation of the observational error is expressed as a constant bound (e.g., ±50 ppm). The mathematical derivation of these corrections is presented in Appendix D.

Computation of the Figure of Merit

The FOM summarizes in a single index the correspondence between the observed and predicted concentrations with respect to the four attributes defined earlier. Consequently, the FOM is a function of the six statistics chosen to measure these attributes.

The first step in computing the FOM entails transforming the six statistics (S_1 through S_6) to a common scale. The lowest value on the scale (which has been designated to be zero) corresponds to very poor agreement between predictions and observations as measured by that statistic and the highest value on the scale (which has been designated to be 10) corresponds to very good agreement. The transformation to the common scale may be different for each component statistic; three distinct transformations have been selected for S_1, S_2, and S_6; and a common transformation has been selected for the three coincidence measures S_3, S_4, and S_5. The transformed scores are denoted as F_1 through F_6, each on a scale from 0 to 10.

By necessity, the specification of the transformation to a common scale includes a judgment factor. The importance of this judgment is not immediately apparent when examining any single transformation. For example, it is clear that if all else were equal, a model with a Pearson's correlation of 0.57 would be more desirable than a model with a Pearson's correlation of 0.30. However, any reasonable transformation (e.g., monotonic increasing from 0 to 1) would arrive at the same conclusion. Consequently, there would appear to be limited importance attached to the choice among reasonable transformations. However, when multiple measures are combined into a single FOM, the choice of the transformations assumes considerable importance. For example, it is much less clear whether, if all else is equal, a model with $S_3 = 0.57$ and $S_6 = 2.3$ ppm is preferable to a model with $S_3 = 0.30$ and $S_6 = 1.7$ ppm. Presumably, different researchers, employing different criteria, could reach dissimilar conclusions about the relative desirability of the models (or, they might reach the same conclusion for different reasons). The criteria used by these researchers might stem from their personal experiences, the thoroughness with which they have examined the problem of comparing models with multiple performance measures, the intended use of the models, or knowledge of expected losses associated with different types of misprediction. Ultimately these criteria affect the evaluation of the tradeoff between ppm of root-mean-squared error and percentage points of correlation, and the subsequent ranking of the models. Thus, the transformations (and the algorithm used to combine the transformed statistics) should be selected to reflect these criteria as closely as possible. In fact, it has been found useful to select an initial transformation, apply the transformation to real models and data, examine the subsequent extent of agreement of subjective evaluations with the FOM, and modify the transformations. This process of refinement also has the advantage of revealing inconsistencies in the internal criteria (and differences among the investigators) and assisted in resolving those inconsistencies. It is reasonable to assume that other investigators may disagree with the choice of transformations, because they are using different underlying criteria. In that event they should develop their own transformations, or present a case for modifying the transformations contained in this report.

For the S_1 statistic a transformation was used that gave equal values to models that underpredicted or overpredicted by a common factor. For example, if one model overpredicted by a factor of 2.0 and another model underpredicted by a factor of 2.0, their transformed S_1 statistics would be equal. This transformation is shown in Figure 1. Algebraically, the transformed statistic $F_1 = 10 * S_1$ for $S_1 \leq 1$, and $F_1 = 10/S_1$ if $S_1 > 1$.

For the S_2 statistic a transformation given algebraically by the following equation was used:

$$F_2 = 10\,(1 - |S_2|/\sqrt{P_1 + P_2})$$

where P_1 is the proportion of observed values greater than the threshold concentration used in defining S_2, and P_2 is the proportion of predicted values greater than the threshold. Note that $S_2 = P_2 - P_1$, so that F_2 is a measurement of the extent to which P_2 and P_1 differ, normalized by division of the square root of $P_1 + P_2$. Figure 2 shows the values of F_2 for different combinations of P_1 and P_2. On the diagonal line of perfect agreement (e.g., $P_1 = P_2$) the F_2 statistic is equal to 10, and if P_1 and P_2 disagree completely (e.g., $P_1 = 1$ and $P_2 = 0$ or $P_1 = 0$ and $P_2 = 1$) the F_2 statistic is equal to 0. Another perspective on F_2 can be gained from Figure 3, which shows F_2 as a function of S_2 and the minimum of P_1 and P_2.

In the computation of F_2, S_2 is divided by the square root of $P_1 + P_2$. This divisor was chosen because it yields an F_2 value that is sensitive to both the absolute difference between P_1 and P_2 and the proportional difference between P_1 and P_2. If the divisor had been chosen to be 1.0, F_2 would have been sensitive only to the absolute difference between P_1 and P_2. For example, a model with $P_1 = 0.01$ and $P_2 = 0.05$ would have received the same F_2 score as a model with $P_1 = 0.62$ and $P_2 = 0.58$. This situation of equal ranking would be undesirable, because the relative error for the first model is so much greater than the relative error for the second model. However, if the divisor had been $P_1 + P_2$, F_2 would have been sensitive to relative error. For example, a model with $P_1 = 0.001$ and $P_2 = 0.01$ would have received the same F_2 score as a model with $P_1 = 0.1$ and $P_2 = 1.0$. Again this situation of equal ranking would be undesirable, because the absolute

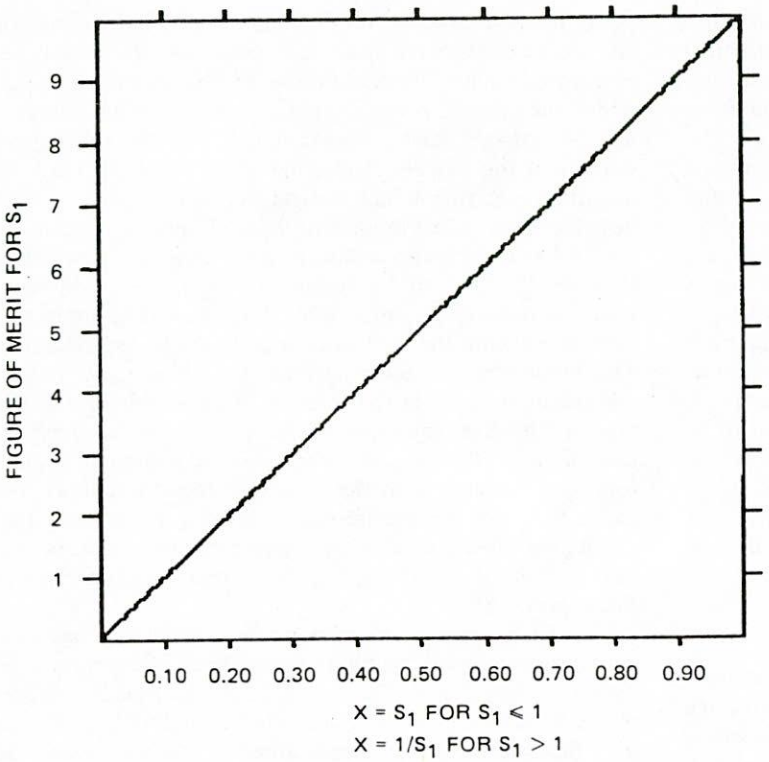


Figure 1. Figure of merit for S_1.

error of the first model is so much smaller than the absolute error of the second model. These considerations underlie the selection of the square root of $P_1 + P_2$, as the divisor, which is "half-way" between 1.0 (e.g., $P_1 + P_2$ raised to the zero-th power) and $P_1 + P_2$ (e.g., $P_1 + P_2$ raised to the first power).

For the correlation statistics S_3, S_4, and S_5 the transformation given in Figure 4 has been selected. For example F_3 is equal to 0 whenever S_3 is less than or equal to 0 and is equal to S_3 otherwise. This transformation reflects judgment that negative correlation is no better than no correlation.

For the root-mean-square error statistic S_6 the transformation used is shown in Figure 5. Algebraically the transformed value F_6 is given by $10/(1 + A * X * X)$ where A is equal to 3.704 and X is equal to S_6 divided by the mean observed concentration. The statistic S_6 was divided by the mean observed concentration because the ratio is scale invariant (e.g., the ratio is not affected by the units used, whether those units are ppm or ppb). The form of the transformation was selected so that small to moderate values of S_6 (e.g., a root-mean-squared error on the order of 10 percent of the observed mean concentration) would be associated with large F_6 values and the transformation would fall-off rapidly thereafter. The constant A = 3.704 was chosen so that F_6 would equal 7.5 when X was equal to 0.3.

Once the six statistics have been expressed in common units, the second task is to combine the statistics into a single FOM. Several strategies for computing the FOM were tested, and a choice between them was made on the basis of the data analysis. One strategy was to formulate the FOM so as to choose the model with the minimum liability. This strategy is based on the presumption that the worst possible single event will occur. For example, if a model performs poorly in predicting exceedances of the air quality standard, the minimum-liability criterion assumes that this characteristic will be the key ingredient in the highway planning problem. In this approach, the FOM was set equal to the worst transformed statistic score.

A second strategy was to formulate the FOM so as to choose the model with the minimum (weighted) average liability. This strategy is based on the presumption that errors accumulate in highway planning problems. In this event the FOM was set equal to the weighted average of the transformed statistic score. In computing this average a weight of one-half was assigned to each exceedance measure, a weight of one-third to each coincidence measure, and a weight of one to the fidelity measure. Thus, the overall FOM is given algebraically by $[(F_1 + F_2)/2 + (F_3 + F_4 + F_5)/3 + F_6]/3$.

The minimum (weighted) average liability definition for the FOM was selected because it was believed that the minimum liability approach would be less likely to produce similar model rankings when applied to different data sets. Moreover it may be overly pessimistic to assume that the worst possible situation would always occur. Chapter Five contains examples of the effect that the minimum liability and minimum average liability criteria have on model rankings.

The Diagnostic Analysis

Diagnostic statistics assist the model user to identify conditions associated with inaccuracies in the model's predictions. A number of diagnostic statistics (including graphical displays) are defined in the following that should be useful in evaluating the causes of model inaccuracies.

Taxonomy of Diagnostic Statistics

The diagnostic statistics found to be particularly useful

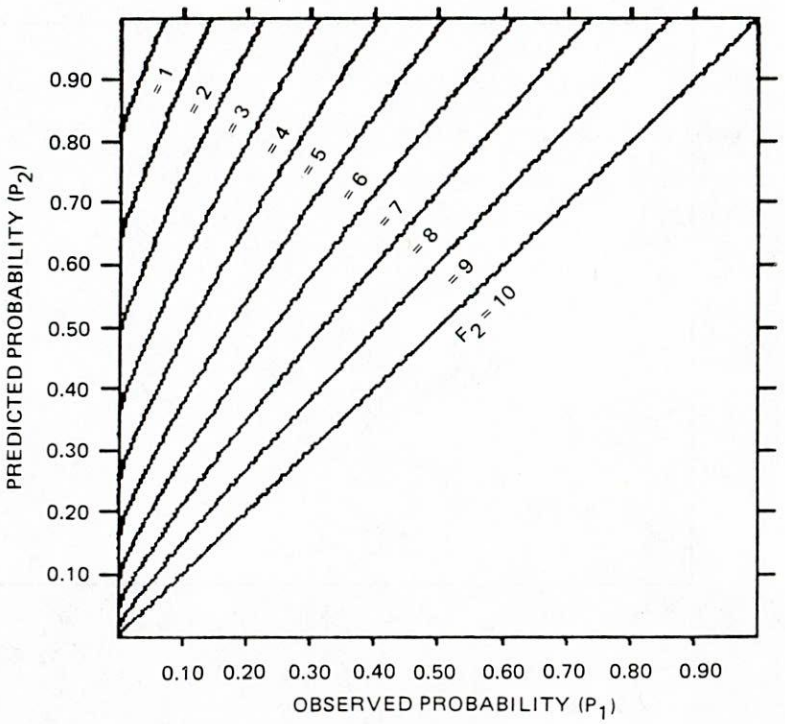


Figure 2. Relationship between F_2, *the figure of merit for* S_2, *and observed and predicted probabilities.*

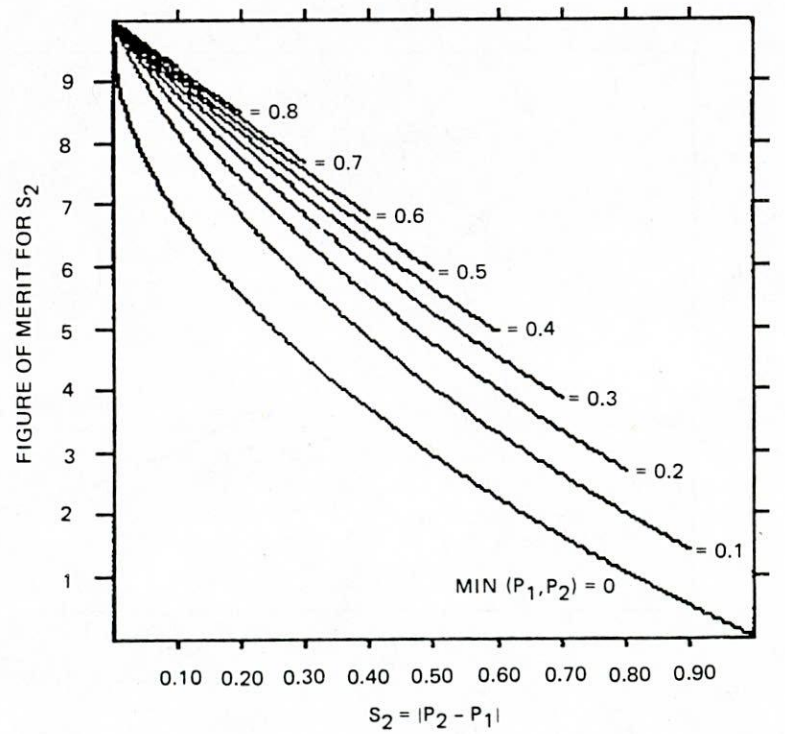


Figure 3. F_2, *the figure of merit for* S_2, *as a function of* S_2 (P_1 *and* P_2 *denote observed and predicted probability, respectively).*

may be classified into six basic categories: (1) time series, (2) spatial, (3) scatterplot, (4) frequency distributions, (5) multiple regression, and (6) specialized displays. Within each category are diagnostic statistics that are appropriate to the examination of the agreement between all observations and predictions, the magnitude of the peak daily observations and predictions, and/or the time difference between the peak daily observations and predictions. This taxonomy is illustrated in Table 3. Descriptions of these statistics are presented later.

Method for Using Diagnostic Statistics

Unlike evaluative statistics such as the FOM, diagnostic

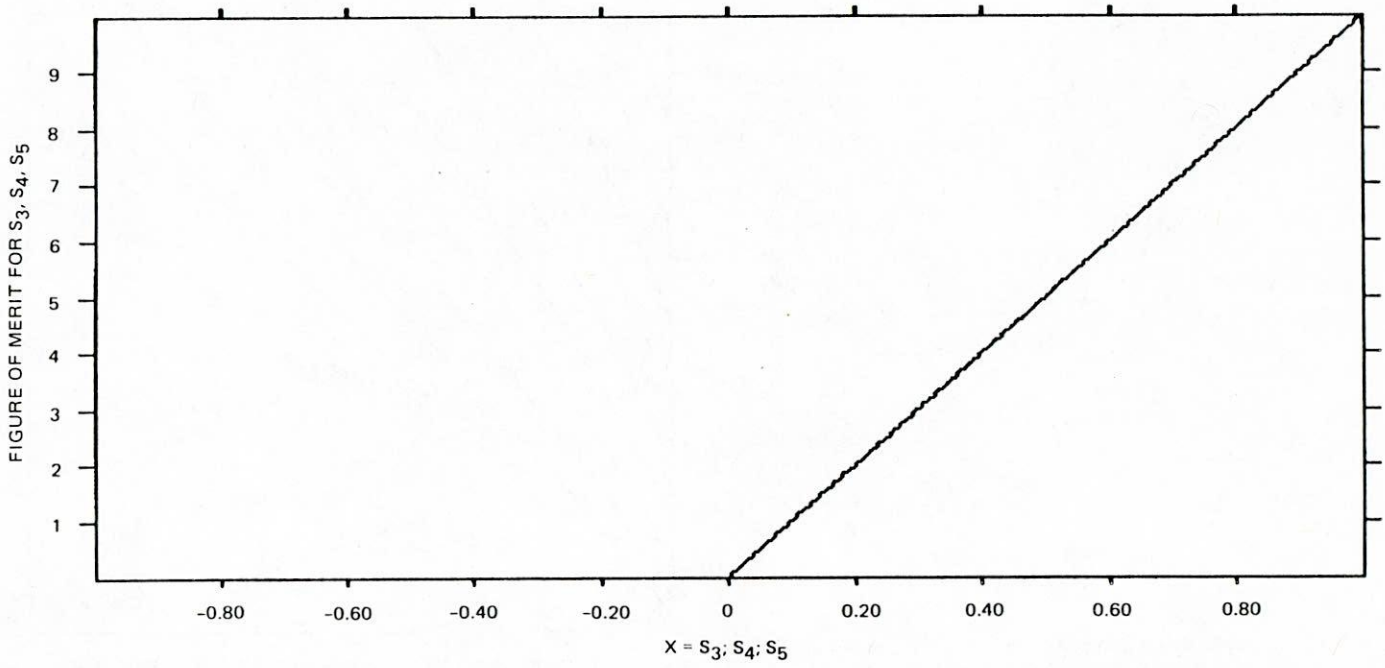


Figure 4. F_3, F_4, *and* F_5, *the figures of merit for* S_3, S_4, *and* S_5.

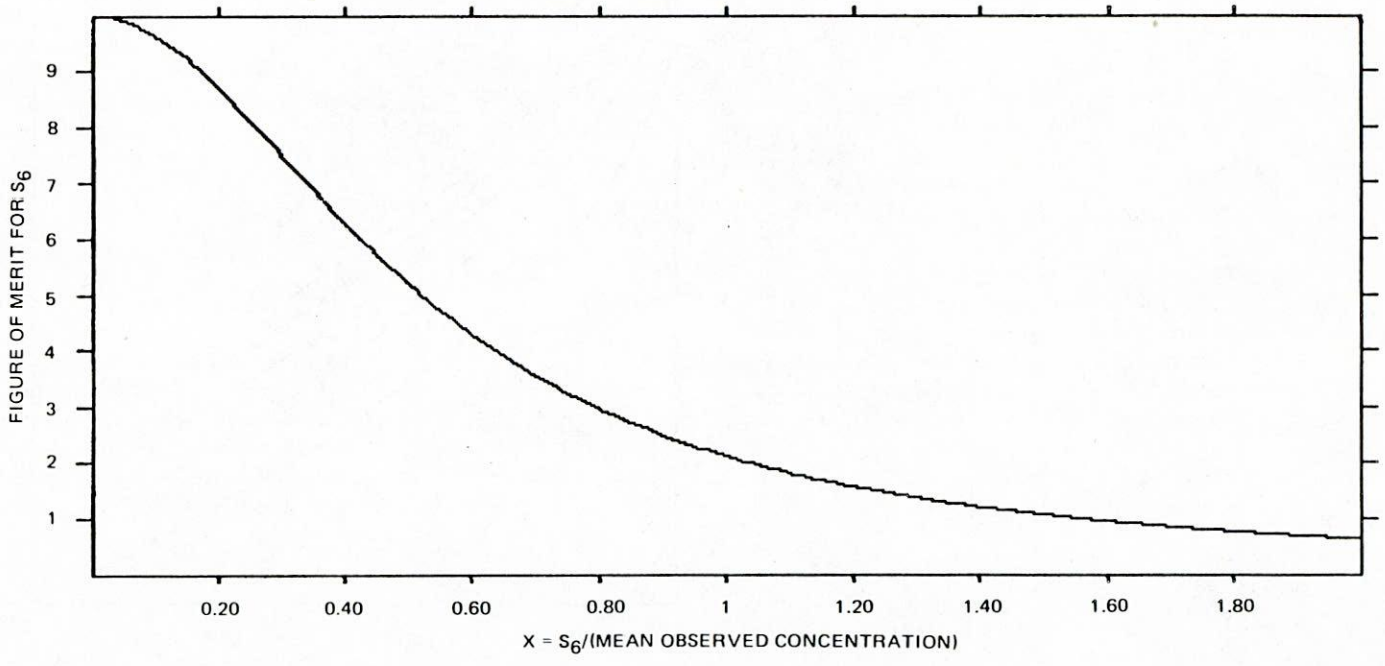


Figure 5. F_6, *the figure of merit for* S_6.

statistics are employed in an iterative and interactive fashion. The first step entails computing the diagnostic statistics on the entire data base, or—in the case of multiple data bases—on the one for which the model yields the lowest FOM. Examination of the diagnostic statistics and the model leads the investigator to suspect, for example, that certain atmospheric or emissions conditions result in poor agreement between observations and predictions. The investigator can then limit the data base to the days or sites in which those conditions occur and recompute the diagnostic statistics. The recomputed diagnostic statistics assist the investigator in determining why the model is breaking down under those conditions, and in identifying other factors that interact with the suspected condition. For example, the model may appear

Table 3. Classification of diagnostic statistics.

Category	Magnitude of Hourly Concentrations	Magnitude of the Peak Concentration	Time of the Peak Concentration
Time series	Superimposed time series of predicted and observed concentrations	Superimposed time series of daily peak predicted and observed concentrations	
	Time series of residuals on a linear scale	Time series of ratio of daily observed and predicted peaks	Time series of the daily difference between the times of the observed and predicted peaks
Spatial maps	Superimposed isopleth plots from observed and predicted concentrations for selected hours	Spatial map of the average ratio of the predicted to observed daily peaks	Spatial map of the average time difference between predicted and observed daily peaks
Scatterplot	Scatterplot of the observed versus the predicted concentrations (both linear and log scales)	Scatterplot of observed and predicted peak concentrations (linear scale)	
Marginal frequency distribution	Quantile-quantile plot for observed and predicted concentrations		
	Histograms of observed and predicted concentrations	Histogram of the peak observed and predicted daily concentrations	
Multiple regressions	Multiple regression of residual error on meteorological, emissions, and other factors		
	Specialized plots of residual error versus two explanatory variables		

to perform poorly when observed concentrations are large. Upon restricting the data base to days with large observed concentrations, the investigator may find that the model performs poorly only for a certain range of wind directions. On the basis of such insights, the investigator may decide to reselect the limiting conditions in order to understand better the model's difficulties. Continuing the example, the investigator may decide that it would be worthwhile to limit the data base on wind direction and emission magnitudes, rather than observed concentration magnitude, and to recompute the diagnostic statistics. This process continues until the investigator understands the conditions under which the model performs most poorly. The discussion that follows examines briefly the particular diagnostic statistics that are recommended, encompassing the six categories defined earlier.

Diagnostic Statistics Related to Time Series

Diagnostic statistics related to time series examine the temporal agreement between model predictions and observed concentrations. Two diagnostic statistics are relevant to the examination of the hourly agreement:

1. The superimposed time series of observed and predicted concentrations.
2. The residual time series (observations less predictions on a linear or logarithmic scale).

A similar diagnostic statistic may be applied to the examination of the agreement of the magnitude of the peak concentrations, namely the time series of the daily ratio of the peak predicted concentration to the peak observed concentration. This diagnostic statistic is also useful for examining the time difference between the predicted and observed concentrations if the daily time difference is substituted for the ratio of predicted and observed daily peak.

Diagnostic Statistics Related to Spatial Agreement

The spatial agreement of model predictions and observations is an important aspect of the model's predictive performance. The spatial agreement for a given time may be examined using superimposed isopleth plots for selected hours of observed and predicted concentrations. The agreement of the magnitude of the peak concentrations may be examined by generating a spatial map of the average ratio of the daily predicted and observed concentrations for each site. The time difference between observed and predicted daily peaks may be examined by generating a spatial map of the average daily time difference for each site.

Diagnostic Statistics Related to the Scatterplot

The scatterplot of observed and predicted concentrations is the most widely used diagnostic statistic for examining hourly agreement. The scatterplots in this study are constructed to include the least-squares regression line.

Diagnostic Statistics Related to the Marginal Frequency Distribution

The examination of marginal frequency distributions allows the investigator to judge whether the model is simulat-

ing the environmental conditions over a length of time rather than at any particular time. The accuracy of the model input should be severely questioned when the predicted and observed marginal frequency distributions are similar and the errors between hourly or peak daily predicted and observed concentrations are large. The marginal frequency distributions may be measured using:

1. A quantile-quantile plot (of observed as a function of predicted hourly concentrations, or observed as a function of predicted peak daily concentrations).
2. Histograms of observed and predicted hourly concentrations and observed and predicted peak daily concentrations.

Diagnostic Statistics Related to Multiple Regression

Multiple regression is the most powerful of the diagnostic statistics. The independent variables in the regression are such meteorological and emissions variables as stability class, wind speed, wind direction, and emissions rate. Two dependent variables will be used. The first—the residual error between the predicted and observed concentration—is useful in determining the conditions under which the model either underpredicts or overpredicts the observed concentrations, and consequently results in a bias. The second—the absolute value of the residual error—is useful in determining the conditions under which the magnitude of the error in prediction increases, even if no bias is present.

For the purposes of the multiple regression, the residual error between the observed and predicted concentrations OC and PC will be defined as

$$\text{Residual} = \frac{OC - PC}{\sqrt{(OC + PC)/2}} \qquad (1)$$

This form of the residual may be considered a compromise between the algebraic error and the relative error. The algebraic error is OC − PC and the relative error may be written either as (OC − PC)/OC or, in a more symmetric form, as (OC − PC)/[(OC + PC)/2]. A shortcoming of the algebraic error is that in general it emphasizes the largest observations and predictions, and slights the smaller ones. The relative error, on the other hand, tends to emphasize the smallest observations and predictions, and slights the larger observations and predictions. The residual as defined in Eq. 1 achieves a reasonable compromise. The nature of this compromise can be visualized in Figures 6 through 8. These figures display the curves along which the algebraic, symmetric relative, and compromise residual errors are constant.

Specialized Plots

Specialized plots, related to the multiple regression, provide a graphical display of the agreement between observed and predicted concentrations as a function of any two independent variables, and can be devised to fit any desired purpose.

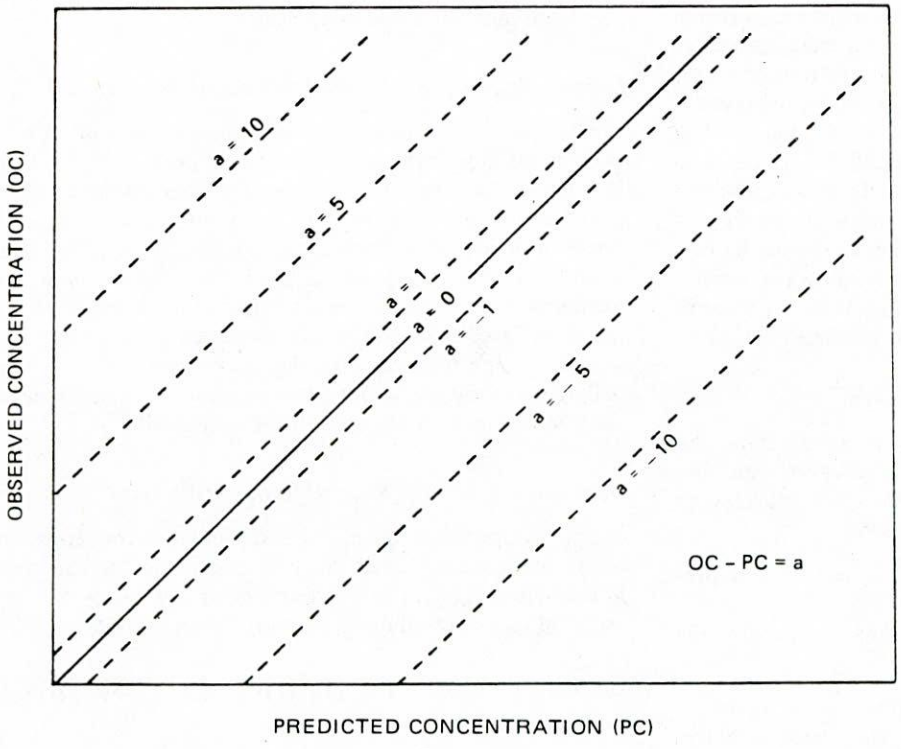


Figure 6. Curves along which algebraic error is constant.

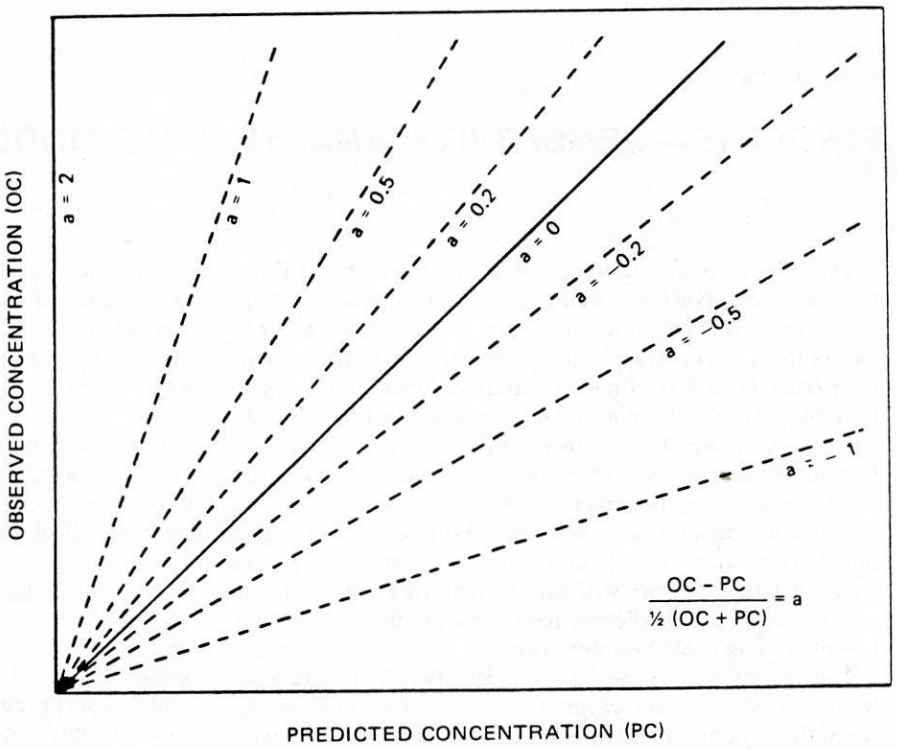


Figure 7. Curves along which symmetric relative error is constant.

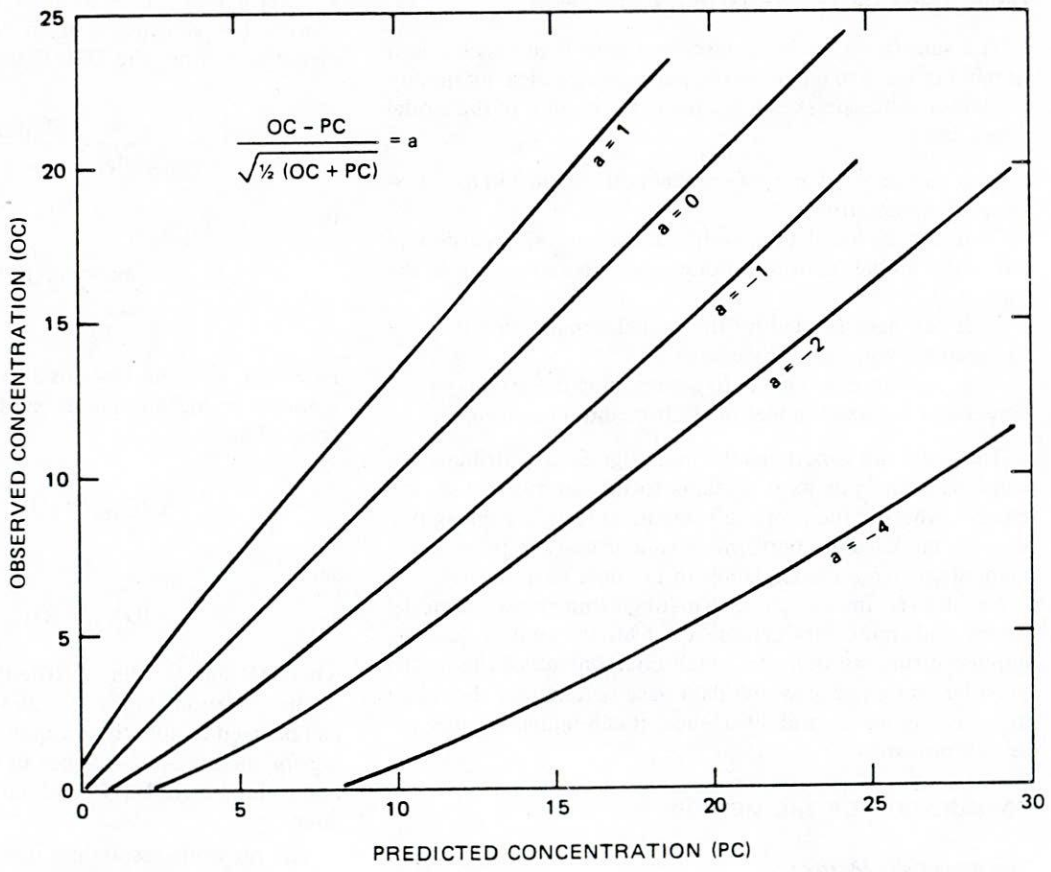


Figure 8. Curves along which residual error is constant.

CHAPTER THREE

FINDINGS—SENSITIVITY ANALYSIS METHODOLOGY

The solution of a multiparameter sensitivity matrix defines the sensitivity (variance and covariance) of model output predictions. This solution is a product of two other matrices. One of these is a matrix of sensitivity coefficients that defines the partial deviatives of model output variables with respect to individual model input variables; the second matrix specifies the user-supplied variances of model input parameters. For application in this report these variances are made to represent the expected error variance.

The same method is applicable to both steady-state and time-varying (i.e., dynamic) models. The matrices previously specified can vary with time so that, for each simulated event, the model prediction for each time interval can be bounded by expected output errors.

The accuracy of the sensitivity analysis is limited in analyzing the effect of model input errors because input errors can only be approximated. Because of this limitation, techniques that reduce the amount of computation are considered. These are briefly discussed in this chapter and illustrated in the example presented in Chapter Five.

OBJECTIVES OF THE SENSITIVITY ANALYSIS

The sensitivity analysis described here is applicable in a number of ways to diagnose the performance of an air quality model (or submodels) as it relates to the quality of the model input data:

1. It can serve to quantify model performance in the presence of input errors.
2. It can be used to quantify the required precision of particular model input to produce an acceptable error in the output.
3. It can help to identify the model components that are the greatest sources of inaccuracy.
4. It can serve as a guide to determining the origin of poor agreement between model prediction and measurement.

The last point is perhaps the most significant attribute of a sensitivity analysis as it pertains to this study; that is, the issue is whether the proposed statistical tests are telling one that the model is not performing well *or* that the poor agreement stems from inadequacies in the data base itself.

Besides serving as a guide to distinguishing between model errors and input data errors, a sensitivity analysis can be applied further to indicate which components of the model must be improved or which data base parameters should be measured more accurately. Hence, it can indicate future research priorities.

FORMULATION OF THE METHOD

The Sensitivity Matrix

Sensitivity is formally defined as the partial derivative of the output of a model to the input parameter(s) in question. In the case of complex models, it is more appropriate to consider incremental changes in output resulting from incremental changes in input, because determining the analytical expressions for partial derivatives becomes too cumbersome.

For relatively inert pollutants such as CO, the true (i.e., errorless) pollutant concentrations, x, can be represented as a function (mathematical model) of emissions input, Q, and meteorological input, G, as follows:

$$[x]_{t,s} = f\left([Q]_{t,s}, [G]_{t,s}\right) - [x'_{err}]_{t,s}$$

where x, x'_{err}, Q, and G are vectors in time, t, and space, s. The x'_{err} term represents an error in the model formulation, which is only a mathematical approximation of the physics and chemistry of the atmosphere. (In the notation used here,the prime will be used to designate errors caused by the model formulation; unprimed error terms will designate the total error.) For purposes of this report, it is assumed that the x'_{err} term represents an error not related to the Q and G arrays. The sensitivity of an element in the x array to an element in either the Q or G arrays is given by

$$\frac{\partial x_i}{\partial Q_j} = \frac{\partial f\{[Q]_{t,s}, [G]_{t,s}\}}{\partial Q_j} \tag{2a}$$

or

$$\frac{\partial x_i}{\partial G_j} = \frac{\partial f\{[Q]_{t,s}, [G]_{t,s}\}}{\partial G_j} \tag{2b}$$

However, Q's and G's are also only mathematical approximations or measurements subject to error. These are expressed as

$$[Q]_{t,s} = [Q]^*_{t,s} - [Q_{err}]_{t,s} \tag{3a}$$

and

$$[G]_{t,s} = [G]^*_{t,s} - [G_{err}]_{t,s} \tag{3b}$$

where Q* and G* denote true values.

The variations in the Q_{err} and G_{err} terms are numbers that can be used to quantify air quality model errors by substituting for incremental changes in Q_j and G_j in the discretized approximation of Eq. 2. Estimating model errors is discussed later.

The previous results are used to define a matrix of sensitivity coefficients, as follows. Let B denote the sensitivity coefficient matrix. Then

$$B = \begin{bmatrix} \frac{\partial \chi_1}{\partial Q_1} & \frac{\partial \chi_1}{\partial Q_2} & \cdots & \frac{\partial \chi_1}{\partial Q_N} & \frac{\partial \chi_1}{\partial G_1} & \cdots & \frac{\partial \chi_1}{\partial G_M} \\ \vdots & & & & & & \\ \frac{\partial \chi_L}{\partial Q_1} & \cdots & \cdots & \frac{\partial \chi_L}{\partial Q_N} & \frac{\partial \chi_L}{\partial G_1} & \cdots & \frac{\partial \chi_L}{\partial G_M} \end{bmatrix} \quad (4)$$

where

x_i = concentration of ith output variable, i = 1,2,..., L;
N = number of emission-related input items;
M = number of meteorological input items; and
L = number of output items.

Thus, B is of order L × (N + M). The ijth element of B expresses the influence of the jth input on the ith output variable. In the context of dispersion models the ith output would be the CO level at a specific receptor.

The B matrix plays a prominent role in determining the variance of the output as a function of the variance of the input parameters. This is discussed later in the section on multiparameter sensitivity. For numerical models, the elements of the B matrix will vary with time. However, treatment of the B matrix as time invariant is usually justified given other inaccuracies inherent in the method. A rigorous method for determination of the B matrix is given by Burns (*2*). Our approach makes use of several simplifying assumptions based on limited assessment of the accuracy of the model input data and on *a priori* knowledge of some of the sensitivity of the individual parameters.

The first step in this type of sensitivity analysis is to identify the factors that will be evaluated (e.g., traffic speed, traffic volume, wind speed, diffusivity, temperature, and source strength input). High and low values are identified for each factor.

The second step is to specify an experimental design (i.e., specifying the computer runs). The experimental design details the levels at which factors will be set during computer simulation runs. A factorial design can be used for this purpose; however, a full factorial design (which requires 2^P simulation runs, where P is the number of model parameters) is prohibitively expensive to perform with the more complicated models. It would, however, allow the estimation of all main factor effects and interactions. As a practical matter, an attempt is made to minimize the number of simulation runs in a manner that is consistent with the achievable accuracy. This will vary from one type of model to another: Chapter Five contains one example of an analysis that required much fewer than the 2^P simulation runs.

Treatment of Model Input Errors

The errors in the model input and validation data will be assumed to follow a Gaussian distribution as shown in Figure 9. The symbols in the figure indicate the error parameters of interest. The distribution of errors about the true value, T, is subject to bias, b, and variability. Precision, σ_e, is a measure of the variability of the distribution.

The foregoing treatment is somewhat idealistic because, in reality, few data bases contain information that allows one to establish a distribution of errors as Figure 9 requires. In fact, for most parameters, typical values for σ_e and b must be estimated with only minimal supporting data. Another complication is that some input data are directly measured whereas others are derived. Therefore, the following factors will be considered in quantifying errors in model input and validation parameters:

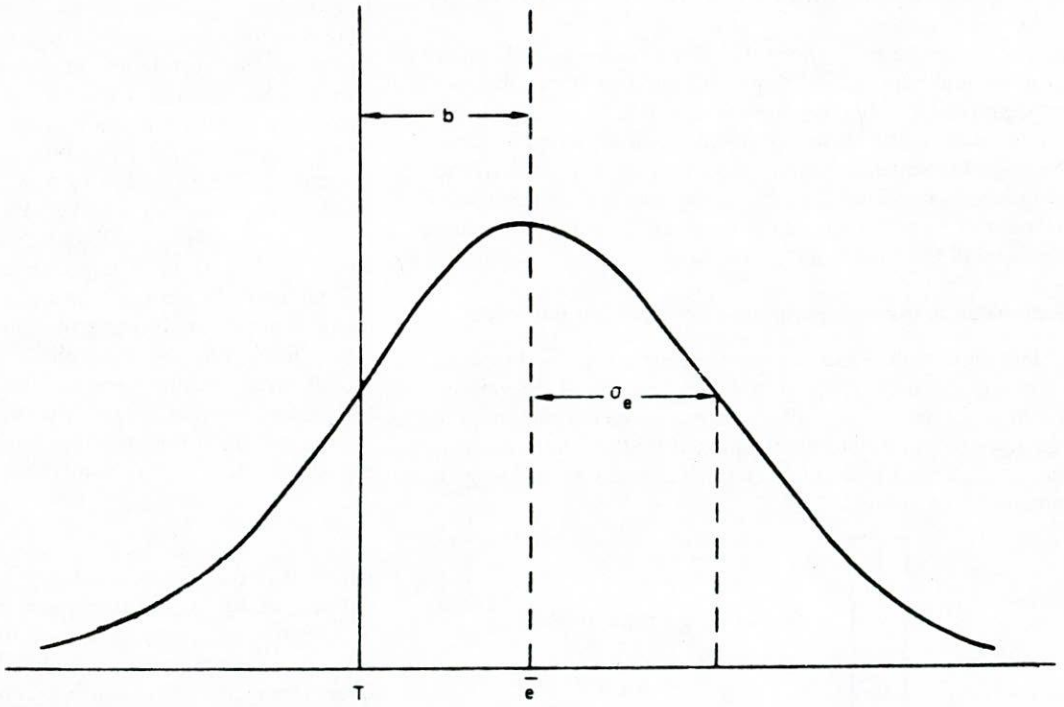


Figure 9. Assumed distribution of error.

1. Magnitude of instrument error (and its representativeness) for directly measured parameters such as winds, traffic counts, vehicle speeds, and ambient CO or tracer measurements.
2. Accuracy of derived input data, such as diffusion coefficients, average vehicle speed, and emission factors.

The "representativeness" of a measurement is a significant point which, coupled with instrument error, adds to the total error for directly measured parameters. Consider the problem of estimating model input parameters for wind speed and direction. Manufacturers claim instrument errors for wind are about 1 percent for speed and 2 to 3 percent for direction. In practice, field installation problems and changes in instrument characteristics with time can add significantly to these error factors. Other factors, such as roundoff of the data or classifying wind directions into sectors, also affect the error. However, all these errors can be reasonably estimated. The representativeness of these winds, however, is somewhat more difficult to evaluate. For some of the data bases, more than one wind measurement is taken. Invariably, multisite wind measurements are not in agreement. The best wind estimate, perhaps, is a spatial interpolation of available measurements leaving out nonrepresentative sites, if necessary. The distribution of wind errors, then, could be approximated from the variation of actual wind measurements about the interpolated value. Some other parameters can be handled in the same way. For instance, errors in dispersion factors would be estimated from variability of source tests.

Because of the nature of certain types of data, some errors must be estimated on the basis of a very subjective engineering judgment. For example, consider "measured" traffic speeds, where verification data do not exist. In this case one must rely on input (published or informal) from the manufacturer of the traffic detector and from the field team that collected the data.

The foregoing paragraphs described some general considerations and approaches to quantifying data errors. Because procedures for collecting the various data bases used in this study differed, the treatment of data base errors must remain flexible. However, it is imperative that the approach to each data base is consistent, to the extent possible. In this way, it is possible to minimize, in the analysis, errors that can arise because of the bias among data bases.

Estimation of the Multiparameter Sensitivity of the Model

Having obtained the sensitivity matrix, it is used to determine the variance of the output as a function of the variance of the input parameters. The variance of the output quantifies the total model output error induced by the combined influence of all the input parameters. We begin by defining the quantities of interest. Let

$$P = \begin{bmatrix} P_1 \\ . \\ . \\ . \\ P_r \end{bmatrix} = \text{vector of input parameters}$$

where r = number of input parameters.

$$\text{Var (P)} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{r1}^2 \\ & & \\ \sigma_{1r}^2 & & \sigma_{rr}^2 \end{bmatrix} = \text{matrix of variances of input parameters}$$

$$x = \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_q \end{bmatrix} = \text{vector of output variables}$$

where q = number of output variables.

$$\text{Var (x)} = \begin{bmatrix} S_{11}^2 & S_{12}^2 & S_{q1}^2 \\ & & \\ S_{1q}^2 & & S_{qq}^2 \end{bmatrix} = \text{matrix of variances of output variables}$$

Both Var (P) and Var (x) are symmetric matrices. Var (P) is not a function of time, but Var (x) is.

The Var (P) matrix contains the variances (σ_{ii}^2) and covariances (σ_{ij}^2) of the input. The variances σ_{ii}^2 are measures of the precision of the input and are the square of the quantity σ_e defined earlier. Thus, the values of the variances must be estimated from such data as wind speed, wind direction, and dispersion parameters. The covariance σ_{ij}^2 is important only if the *ij*th pair of input items is correlated. Such correlation would exist for such derived input as emissions, which is calculated from traffic volume, emission factors, traffic speed, and so on. Thus, the covariance σ_{ij}^2 quantifies the error in the emissions due to traffic volume and traffic speed, respectively. If the *ij*th pair is uncorrelated, $\sigma_{ij}^2 = 0$; an example of ordinarily uncorrelated input is emissions and wind speed.

The Var (x) matrix contains the variance and covariance of the output variables. The expected error of the *i*th output variable is given by s_{ij} (i.e., the square root of the variance). In this case, the *i*th output would correspond to the concentration at the *i*th receptor. The covariance s_{ij} approximately describes the correlation of the errors of pairs of output variates. Knowledge of the correlation of such errors can be extremely helpful because it gives an indication of the limits of the model. Consider two outputs whose errors are highly correlated. The precision of one cannot be improved independently of the other. Conversely, a low or zero covariance of the errors implies a high degree of independence; one would be free to try to improve either variable separately.

Each entry of Var (x) shows the results of the propagation of the input errors through the model. In general, Var (x) is a function of time, so the input errors propagate through the model in an evolutionary fashion. It is of interest to find the expression that relates Var (x) to Var (P) in order to be able to calculate how the input errors propagate. Burns (*2*) has shown that Var (x) is given by

$$\text{Var (x)} = B \cdot \text{Var (P)} \cdot B^T \qquad (5)$$

where B is the matrix of sensitivity coefficients previously defined, and B^T is the transpose of B.

Note that Eq. 5 could be used in two ways. The first way allows one to estimate the output errors associated with input data errors. Thus, this approach considers input errors to be fixed. The second way would allow one to determine the level of precision required by the input in order to satisfy a

prespecified accuracy bound in the output. For example, one could specify that a particular output variable(s) must be within ±10 percent of its mean value. Then, using Eq. 5 iteratively, one could find the precision level required of the input to satisfy the desired output accuracy. This would be extremely useful in planning a field experiment.

The expected error in each output variable induced by the combined effect of all the input uncertainties is given by

$$e_i = s_{ii}$$

where s_{ii} is the square root of the *i*th diagonal element of Var (x), and i denotes the *i*th output variable. For example, if one had N receptors, i = 1,2,...,N, and e_i would be the total error at each receptor. This error includes the effect of input parameters such as diffusivities and emissions and also of the measurement errors in the initial concentrations (if applicable).

INTEGRATION OF SENSITIVITY ANALYSIS AND STATISTICAL METHODS

Ideally, results from the sensitivity analyses are integrated into results for other statistical tests. The standard deviation from the diagonals of the Var (x) matrix described earlier can be thought of as the expected error in the predicted concentrations, which could be handled in much the same manner as the measurement errors associated with observed concentration. The discussion in Chapter Two shows how the measurement errors can be incorporated into various statistics.

Figure 10 graphically shows how results for a sensitivity study can be represented on a simple scatterplot. In Figure 10, every comparison between calculations and measurements has upper and lower sensitivity bounds (limits as functions of model input errors). In this example, only points labeled "2" and "6" fail to intersect the line of ideal fit. The conclusion is that input data errors alone cannot explain the disagreement for these two points. One can take several approaches to establishing the error bars for each comparison in the data base. On one extreme, the Var (x) matrix can be solved every time a prediction is made. At the other extreme, the Var (x) matrix is solved only once to give an average error for the entire set of comparisons. Striking a compromise between accuracy and cost to achieve it, one might work with a limited number of Var (x) solutions, each of which is related to some constraint on the model input parameters (e.g., low wind speeds versus high wind speeds).

The nonlinearity of the sensitivity matrix poses some problems in integrating the sensitivity analysis with the rest of the statistical methodology. For instance, consider the steady-state Gaussian models for which the output varies inversely with the wind speed. Hence, the sensitivity varies inversely with the square of the wind speed. A given percentage error in the wind speed will result in much larger model-output errors at lower wind speeds (say, 2 m/s) than at higher wind speeds (say, 8 m/s). Nevertheless, as a practical matter, for certain statistics it may be acceptable to consider a single average modeling error induced by all model input errors.

Three alternatives in relating the Var (x) matrix to model output errors are follows:

1. Solution of the sensitivity matrix for each model simulation time increment. This may be warranted if one could accurately quantify the model input errors. As discussed later, these errors are only roughly estimated.

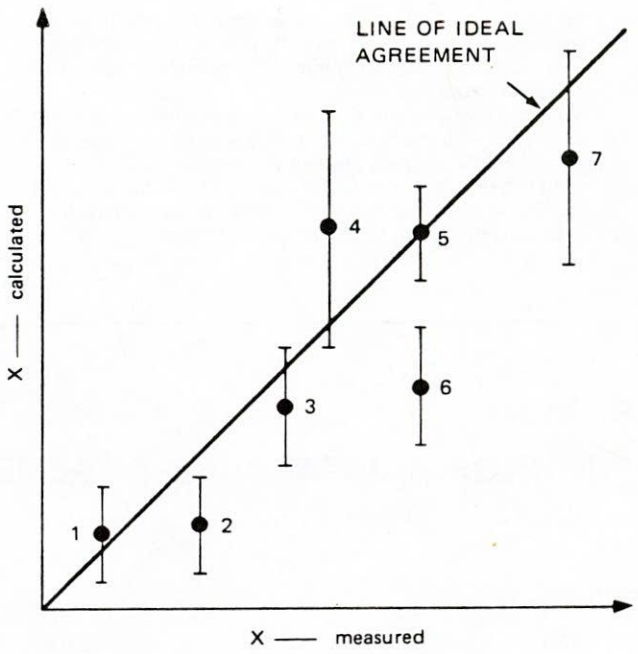


Figure 10. Scatterplot with sensitivity bars.

2. Solution of the sensitivity matrix for a limited set of model input values to determine an "average" model output error. The average error would be compared with average differences between model predictions and observations. This approach is justifiable if it is consistent with the accuracy limitations previously discussed.

3. Combinations of the foregoing approaches could be used. For instance, one may choose to work with the average error for most of the model evaluation. However, for certain large disagreements between observations and predictions, one may choose to solve the sensitivity matrix for those individual cases.

Some of these concepts are discussed and illustrated in Chapter Five.

STEP-BY-STEP SENSITIVITY ANALYSIS PROCEDURE

The following procedure varies somewhat according to the type of model and what is already known about its sensitivities:

Step 1: Quantify the error bounds associated with each of the model input parameters. (This includes the standard input parameters (e.g., emissions, wind speed and direction, atmospheric stability class) as well as certain model constants that are user-adjustable.) These error bounds can usually be determined from values in the literature.

Step 2: Given the error bounds above, determine which of the model input parameters vary in a manner that could produce a significant variation in the model output.

Step 3: The parameters identified in Step 2 become elements in the Var (P) matrix. The step requires the conversion of the error bounds to error variances in forming the Var (P) matrix.

Step 4: For the same parameters, form the B matrix (or B matrices) for the receptor locations of interest. Quantify

the elements in the matrix by appropriate techniques, including analysis of the partial derivatives of the model equation and numerical estimation of the partial derivatives of this model equation.

Step 5: Solve the model sensitivity equation and present the results with the model output values. (Each model output value has an upper and lower sensitivity limit.) To accomplish this, the sensitivity equation can be incorporated into the computer code along with the model formulation. Alternatively, it can be solved separately.

Step 6: Interpret the results in the context of evaluating the model's overall performance.

The above procedure briefly highlights the steps of the sensitivity analysis, but does not emphasize all the options that an individual might select, because such subjective judgment will differ among individuals. However, the calculation of the sensitivity matrix, which is fundamental to the process, requires the execution of the six steps previously listed.

CHAPTER FOUR

FINDINGS—DATA BASE ASSEMBLY

GENERAL

A data base was assembled for use in evaluating the performance of highway air-pollution dispersion models. The data base, which included data from at-grade, elevated (above-grade), and depressed (below-grade) highways, was composed of five data sets provided by SRI International, General Motors Corporation, New York State Department of Environmental Conservation, Texas A&M University, and the California Department of Transportation.

The five data sets are distinguished by a fine level of detail in the measurement of meteorological, pollutant, and traffic data. A user's guide to the data base has been prepared that describes in detail the contents, organization, and format of the data base. The user's guide is included in Appendix G of this report.

The purpose of the data base was to provide input data for the dispersion models as well as the observations with which model predictions will be compared. For this reason, the data base comprises measurements collected in special experimental programs expressly designed to describe highway-related air pollution. This chapter reviews in condensed fashion the data base components and the organization of the archive, and discusses supplemental data needs. A detailed description of the data base is given in Appendix G of this report.

COMPONENTS OF THE DATA BASE

The data base is composed of five data sets provided by:

- SRI International (SRI)
- General Motors Corporation (GM)
- New York State Department of Environmental Conservation (NYS)
- Texas A&M University (Texas A&M)
- California Department of Transportation (CALTRANS)

Table 4 gives the contents of the five individual data bases. Four of the five data bases contain CO measurements, and three have SF_6 data. The SRI and NYS data have both CO and SF_6 data, but GM did not measure CO. The SRI data base has the unusual feature that two different tracer gases, SF_6 and Freon 13-B1, were released by moving vehicles traveling in opposite directions.

All five data bases contain detailed coverage of meteorological conditions, in both time and space. The geographical separation of the various sites suggests that a wide variety of meteorological regimes is represented in the data (one reason for including all five data sets in the model evaluation data base). Judging by the measurement periods, all the seasons are represented to some degree. The precise dates and times of the measurements are documented in Appendix G.

The traffic data are also very detailed, including vehicle speed per lane and vehicular category, in some cases. This detail should lead to more accurate estimates of vehicular source strength. The variety of site locations implies that there will be a wide assortment of age distributions and of mix of vehicle types. Age distribution and vehicle type considerations do not apply to the GM data set because the vehicles consist solely of 1975 model-year passenger cars.

Site type and location are the most important factors affecting the composition of the data base, inasmuch as fixing the site determines the weather, traffic, and other conditions. As Table 4 shows, the combined data sets include seven at-grade and three each of elevated and cut sites, allowing the dispersion models to be tested under a variety of highway configurations.

The five data sets have been evaluated to determine the adequacy of the data for model evaluation purposes. This evaluation included assessing the quantity and quality of the data, as well as identifying the pollution and meteorological conditions represented in each data set. Data quality concerns included assessing the accuracy and precision of the measurements, identifying anomalies in the data, and evaluating the experimental design of the monitoring program and statistical summaries. The data evaluation process included compilation of data inventories and statistical summaries. The inventories and the data summaries, which appear in Appendix G, will assist users to select data for testing models.

DATA-BASE ORGANIZATION

The data archive is divided into four parts:

Table 4. Contents of data bases.

Parameter	Agency				
	SRI	GM	NYS	Texas A&M	CALTRANS
Aerometric Data					
CO	X		X	X	X
SF_6	X	X	X		
Freon 13-B1	X				
Sulfate	X	X			
Meteorological Data					
Horizontal wind speed and direction	X	X	X	X	X
Vertical wind speed	X	X		X	X
Temperature	X	X	X	X	X
Temperature gradient	X	X	X	X	X
Solar radiation	X		X	X	
Relative humidity			X	X	
Precipitation			X	X	
Wind fluctations	X			X	X
Temperature fluctuations			X		
Wind profile	X		X	X	
Cloud cover					X
Ceiling height					X
Pressure		X			
Traffic Data					
Traffic volume	X	X	X	X	X
Vehicle speed	X	X	X	X	X
Vehicle length or type	X			X	
Site Types					
Number of at-grade sites	1	1	1	4	
Number of elevated sites	1			1	1
Number of cut sites	1			1	1
Period of Measurements	Jan-Feb 75 Jul 75 Aug-Sep 75	Oct 75	Apr- May 77	May 76- Dec 77	Apr-Jun 74 Aug 74-Apr 75

1. Tape Index.
2. Format Definitions.
3. Table of Contents of Data Files.
4. Data Files.

The four parts of the archive appear on the tape in the order shown; their descriptions are given in the following.

Tape Index

The index is intended to provide a quick-reference guide to the file structure of the tape. The tape index (see Table 5) gives all the files, showing the name of each file and the corresponding number of blocks in the file, a block being a set of 3200 characters. A user can consult the index to find the position of a file on the tape; knowing the file position will allow the user to skip to the desired file by using job control commands indigenous to the user's computer installation. (Note that the tape index itself is the first file on the tape.)

From Table 5 it can be seen that the archive consists of 30 files as follows:

- Tape Index — File 1
- Format Definitions — File 2
- Table of Contents of Data Files — Files 3 to 16
- Data Files — Files 17 to 30.

End-of-file marks separate all the files.

Format Definitions

A file of format definitions (File 2 in Table 5) has been included to make the tape as self-documenting as possible. This file contains the format specifications of all the files on the tape. Thus, four kinds of format are defined, one each for the four types of file in the archive (tape index, format definitions, table of contents of the data files, and the data files).

A uniform format has been used in the data files. This format follows the SAROAD specifications. (SAROAD is an acronym for the Storage and Retrieval of Aerometric Data system devised by EPA. See Ref. (*3*) for a description of the system.) An advantage of a SAROAD format is that each data record is self-contained, and thus can be read independently of the others; the SAROAD format is well known and its use should facilitate access to the data.

For the data files, records consist of 80 characters (a card image), with 12 hours of data per record, and two records per parameter per day. Records will be stored in blocks of 40 (i.e., 3200 characters/block). Blocks are not designed so that each day of data has an integral multiple of blocks, because, in general, this would require blank-filling the last block of a given day, and one does not expect to have the same number

Table 5. Index of data archive.

File Number	Block Number	File Name
1	1	Tape index
2	3	File Descriptor
3	2	Calt Santa Monica 1 table of contents
4	3	Calt Santa Monica 2 table of contents
5	10	Calt San Diego table of contents
6	1	SRI Hiwy 101 table of contents
7	2	SRI Hiwy I280 B table of contents
8	1	SRI Hiway I280 A table of contents
9	2	NY Huntington table of contents
10	2	GM Mich Prov Grnd. table of contents
11	2	TA&M Dallas-Forst table of contents
12	1	TA&M San Antonio table of contents
13	2	TA&M El Paso table of contents
14	1	TA&M Dallas Motley table of contents
15	2	TA&M Houston-Katy table of contents
16	1	TA&M Houston-Link table of contents
17	111	Calt Santa Monica 1 data file
18	65	Calt Santa Monica 2 data file
19	373	Calt San Diego data file
20	68	SRI Hiwy 101 data file
21	46	SRI Hiwy I280 B data file
22	42	SRI Hiwy I280 A data file
23	43	NY Huntington data file
24	20	GM Mich Prov Grnd. data file
25	22	TA&M Dallas-Forest data file
26	14	TA&M San Antonio data file
27	27	TA&M El Paso data file
28	25	TA&M Dallas Motley data file
29	38	TA&M Houston-Katy data file
30	17	TA&M Houston-Link data file

of blocks per day. Accordingly, data for a new day follow immediately within the same block, after the data for the preceding day. However, the last block in a file has been blank-filled, if necessary, to retain the constant block size of 3200 characters.

The SAROAD format is not applicable to the records contained in the tape index, format definition, and table of contents. Thus, each of these parts of the archive has its own format. Regardless of the format, the block size of 3200 characters has been retained.

Table of Contents of Data Files

The table of contents of the data files is provided to assist the user in locating the data for a particular day in a given data file. Table 6 gives an example for the Santa Monica data file. Referring to Table 5, this table of contents appears in File 3, and the data to which it refers appear in File 17. Table 6 gives the number of the data file and the number of the first block and record in the file where the data for each day begin. For example, the data for April 10, 1974, begin in record 1 of block 1, whereas the data for April 11 begin in record 21 of block 3. One record corresponds to an 80-character line in the block. Hence, the data for April 11 begin on the 21st line of block 3.

Table 6 shows that the header of the table of contents lists the number of the data file (i.e., Data File 1). This data file is File 17 in Table 5. The header also shows the file name and the number code assigned to this site (i.e., 50108). This number code appears in every record of the data file.

The table of contents has been designed so that, if desired, the table can be searched with a computer program provided by the user. As shown in Table 5, there is a table of contents for each highway site.

Table 6. Partial table of contents for CALTRANS Santa Monica site 1.

DATA FILE NUMBER 1

FILE NAME CALT SANTA MONICA 1 50108

SITE TYPE BELOWGRADE

DATE	BLOCK NO.	RECORD NO.
740410	1	1
740411	3	21
740412	5	1
740413	8	21
740415	8	37
740416	11	17
740417	13	33
740418	14	13
740419	16	33
740420	16	37
740421	19	13
740422	19	29
740423	22	9
740424	24	29
740425	25	5
740426	27	25
740427	28	5
740428	28	21
740429	28	37
740430	31	11
740501	33	31
740502	36	11
740503	38	31
740504	41	11
740505	41	27
740506	42	3
740507	44	23
740508	47	3
740509	49	23
740510	52	3

Data Files

There are 14 data files, one for each highway site. Within each file, the data are ordered by month, day, and parameter as follows:

Month M
  Day 1
    Parameter 1: 24 hours
    .
    .
    .
    Parameter N: 24 hours
  Day 30 (or 31)
    Parameter 1: 24 hours
    .
    .
    .
    Parameter N: 24 hours
Month M + 1.

Thus, all the measurements made on a particular day will be together, and the day can be regarded as the basic unit of data. This ordering scheme is particularly convenient, because, in general, one wishes to use the data for a given day for model verification purposes.

SUPPLEMENTAL DATA NEEDS

The data archive assembled in this project contains some limitations for model evaluation purposes. These limitations are discussed in the following, and suggestions are offered for consideration in the planning and design of future measurement programs.

- The archive contains data for seven at-grade sites, and for only three each below-grade and above-grade sites. Thus, more data are needed for the latter two highway geometries.
- In the present data base, parallel and nearly parallel winds and low wind speeds are relatively scarce. Such wind conditions tend to be troublesome for models; hence, it is desirable to test the models using a larger sample of such cases. Thus, additional data are needed for low wind speeds and parallel and nearly parallel winds.
- The accuracy and precision of the measurements are inadequately characterized in four of the five data sets. This is not meant to imply that the instruments were inaccurate or imprecise, but rather that the existing information is generally inadequate to describe quantitatively the accuracy and precision of the measurements. Because uncorrected bias in the measurements will appear as modeling errors in the model evaluation, it is recommended that calibration records be made an integral part of the data base generated in future programs. Moreover, if such records exist for the five data sets studied they should be appended to the present data archive.
- Because the emission input used by the models is a sensitive function of the traffic data, it is recommended that future programs devote considerable effort to characterizing the traffic parameters (volume, speed, mix, and operating mode), such measurements preferably being concurrent with the air-quality and meteorological data.
- To improve emissions estimates, it is recommended that controlled releases and subsequent measurement of concentrations of tracer gasses such as SF_6 be included in future experiments whose data are intended for model evaluation purposes.

CHAPTER FIVE

INTERPRETATION, APPRAISAL, APPLICATIONS

DEMONSTRATION OF EVALUATION METHODOLOGY

This chapter describes the application of the model evaluation procedures previously developed by conducting a preliminary assessment of the performance of six selected highway air-pollution dispersion models. The model evaluation was conducted using a subset of the data base assembled in this project, which is described in Chapter Four and in Appendix G of this report. It is emphasized that the performance evaluation of the selected models was not intended to be definitive. Rather, the assessment was meant solely to illustrate the application of the evaluation methodology in a realistic context by using data collected at a highway site.

The following first discusses the selection of the dispersion models and of the subset of the data base chosen for the application. The results of the model evaluation are then described.

SELECTION OF DISPERSION MODELS

Six models were selected: four are Gaussian and two are numerical. The models were selected following a review of various candidate models that, although thorough, was not intended to be exhaustive. Consequently, selection of a model does not constitute an endorsement, nor is it meant to slight those models not selected. The model selection criteria are discussed in the next section. Subsequent sections list the models considered and identify those selected.

Criteria for Model Selection

Selection of the models was guided by the following criteria:

- *Availability of model and its documentation* — The documentation and computer program for a model must be readily obtainable. This requirement rules out proprietary or undocumented models.
- *Diversity in the modeling approach* — Because of the constraint to selecting only a few models, it was considered desirable to choose models that use different modeling techniques.
- *Performance in previous tests* — Models should have been previously tested with relatively good results.
- *Ease of model implementation and application* — The model must be relatively easy to implement and operate. (These are necessary practical constraints dictated by time and financial restrictions.)
- *Project specifications* — Some models were automatically selected because the scope of work of the project required it.

The application of these criteria to the model selections is discussed below.

Selection of Gaussian Models

The 12 Gaussian models considered are given in Table 7.

Table 7. Gaussian models.

Model Name	Agency	Reference
HIWAY-II	U.S. Environmental Protection Agency (EPA)	(7, 8)
CALINE3	California Department of Transportation	(9)
ROADMAP	SRI International	(4)
GMG	General Motors Corporation	(10)
TRAPS II	Texas A&M University	(11, 12)
AIRPOL-4	Virginia Highway and Transportation Research Council	(13, 14)
TSC/EPA	Transportation Systems Center	(15)
WHM	Walden Research Corporation	(6, 15)
ESL	Environmental Systems Laboratory	(6, 15)
KSC	Kaman Sciences Corporation	(15)
SCI	Systems Control, Incorporated	(15)
TRC	The Research Corporation of New England	(15)

Four of these were selected for evaluation: HIWAY-II, CALINE3, ROADMAP, and GMG. HIWAY-II and CALINE3 were chosen pursuant to the specifications of the project scope of work. Although several of the other candidate models satisfy the selection criteria, ROADMAP and GMG were selected largely on the basis of their good performance in previous tests (*4, 5*). The process that led to these selections is discussed in more detail below, and the features of six of the models are compared.

Of the 12 models considered, only 7 satisfy the availability criterion: TRAPS II, AIRPOL-4, and TSC/EPA, in addition to the above four models selected. The models WHM, ESL, KSC, SCI, and TRC are neither current nor available; in fact, it appears that the ESL, KSC, and SCI models no longer exist. The TSC/EPA model was eliminated based on the results of Downey et al. (*6*), which show that its performance is similar to that of CALINE1 for at-grade highways. Because CALINE1 has been superseded by better models (CALINE2 and, most recently, CALINE3) it seemed reasonable to exclude TSC/EPA from further consideration. This left ROADMAP, GMG, TRAPS II, and AIRPOL-4 as the candidates for the two remaining selections. In a recent test, GMG outperformed AIRPOL-4 in a comparison using data from an at-grade site in New York (Rao et al., 1979). The ROADMAP model has not been compared with other models thus far, but it has been tested with good results using SRI data (Dabberdt, 1977). Both AIRPOL-4 and TRAPS II contain a number of unique and interesting features in their design, and merit further testing in the future.

Table 8 compares various characteristics of HIWAY-II, CALINE3, ROADMAP, GMG, TRAPS II, and AIRPOL-4. The table shows that the diversity criterion is satisfied by the four models selected and, indeed, by all six models. The models are collectively applicable to a variety of highway configurations; each has a different formulation; and each accounts for vehicle-induced turbulence in a way that is unique for each model. All the models are fairly well documented and are easy to implement and use. All but GMG are available in a user-oriented package of algorithms. (The GMG model is considered to be research-oriented because it lacks a user's guide and its computer program, although easy to use, is not designed with a view toward highway-planning applications.) All the models share essentially the same types of input, and all the input data are normally available, with one exception: GMG requires a buoyancy flux parameter that is not readily available. (Chock (*10*) has derived such a parameter using data from the GM experiment, and has suggested a method for adapting this derived value to other applications (*16*).) Only CALINE3 and TRAPS II explicitly incorporate a terrain roughness parameter. For CALINE3, the parameter is an input; for TRAPS II it is built into the program. The user's guide for CALINE3 (*9*) provides a table of roughness parameters for different kinds of terrain.

Selection of Numerical Models

The numerical models considered are given in Table 9. The two selected are MROAD2 and LAMB.

The scope of work of the project required selecting one of either MROAD2 or MROAD3, DANARD, or RAGLAND. A recent comparison of MROAD2, DANARD, and RAGLAND using the New York data base (*5*) revealed that MROAD2 and DANARD give similar results, and that both outperformed RAGLAND. Thus, either MROAD2 or DANARD was suitable. MROAD2 was selected because it was readily obtainable from the Oregon Department of Highways, but DANARD certainly merits inclusion in any future projects aimed at evaluating models. MROAD3 was rejected because it does not satisfy the performance and implementation selection criteria.

The LAMB model was selected, first, because its performance in previous verification attempts using GM data has been very good and, secondly, because the LAMB model is unlike any other numerical model considered in that it uses the Monte Carlo method for solving the integral equation that describes the mean concentration field. The model thus differs substantially from finite-difference K-theory models such as MROAD2. Rather than test two versions of a K-theory model, it was decided to test one such model (MROAD2) and another that uses a significantly different approach.

Of the other models given in Table 9, GMN, AVQUAL, CEM, EGAMA, INTERA, and LOC are proprietary and unavailable. This left LAMB, ESK, and ROADS as potential candidates for the second selection. Both ESK and ROADS are finite-difference models that were passed over in favor of LAMB for reasons already stated. However, ESK has several unique features, notably a new approach for incorporating the effects of vehicle-induced turbulence, that make it a strong candidate for inclusion in future model evaluation efforts. In recent tests conducted using New York data, both MROAD2 and DANARD outperformed ROADS (*5*).

From a user's standpoint, none of the numerical models is packaged as well as the Gaussian models previously discussed. The technical documentation for MROAD2 is inadequate, but LAMB's is better. The technical report describing the LAMB model contains a brief user's guide describing the input structure of the model, but the version of the model received by the research team did not always conform to the published description. MROAD2 has a very brief user's

Table 8. Features of six Gaussian models.

Model Name	Roadway Configuration			Minimum Wind Speed (m/s)	Special Treatment For Parallel Wind	Treatment of Line Source			Vehicle-Induced Turbulence	Terrain Roughness	Special Input	Remarks
	Cut	At-Grade	Elevated			Virtual Point Source	Finite Line	Infinite Line				
HIWAY-II	X	X		$\geq$ 0.3	Yes	X			Implicit in σ_z and σ_y	No	Width of median, width of top of cut section	Applications-oriented program and documentation.
ROADMAP	X	X	X	$\geq$ 0.5	Yes			X	Implicit in σ_z and z	No		Applications-oriented program and documentation.
CALINE3	X	X	X	$\geq$ 1.0	No		X		Implicit in mixing zone width and σ_z correction	Yes	Surface roughness parameter; σ_θ for special applications; settling and deposition velocities	Applications-oriented package. Includes treatment for bridges.
GMC		X		0	Yes		X		Implicit in σ_z and a wind correction factor	No	Buoyancy flux	Does not require σ_y. Uses only three stability conditions. Research-oriented computer program and documentation; no user's guide available.
TRAPS II	X	X	X	> 0.5 at 10 m	No	X			Implicit in fitted equation for CO concentration at edge of roadway	Yes	Height of wind speed measurement and wind speed at given height; roughness height	Uses non-Fickian solution of dispersion equation in a semiempirical approach. Application-oriented computer program and documentation.
AIRPOL-4	X	X	X	0	No	X			Implicit in σ_z and σ_y	No		Applications-oriented program and documentation.

Table 9. Numerical models.

Model Name	Agency	Reference
MROAD2, MROAD3	Oregon State Department of Highways	(17, 18)
LAMB	U.S. Environmental Protection Agency	(19)
DANARD	University of Waterloo	(20)
RAGLAND	University of Wisconsin	(21)
ESK	U.S. Environmental Protection Agency	(22, 23)
ROADS	Oregon Graduate Center	(24)
GMN	General Motors Corporation	(25)
AVQUAL	AeroVironment, Incorporated	(26)
CEM	Center for Environment and Man	(15)
EGAMA	Environmental Research and Technology, Incorporated	(27)
INTERA	Intera	(15)
LOC	Lockheed Missiles and Space Company	(15)

guide that was marginally useful. Brief descriptions of MROAD2 and LAMB are given in the following.

MROAD2

MROAD2 is a K-theory model that solves the two-dimensional advection-diffusion equation using finite differences. The governing equation is:

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + w\frac{\partial C}{\partial z} = \frac{\partial}{\partial x}K_x\frac{\partial C}{\partial x} + \frac{\partial}{\partial z}K_z\frac{\partial C}{\partial z} + S \quad (6)$$

where

C = pollutant concentration;
u,w = mean wind speed in x and z directions, respectively;
K_x,K_z = turbulent diffusion coefficients in x and z directions, respectively; and
S = emission source strength.

The two-dimensional formulation implies that the model is applicable to a vertical cross section of a highway. The model requires specification of the wind field and the diffusivity field at every point in the (x,z) plane.

LAMB

This model differs from the K-theory approach in that it tracks the movement to a receptor of particles released over the roadway. The movement of a particle is governed by a probability density function (pdf) of the form $p(r,t/r_s,t')$, which is the pdf that a particle released at point r_s at time t' will be found at point r at time t. The governing equation of the model is:

$$C(r,t) = \int_0^t \int_{-\infty}^{\infty} p(r,t|x_s,y',z_s,t')\, S\,(y',t')\, dy'\, dt' \quad (7)$$

where

$C(r,t)$ = mean pollutant concentration at point $r = (x,y,z)$ at time t;
$p(r,t|x_s,y',z_s,t')$ = pdf that a particle released at time t' at point x_s,y',z_s arrives at point r at time t; and
$S(y',t')$ = function describing the distribution of vehicle emissions.

When solved under certain particular assumptions, Eq. 7 yields the K-theory model; those assumptions have been foregone in the LAMB model. Instead, particular forms of the pdf and S function are used, and the integral equation is solved using a Monte Carlo approach. The resulting solution consists of an ensemble of simulated particle trajectories that are averaged to yield the desired mean concentration.

DATA-BASE CONSIDERATIONS

Demonstration Data Set

For this application, the data set for one at-grade highway site from the Texas data base was selected. The site is one of two located in Houston (the other is for a below-grade roadway), and has been designated as Houston 1. Appendix G contains a complete description of the site as well as a data inventory.

The selection of this data set for demonstration purposes was guided by several factors. First, the demonstration data set must be selected in a manner consistent with the project's scope and associated time and financial constraints. Thus, although it would be ideal to apply the evaluation procedure to the entire data base, rather than to a subset, such an undertaking would substantially exceed the resources of the project. The Houston 1 data set satisfies this criterion because its size is manageable, yet sufficiently large to obtain meaningful statistics.

Second, the data set should be well documented, especially with respect to the probable magnitude of measurement errors. All the data sets in the archive are reasonably well documented, but the Texas data base is noteworthy because it includes measurement error information; such information will be used to obtain performance statistics adjusted for observational uncertainty.

Third, for this demonstration it is desirable to use data that have not been previously used to test the six selected models. In this way knowledge is gained about the performance of the models when applied to a variety of data bases. It is believed that none of the six models has been tested using the selected data.

Figure 11 shows the instrument layout at Houston 1. As the figure shows, the Houston 1 site contains 12 instruments that monitored CO. The evaluation procedure will compute the performance statistics for each individual monitor and for all the monitors combined. Table 10 gives the number of CO measurements available for comparison with predictions at each monitor. The amount of data is smaller than that shown in the inventory in Appendix G because simultaneous measurements of several variables are required (e.g., CO, wind

speed, traffic density, and the like). The simultaneity constraint reduces the size of the data set because for a given hour, missing data for any one variable create a gap in the entire data record for that hour. The size of the data set is further reduced for the evaluation statistics that estimate correlations (S_3, S_4 and S_5) because these use the logarithm of the concentration, and measurements reported as zero concentration must be deleted. Table 10 shows that monitors 8 and 12 reported a total of 26 hours with CO equal to zero. Hourly CO concentrations at Houston 1 ranged from 0 to 19 ppm. Thus, the one-hour ambient air quality standard for CO, which is 25 ppm, was never exceeded in the test data set for Houston 1. (Recall that the Houston 1 data set for CO is larger than the test data set described in Table 10. The 0- to 19-ppm range applies only to the test data set.)

Note that the frontage roads shown in Figure 11 could not be included in the analysis because no traffic data are available.

Table 10. Number of hours of CO data for each monitor at Houston 1 site.

Monitor Number	Number of Hours	
	Raw Data	Log-Transformed Data
1	79	79
2	77	77
3	78	78
4	70	70
5	76	76
6	73	73
7	45	45
8	76	52
9	64	64
10	60	60
11	64	64
12	40	38
Total	802	776

Estimating Carbon Monoxide Emission Factor

Each of the dispersion models being tested requires the user supply, as part of the model input data set, a pollutant emission rate for the roadway as a whole or for each lane of the roadway being modeled. To compute these emission rates, an emission factor is multiplied by the volume of vehicles traveling the roadway. At the present time two mobile-source emission estimation procedures are accepted by the EPA, the emission methodology based on the Federal Test Procedure (FTP) (*28, 29*) and the modal methodology (*30*). A third type of procedure for computing emission factors, one not currently endorsed by the EPA, involves the concept of mass balance (*11, 13*). For this study, the FTP method of emission factor estimation was selected. Each of these three-methodologies for the computation of mobile source emission factors will be outlined briefly below, followed by a discussion of both the relative merits and the disadvantages of each and the rationale for choosing the procedure to be used in the current study.

In the FTP methodology, motor vehicles are divided into six groups: light-duty vehicles (automobiles), two classes of light-duty trucks, heavy-duty gasoline-powered vehicles, heavy-duty diesel-powered vehicles, and motorcycles. An emission factor is computed separately for each vehicle group, weighted according to the percentage of the total vehicle-miles driven by that group; the weighted emission factors for all groups are combined to produce a composite emission factor.

The equation that describes the FTP emission factor for light-duty vehicles (LDV) is:

$$Enpstwx = SUM(Cipn*Min*Ripstwx*Aip*Lp*Uipw*Hip) \quad (8)$$

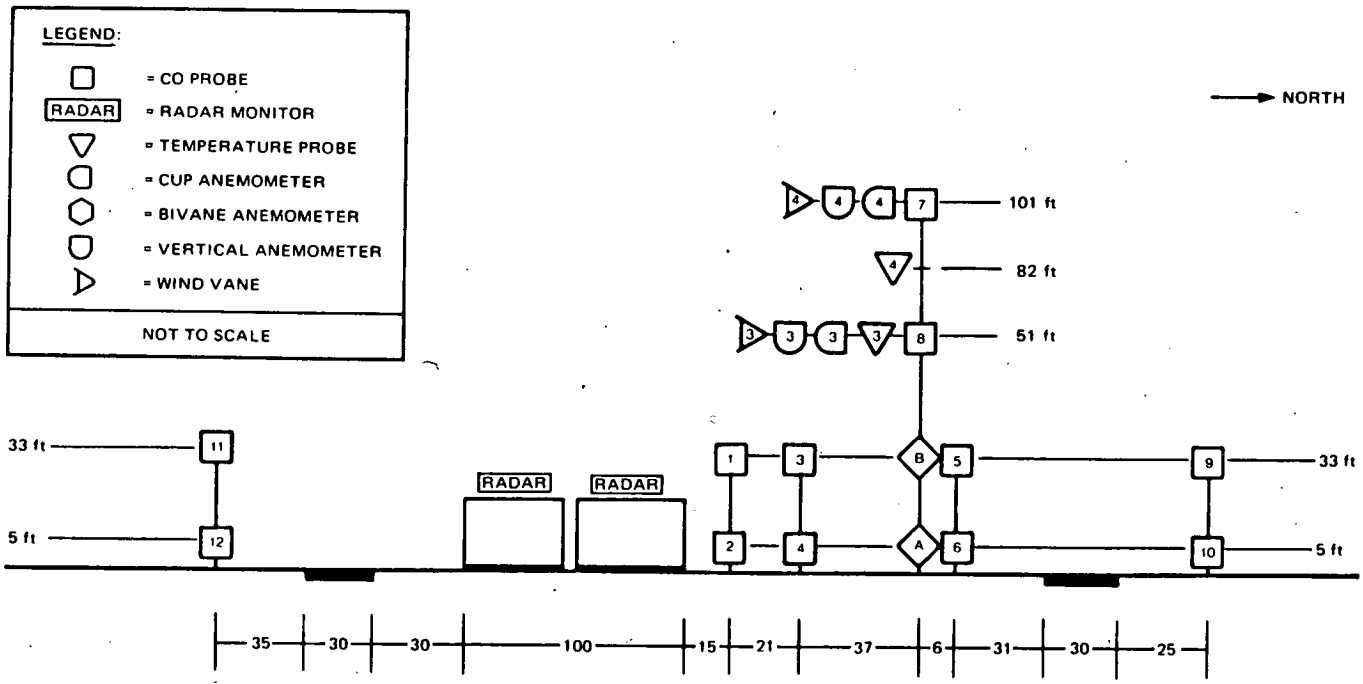


Figure 11. Instrument layout at Houston 1 site.

where all lower case letters are subscripts and

SUM ( ) = summation over model year (i), from the calendar year for which emission factors are being calculated (i = n) to the calendar year 19 years previous (i = n − 19);
Enpstwx = composite emission factor in gm/mi for calendar year n, pollutant p, average speed s, ambient temperature t, fraction cold-start operation w, and fraction hot-start operation x;
Cipn = the FTP (1975 Federal Test Procedure) mean emission factor for the *i*th model year light-duty vehicles during calendar year n and for pollutant p;
Min = the fraction of annual travel by the *i*th model year LDVs during calendar year n;
Ripstwx = the temperature, speed, and hot/cold correction factor for the *i*th model year LDVs for pollutant p, average speed s, ambient temperature t, fraction cold-start operation w, and fraction hot-start operation x;
Aip = the air-conditioning correction factor for the *i*th model year LDVs for pollutant p;
Lp = the vehicle load correction factor for pollutant p;
Uipw = the trailer towing correction factor for the *i*th model year LDVs for pollutant p and fraction of cold-start operation x; and
Hip = the humidity correction factor for the *i*th model year LDVs for pollutant p.

For the light-duty truck classes (LDT), the emission factor equation has the above functional form with the exception of the trailer towing factor, which is deleted. Of course the values of the components of the equation differ in some cases for LDV and LDT.

The equation describing the emission factors for heavy-duty gasoline-powered vehicles (HDG) and heavy-duty diesel-powered vehicles (HDD) is:

$$\text{Enpsoq} = \text{SUM(Cipn*Min*Vips*Pipnoq)} \qquad (9)$$

where

Vips = the speed correction factor for the *i*th model year HDG vehicles for pollutant p and average speed s; and
Pipnoq = the truck characteristic correction factor for the *i*th model year HDG vehicles for pollutant p, calendar year n, truck weight o, and weight/power ratio q.

The individual components of Eq. 9 differ for HDG and HDD vehicles. Finally, the emission factor for motorcycles (MC) is given by:

$$\text{Enpstwx} = \text{SUM(Cipn*Min*Ripstwx)} \qquad (10)$$

where each component of the equation applies to motorcycles.

The FTP methodology described here has been computerized by EPA (*29*), and a version of the program (called MOBILE1) was used to compute the emission factors.

The basic premise of the modal model is that the instantaneous emission rate, $\dot{e}(t)$, can be described as a function of vehicle speed v and acceleration a. Because speed and acceleration are functions of time, the emission rate function can be expressed as

$$\dot{e}(t) = \dot{e}[v(t),a(t)] \qquad (11)$$

Acceleration to or from a given speed is assumed to be a perturbation to the steady-state emission rate, and two equations are used to describe the emission rate:

$$\dot{e}_A(v,a) = b_1 + b_2 v + b_3 a + b_4 av + b_5 v^2 + b_6 a^2 + b_7 av^2 + b_8 a^2 v + b_9 a^2 v^2 \qquad (12)$$

and

$$\dot{e}_s(v,o) = b_{10} + b_{11} v + b_{12} v^2 \qquad (13)$$

where Eq. 12 describes the nonzero acceleration emission rate and Eq. 13 describes emission rates for steady speeds. Note that the idle emission rate equals the coefficient b_{10}.

The instantaneous emission rate function for a given vehicle and pollutant is given by:

$$\dot{e}(v,a) = h(a)\dot{e}(s) + [1 - h(a)]\dot{e}_A(v,a) \qquad (14)$$

where h(a) is a weighting function dependent on acceleration and bounded by the values of 0 and 1, which allows for a smooth, continuous transition from steady speed to acceleration and deceleration emission rate functions.

Inputs to the original modal model include calendar year, vehicle model-year mix for various locations (low or high altitude, in California or outside California) and second-by-second speed data for the driving sequence being modeled. The two preceding equations are solved for each second, the results are taken as one-second integrated emissions, and the model output is the total emissions over the driving sequence. Eighteen vehicle groups can be included in the vehicle mix; deterioration as a function of calendar year is included in the emission calculations. Temperature and cold-start correction factors are also included.

The basic assumption of the mass balance concept for computing emission factors is that the difference between the amount of pollutant flowing past vertical planes located downwind and upwind of the roadway equals the amount generated by the traffic on the roadway. Basic to this theory is the assumption that there are no other pollutant sources or sinks for the material between the two planes. (Because highways are often approximated as line sources, the planes on each side of the roadway are reduced to lines, and calculations are made from measurements made at several heights on towers located on each side of the roadway. A major restriction is that the roadway pollutant plume must be entirely defined within the height of the towers.)

To compute emission factors, the upwind concentrations measured at each height on the tower are subtracted from the downwind concentrations at the corresponding heights. The product of each concentration and the component of the wind vector normal to the roadway is plotted as a function of

height, and the resulting curve is graphically integrated to obtain pollutant mass per unit time and per length of roadway. Division of this quantity by vehicle volume on the roadway yields the emission factor.

Several advantages and disadvantages are common to both the FTP and modal models. Both methodologies are endorsed by the EPA for use in estimating hydrocarbon, carbon monoxide, and nitrogen oxides mobile-source emissions. The methodologies were both derived from a large data base that is continually being updated. On the negative side, however, neither the FTP nor the modal model has been comprehensively validated. Also, estimates made with the models are based on the average emissions measured for the test vehicles. The emissions from the vehicle population on the roadway being modeled may differ from the average emissions of the test vehicles.

One of the chief advantages of the FTP model is the ease with which it can be used. Most of the required input data are generally available, and national average values supplied with the model may be used for many of the variables for which more site-specific data cannot be found. However, some of the input parameters are difficult to estimate, and error in such parameters can produce error in the emission factors. (See Visalli (*36*) for a discussion of the effect of variations in the vehicular age and classification distributions.) This same problem may exist with the modal model, although to a lesser extent.

The advantage of the modal model is that it can simulate emissions from a specific driving sequence, with any set of accelerations, decelerations, idling, and steady-speed driving. However, to gain this privilege, the user must input a second-by-second speed/time profile for each direction of the roadway. In practice, such a profile is a part of only very few data sets.

The major advantage of the mass balance technique is that the emission factor is derived from data characteristic of the vehicle population and the vehicle operating conditions on the roadway to be modeled. However, this advantage is somewhat mitigated by the fact that the measured concentrations at the towers could (and in a major metropolitan area certainly do) arise partially from emissions on surrounding roadways. Because carbon monoxide levels vary rapidly over short spatial distances, the assumption that the background level of pollutant at the tower on one side of the roadway is equal to that at the tower on the other side of the roadway is not necessarily valid, particularly as the wind becomes less normal to the roadway.

Examination of the available methodologies for estimating mobile source emission factors has shown that for a number of reasons the FTP methodology is most applicable to this study. First, and probably most importantly, by far the majority of highway dispersion model users choose the FTP methodology to obtain emission factors because of the ease of application of MOBILE1 and because the methodology is approved by EPA. Because this study should be directed toward producing results that will be pertinent to the application of model users, the use of the FTP methodology is appropriate.

Also, the modal model input data requirements include a second-by-second speed/time profile that is not available in the data bases being used in this study. Therefore, use of the modal methodology is not deemed advantageous for the current model verification program.

Finally, the mass balance emissions methodology has not been used for a number of reasons:

- The methodology can only be applied at sites that have been heavily instrumented, such as field experiment sites.
- The results of the methodology are site-specific, and sufficient data do not exist to generalize the results either for other sites, or for a wide range of wind/roadway angles and meteorological conditions. Thus, use of the mass balance technique in sensitivity tests, for instance, would not give the general model user information about the emissions methodology most likely to be used.
- In the mass balance computations made by Bullin (*31*) the variation in computed CO emission factors is seen to be considerable and in some cases inexplicable. For instance, a CO emission factor variation of as much as a factor of two or more is shown between computations made at one site for two times of day, separated by about 20 min and having nearly equal wind/roadway angles and 5-min traffic volumes.
- From a statistical point of view, it is believed that emissions estimates obtained using the mass balance method cannot be used as input for dispersion models whose predictive performance is to be evaluated. The problem stems from the fact that the CO emissions obtained by the mass balance method are based on the same set of measured CO concentrations against which the model's predictions are to be compared. Thus, the emissions input of the dispersion model is a function of the CO concentrations that the model is trying to predict. Under such circumstances, the model's performance would be biased because the model input is not independent of the output.

The FTP methodology requires for each roadway the input data listed as follows:

- Calendar year.
- Vehicle type distribution.
- Model year distribution (for each vehicle type).
- Ambient temperature.
- Average route speed.
- Percentage of cold-starting and hot-starting vehicles (for LDV, LDT, and MC).
- Fraction of vehicles operating under an additional 500-lb load (for LDV and LDT).
- Fraction of LDV towing a trailer (1000 lb).
- Vehicle weight and engine displacement (for HDG and HDD).

These data are part of the Texas data base (see App. G).

It is recognized that estimates of CO emissions obtained using MOBILE1 can be inaccurate. To the extent that such a basic model input is unreliable, it will be difficult to evaluate the performance of a dispersion model in an absolute sense. However, the performance of the various models can still be assessed in a comparative sense. Moreover, the problem of assessing model performance in the presence of inaccurate emissions input can be addressed by the use of sensitivity analysis, which quantifies the effect of input and other errors on the model's predictions.

The principal objective of this project was to develop methodology for evaluating model performance, hence the methods developed are independent of the emissions model used. As new, more accurate emissions models are developed, these methods can be used anew to examine the predictive performance of the dispersion models using the new emissions input.

PRELIMINARY EVALUATION OF SELECTED MODELS

Structure of the Model Evaluation Procedure

The model evaluation procedure (MEP) is designed to meet the needs of a user who wants to assess the performance of a dispersion model (or models) that estimates highway-related air pollution. The MEP described below provides the logical structure that begins with the definition of a highway-related problem, and ends with a measure of the performance of the model (or models) in the context of the problem.

The MEP developed in this study consists of four steps, depicted in Figure 12. The first step defines the parameters of the problem under consideration. The second step specifies the model or models to be tested. The third step consists of selecting a subset of the full data base for use in the model evaluation, hence the apellation "working data base." It should be noted that the working data base could comprise the full data base. The working data base, as well as the full data base, may consist of data from several highway sites. If so, the component statistics and the FOM for each model being tested would be computed separately for each site. The fourth step consists of selecting one or more of three types of model evaluations that can be performed. The feedback between Steps 3 and 4 results from the fact that the required number of data points in the working data base may be influenced by the type of model evaluation selected. The internal structure of each of these steps is described as follows.

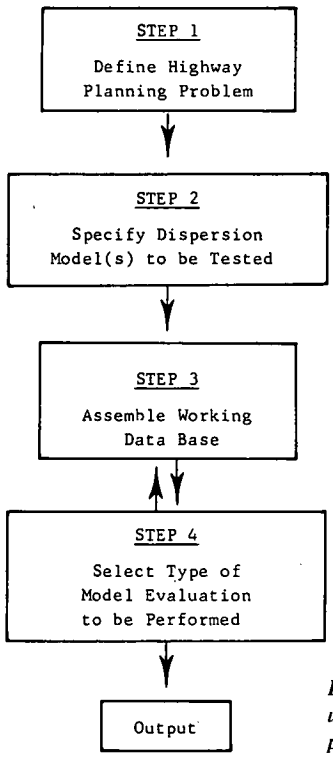


Figure 12. Basic structure of evaluation procedure for highway dispersion models.

Step 1

Because different models used to predict highway-related air pollution do not necessarily perform equally well in all circumstances, the first step in the MEP defines the parameters of the planning problem that may influence the absolute and relative predictive performance of the models. These parameters will also help to define the composition of the working data base. There are several parameters of interest, including: type of highway configuration—at grade, elevated (above-grade), depressed (below-grade); distance scale of roadway impact—next to roadway, local neighborhood, regional; meteorological conditions—atmospheric stability, wind speed and direction, inversion height, terrain roughness; time of year; and traffic count or emission rate.

Step 2

The air pollution dispersion models to be tested are specified in Step 2 (which may be concurrent with Step 1). Because the MEP entails comparing model predictions against observations, it is only necessary that the models be specified one at a time. Because the FOM used for judging model performance is an absolute index (e.g., "the model scored 8.5 on a 10-point scale") rather than a relative index (e.g., "the model was the third best among 12 tested models"), the user may find that the first model evaluated is adequate for the problem under consideration. In this event it is not necessary to evaluate other models, unless of course the user is concerned with identifying the best model or comparing the accuracy of several models. (Other researchers (Darling et al., *15*) have investigated the clustering of models by comparing the predictions of a model against predictions of other models. The clustering approach addresses the questions of which models are similar rather than which agree best with observed data, and requires the specification of more than one model at a time.)

Step 3

Having defined the parameters of the problem and the models to be tested, the third step is to assemble the relevant portions of the full data base into a working data base. For example, if the user is interested only in an elevated highway configuration, only this configuration would be included in the working data base.

As part of Step 3, a summary of the working data base should be prepared. This summary should include information on the number of observed concentrations contributed by the individual data bases, the distribution of these data points by type of highway configuration, distance scale, time of the year, wind direction, wind speed, estimated emission rate, and a histogram of the observed concentrations. This characterization of the working data base is important both for documentation purposes and for assisting the user in deciding whether the data base is representative of the problem being considered. For example, it may be that the working data base is evenly distributed over the stability classes A through F, when the area under consideration predominantly experiences stability classes A through D

and only occasionally experiences stability classes E and F. In this event, the user may desire to compute the FOM separately for the two ranges of the stability class by subdividing the working data base, and later compute a weighted average of the two FOMs (where the weights would be the expected frequency of occurrence of the two ranges of the stability class in the area under consideration).

Step 4

The fourth step is to select and perform the desired type or types of evaluation analysis. As mentioned in Chapter Two, a distinction has been made between three types of test procedures that examine different, but complementary, aspects of model performance, namely:

- Accuracy analysis—measures the predictive performance of the model.
- Diagnostic analysis—identifies conditions associated with inaccuracies in the model's predictions.
- Sensitivity analysis—quantifies the model's response to uncertainty in the input data.

The components of Step 4 are shown in Figure 13.

In highway planning problems, the first function, evaluating the model's accuracy, is usually the most crucial. If the model is found to be inaccurate it may be discarded, or a diagnostic analysis can be performed to uncover the conditions in the working data base associated with the prediction inaccuracies. For example, the diagnostic analysis may reveal that the FOM is being adversely affected by large differences between observations and predictions on days with low wind speed or by a time lag of several hours between observations and predictions. In the former case the working data base could be fashioned to exclude periods of low wind speeds (and another model used for these excluded periods). In the latter case, a modification of the model might be attempted. Even in cases where the FOM is high (implying that the accuracy of the model is high), it might be advantageous to perform a diagnostic analysis to alert one to potential hazards in using the model. Finally, the sensitivity of the model to errors in the meteorological, emissions, and other input should be established, particularly if the user suspects that the input errors are more serious in the geographical area under consideration than in the working data base. The sensitivity of the model to input errors may also account for inaccuracies in the predictions, so that this avenue should also be explored if the FOM is low. Sensitivity analysis may be used to identify potential modifications in a model that is particularly sensitive to a specific input parameter or to establish guidelines on the allowable errors in model input for future data bases.

To demonstrate the application of the model evaluation techniques, one cannot strictly adhere to the sequence of steps contained in the MEP. Thus, the focus of the subsequent discussion will be on Step 4 in order to illustrate the use of accuracy, diagnostic and sensitivity analyses. Consequently, for purposes of this report, it is assumed that the highway planning problem indicates a need to test the models under at-grade roadway conditions, and that this has led to specifying the Houston 1 data set to be the working data base. Of course, the models to be tested in this study were selected for reasons that are not necessarily relevant to the MEP.

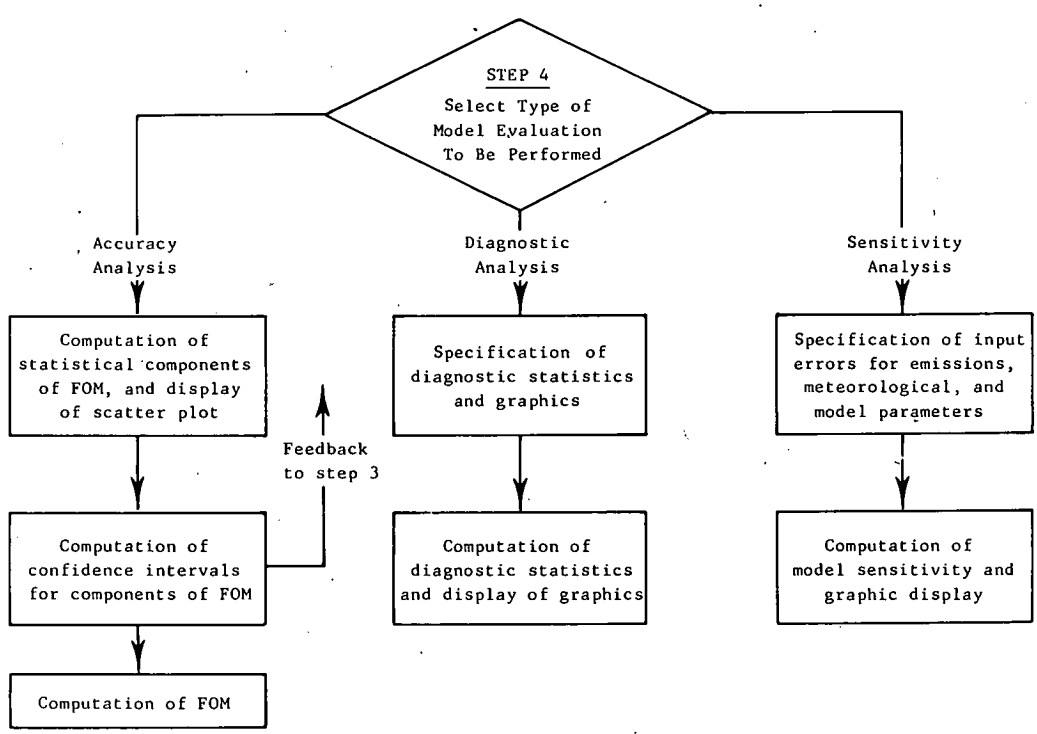


Figure 13. Components of step 4 of model evaluation procedure.

Statistical Methodology Measurement of Performance

For ease of reference, the six performance statistics defined in Chapter Two are briefly described as follows:

- S_1 ratio of the largest 5 percent of the observed concentrations to the largest 5 percent of the predicted concentrations.
- S_2 difference between the predicted and observed frequency of exceedance of a CO threshold concentration.
- S_3 Pearson's correlation coefficient between observed and predicted concentrations for the entire data set.
- S_4 temporal correlation coefficient.
- S_5 spatial correlation coefficient.
- S_6 root-mean-square error of observed and predicted concentrations.

Associated with each S_1 is an FOM denoted by F_i, i = 1, 2, . . . , 6; F_i is a function of S_i. The value of each F_i ranges from 0 to 10, where 0 denotes essentially no agreement between model predictions and observation, and 10, the highest level of agreement, but not necessarily perfect agreement. The 0 to 10 scale is both arbitrary and convenient. A composite FOM is defined as a function of the individual F_i; its value also ranges from 0 to 10.

The six selected models were run and their predictions of hourly CO concentrations were compared with the observed CO levels in the Houston 1 data set. In the following, performance of the models is assessed in terms of S_i and F_i. The meaning of S_i and F_i is examined, different ways of obtaining a composite FOM are investigated, and various approaches for ranking models according to performance are considered. Also demonstrated is the use of some diagnostic tools to complement the performance measures S_i and F_i.

Table 11 shows S_1 through S_6 and their corresponding FOMs for the six models; their values are unadjusted for the effect of observational error, a topic that will be examined later. The optimum value of S_1 is one: $S_1 > 1$ indicates a tendency to underpredict high concentrations, and $S_1 < 1$ a tendency to overpredict. The table shows that five of the six models tended to underpredict, the exception being the LAMB model, which overpredicts by a considerable margin. CALINE3 and GMG differ only slightly in their performance as measured by S_1; both tend to underpredict by about 25 percent, which is a reasonably good performance. The value of F_1 for these two models is 8.1 and 7.9, respectively, indicating that both rank in the upper quarter of the FOM scale. ROADMAP underpredicts by a 46 percent margin, which places it in the top third on the FOM scale. MROAD2 and HIWAY-II underpredict by more than 100 percent, which puts them in the bottom half of the FOM scale for F_1. The LAMB model not only overpredicts, but does so by a large margin. This indicates that the model is not working properly, an inference that will be confirmed by its performance in other areas. The reasons for, and implications of, the LAMB's model performance are discussed later.

The apparent tendency to underpredict indicated by S_1, for five of the models, could be caused by low emissions inputs. Bullin et al. (*11, 31*) indicate that MOBILE1 emissions estimates tend to be low, which could explain the results obtained. It is emphasized that a model's tendency to overpredict or underpredict could be caused by inaccuracy in the model inputs, particularly the emissions. Consequently, the statements in the preceding paragraph about the numerical margin of overprediction and underprediction are not intended to imply that such behavior is inherent to the model because the numerical margin will change if a different set of inputs is used. Rather than focus on the absolute numbers, attention should be paid to the different tendencies exhibited by the various models for a common set of inputs. These caveats should be borne in mind in the discussions that follow.

Statistic S_2 measures a model's ability to predict the probability of exceeding a CO concentration threshold specified by the user. The value of S_2 ranges from −1 to +1. S_2 = 0

Table 11. Model evaluation statistics and figure of merit.

Model Name	Statistic[1]						Figure of Merit (FOM)						Overall FOM[c]
	S_1	S_2[a]	S_3	S_4	S_5	S_6[b]	F_1	F_2	F_3	F_4	F_5	F_6	
HIWAY-II	3.12	-0.0349	0.324	0.224	0.400	3.23	3.2	8.2	3.2	2.2	4.0	2.1	3.6
CALINE3	1.24	-0.0150[d]	0.415	0.526	0.337	2.15	8.1	9.4	4.1	5.3	3.4	3.8	5.5
ROADMAP	1.46	-0.0158[d]	0.154[d]	-0.0742[d]	0.515	3.11	6.9	9.2	1.5	0	5.2	2.3	4.2
GMG	1.26	-0.0212[d]	0.211	0.0894[d]	0.338	3.23	7.9	9.1	2.1	0.9	3.4	2.2	4.3
MROAD2	2.14	-0.0324	-0.00021[d]	0.189[d]	-0.191	3.47	4.7	8.4	0	1.9	0	2.0	3.1
LAMB	0.0261	0.943	0.145[d]	-0.0601[d]	0.430	143.0	0.3	0.6	1.5	0	4.3	0	1.1

[1] Significant at the 0.05 level or better unless otherwise specified.

[a] CO threshold equal to 8 ppm.

[b] Units are ppm.

[c] FOM = $[(F_1 + F_2)/2 + (F_3 + F_4 + F_5)/3 + F_6]/3$.

[d] Not statistically significant.

denotes perfect agreement between prediction and observation, and ±1 correspond to total disagreement. $S_2 > 0$ indicates a tendency to overpredict, and $S_2 < 0$ a tendency to underpredict. The threshold concentration of greatest interest is the 1-hour NAAQS for CO, which is 35 ppm. (Alternatively, the 8-hour NAAQS of 9 ppm would also be of interest if one were investigating concentrations averaged over 8 hours.) As noted earlier, the highest CO level in the Houston 1 data set is 19 ppm. Thus, when 35 ppm is used as the threshold five of the six models, LAMB being the exception once again, yield $S_2 = 0$. This is not very enlightening because both the observed and predicted probabilities of exceeding 35 ppm are zero. Thus, the result is hardly surprising in view of the tendency of these five models to underpredict high concentrations.

To demonstrate the application of S_2, a CO threshold of 8 ppm was used in addition to 35 ppm. The observed frequency of exceedance of this threshold in the Houston 1 data set is 0.0362, and S_2 is the difference between this observed and the predicted frequency. Table 11 indicates that S_2 is negative for the first five models, which is in keeping with their tendency to underpredict the high levels of CO. The tendency of the LAMB model to overpredict is apparent in the large positive value of S_2, which is further evidence of the problems with the LAMB model in its present configuration.

CALINE3 and ROADMAP produced essentially equivalent results with respect to S_2, with GMG a close second. In fact, S_2 is not statistically significant for these three models, which implies that the value of S_2 is statistically indistinguishable from zero. This is reflected in the closeness of the respective values of F_2, which range from 9.1 to 9.4, for these three models. MROAD2 and HIWAY-II also produced basically equivalent results for S_2, but the performance is somewhat poorer than that of CALINE3, ROADMAP, and GMG, and indicates an enhanced tendency to underpredict that is consistent with the indications previously provided by S_1. Nevertheless, MROAD2 and HIWAY-II rate fairly high on the FOM scale, with F_2 being 8.4 and 8.2, respectively. The relatively high value of F_2 is a consequence of the fact that one is dealing with a very low observed probability, and any low predicted probability will yield a respectable F_2 rating. For example, if the predicted probability were equal to zero, $S_2 = -0.0362$, corresponding to $F_2 = 8.1$. Thus, only large absolute deviations from the observed probability will yield a low F_2 rating. This is the case with the LAMB model, which has S_2 close to 1, and, consequently, a low F_2.

Statistics S_3, S_4, and S_5 are correlation coefficients between observation and prediction, computed using the natural logarithm of the concentrations. They describe the amount of coherence between observed and predicted concentrations. Their values range from −1 to +1, with +1 denoting perfect coherence and zero corresponding to no coherence between observation and prediction. Negative correlations indicate an inverse relationship between observation and prediction (i.e., CO is predicted to decrease when it actually increases and vice-versa). Negative correlations are considered regardless of their value or statistical significance, to be indicative of the presence of serious modeling errors. Consequently, a positive FOM is obtained only for positive values of S_3, S_4, and S_5.

From Table 11 it is seen that S_3 is less than 0.5 in all cases, which indicates a substantial amount of scatter in the data. CALINE3 has the highest S_3 at 0.415, which is a mediocre performance in absolute terms. HIWAY-II and GMG are second and third, respectively, and—surprisingly perhaps—ROADMAP and LAMB have similar values of S_3. MROAD2 shows no significant correlation between observation and prediction. Judging from S_3, none of the models perform as well as one would like, as reflected in the low values of F_3 assigned to the models. This contrasts with the good performance of most of the models as measured by S_1 and S_2. The decline in performance is partly a consequence of the fact that S_3 treats the entire data set, whereas S_1 and S_2 are concerned only with high concentrations. It is also indicative of the change from macrostatistics to the more demanding microstatistics. Later, under the diagnostic analysis discussion, scatterplots of observation and prediction will be examined that will provide insight about S_3.

S_4 is the temporal correlation coefficient; it is computed by averaging the correlation coefficients for the individual monitors. Only CALINE3 and HIWAY-II have statistically significant values of S_4, and both coefficients indicate a relatively low level of linear agreement between observations and predictions. Clearly, CALINE3 is considerably better than the other models. The generally mediocre performance of the models in this category is indicated by the values of F_4, which range from 0 to 5.3. Because S_4 is the average of the correlation coefficients for the individual receptors, a low (or high) correlation at a single monitor can cause S_4 to be too low (or too high). Thus, the correlation coefficients of the various receptors should be examined to determine whether S_4 is being affected by erratic behavior at a particular monitor or subset of monitors. In the discussion of diagnostic analysis, a graphic display will be introduced that allows one to make such an assessment.

S_5 is the spatial correlation coefficient, which measures the model's ability to follow the spatial fluctuations of CO on an hourly basis. For each hour, the correlation coefficient is computed over all receptors, and S_5 is obtained by averaging all these correlations over all the hours. As with S_3 and S_4, the values of S_5 are relatively low, ranging from −0.191 to 0.515. The ROADMAP model has $S_5 = 0.515$, which is a considerable improvement over its performance as measured by S_4. HIWAY-II, GMG, and LAMB also have values of S_5 that show an improvement over S_4, which indicates that these four models do a better job of matching spatial rather than temporal patterns. CALINE3, by contrast, shows a decline in S_5 relative to S_4. MROAD2 has $S_5 = -0.191$, which shows a lack of ability to match spatial patterns. On the FOM scale, F_5 shows that the performance of the models ranged from 0 to 5.2, which continues the pattern of mediocre performance indicated by F_3 and F_4.

Statistic S_6 is the root-mean-square (rms) error between observed and predicted concentration. The theoretical range of S_6 is zero to infinity, with zero denoting a perfect match between all observed and predicted concentrations. Table 11 shows that five of the six models, LAMB excepted, have similar values of S_6 that cover a narrow range between 2.15 and 3.57 ppm. CALINE3 has the lowest S_6, ROADMAP is second, HIWAY-II and GMG are tied for third, and MROAD is fourth. The LAMB model has $S_6 = 143$ ppm, which clearly shows a serious problem of inaccuracy in the model.

Examined in isolation, the values of S_6 that range from 2.15

to 3.47 ppm appear small, and it is necessary to put them in perspective by comparing them with the observed concentrations. This is one function of F_6, which uses the ratio of S_6 to the mean observed concentration as the variable. The values of F_6 given in Table 11 indicate that the global accuracy of the predictions is relatively low, all the values of F_6 being less than 5. (A value of $F_6 = 5$ corresponds to a ratio of S_6 to mean observed concentration of 0.5.) Thus, what appeared to be low rms errors are shown by F_6 to be associated with a performance that is below average. As it should, the inflated S_6 of the LAMB model yields F_6 approximately equal to zero.

Table 12. Model evaluation statistics and figure or merit adjusted for measurement error[1].

Model Name	Statistic[1]						Figure of Merit (FOM)						Overall FOM[c]
	S_1	S_2[a]	S_3	S_4	S_5	S_6[b]	F_1	F_2	F_3	F_4	F_5	F_6	
HIWAY-II	3.03	-0.0303	0.391	0.274	0.528	3.07	3.3	8.5	3.9	2.7	5.3	2.3	4.1
CALINE3	1.21	-0.0103	0.505	0.652	0.468	1.91	8.3	9.6	5.1	6.5	4.7	4.4	6.3
ROADMAP	1.41	-0.0113	0.188	-0.0915	0.716	2.95	7.1	9.5	1.9	0	7.2	2.5	4.6
GMG	1.23	-0.0165	0.256	0.111	0.463	3.07	8.1	9.3	2.6	1.1	4.6	2.4	4.6
MROAD2	2.07	-0.0271	-0.00025	0.232	-0.234	3.32	4.8	8.7	0	2.3	0	2.1	3.7
LAMB	0.0253	0.947	0.177	-0.0741	0.592	143.0	0.3	0.6	1.8	0	5.9	0	1.4

[1] Error = ±1 ppm.

[a] CO threshold equal to 8 ppm.

[b] Units are ppm.

[c] FOM = $[(F_1 + F_2)/2 + (F_3 + F_4 + F_5)/3 + F_6]/3$.

The poor performance of the LAMB model was unexpected because the model had been previously evaluated (by both the developer of the model (*19*) and at SRI in the course of transferring it to a local computer) using the GM data base with excellent results (*19*). Following a discussion with Dr. Lamb, it was learned that the current version of the computer program contains several features that pertain only to the GM data base. Hence, the computer program in its present form is not applicable to all data bases.

Thus, the LAMB model's poor performance can be attributed to the fact that, in its current form, the model is data-base dependent. In view of the previous results obtained by Dr. Lamb (*19*), the authors of this report believe that the problem does not lie in the mathematical formulation of the model, which is perfectly general. Rather, the problem is in the way the computer program treats the emission input and the receptors. Thus, the results of the evaluation do not provide a fair test of the capabilities of Dr. Lamb's approach.

The evaluation of the LAMB model is instructive nevertheless because it demonstrates one of the pitfalls that highway planners face in selecting and applying a model to a particular problem. Usually, it is assumed that the model is independent of the data base that was originally used to "validate" the model. Experience gained on this project shows that such an assumption is not always true, and it is well-advised to test the model using a data base different from that used previously to develop the model.

Effect of Measurement Error

The statistics given in Table 11 were not adjusted for the presence of uncertainty in the CO measurements, which Bullin et al. (*11*) define as ±1 ppm. Table 12 gives the value of the six statistics and associated FOM after adjusting for measurement error. In general, the adjustment resulted in an improvement in the performance of the models. The greatest improvement occurred in S_5, which went from a positive range of 0.337 to 0.515 to an adjusted range of 0.463 to 0.716, the value of 0.716 corresponding to the ROADMAP model. The FOM reflects best the improvement in the performance, with more models approaching or exceeding the midpoint value of 5. Thus, as intended, adjusting for the presence of observational uncertainty results in enhanced measures of performance for the models.

Diagnostic Analysis

According to the six measures of model performance, S_1 to S_6, five of the six models tend to underpredict high concentrations, and one overpredicts by a wide margin. Moreover, as measured by S_3, S_4, and S_5 the coherence between observed and predicted concentrations ranges from nonexistent to fair, where fair is considered to be a correlation of approximately 0.5. Statistic S_6 revealed a general lack of agreement between observations and predictions, which is exemplified by relatively large rms errors. One now turns to some diagnostic analyses in an attempt to gain insight about the reasons for such performance. Diagnostic analyses take many forms and use a variety of tools. The approach depends heavily on the use of graphic displays, particularly the scatterplot of observations and predictions. No claim is made of exhaustiveness: the aim is simply to obtain an intuitive understanding of a few aspects of model behavior.

The modeling problem is to attempt to reproduce a time series of CO concentrations at various receptors; the example in this report has 12 receptors. The computer program that calculates S_1 through S_6 computes S_1, S_2, S_3, and S_6 for each receptor, and all six statistics for all the receptors combined (S_4 and S_5 are averages defined for all the receptors taken together). The values of the statistics in Tables 11 and 12 are for all the receptors combined. The multidimensional character of the problem suggests that one should examine the behavior of S_1, S_2, S_3, and S_6 at the individual receptors to see whether any receptor or cluster of receptors stands out from the others and thus may have some special property that affects the performance of the model.

Figure 14 displays S_1 for the different models and receptors. The receptors are numbered from 1 to 12, the number corresponding to the CO probes shown in Figure 11. The values of S_1 for each model are displayed on a horizontal line on which the receptor number is placed at the value of S_1 for that receptor. Because of space limitations, or in case of tie, the receptor numbers are sometimes displaced vertically from the horizontal line. The "⊙" symbol is located at the value of S_1 for all the receptors combined. The figure shows that all the receptors cluster tightly for CALINE3. Thus, CALINE3 does not show much spatial variability in S_1. By contrast, the other models show substantial variations in S_1 for the different receptors. HIWAY-II does worst for monitors 7, 8, 11, and 12. Referring to Figure 11, monitors 7 and 8 are the highest, and 11 and 12 are generally upwind of the roadway; it appears that HIWAY-II has difficulty in predicting CO under these conditions. ROADMAP shows numbers only for 10 receptors, because it is not formulated to include upwind receptors. Figure 14 shows that ROADMAP has two clusters of monitors: One cluster has receptors 1, 3, 5, and 9 and has the lowest values of S_1. (Figure 11 shows that these four receptors are located at the same height, 33 ft.) The other cluster for ROADMAP contains receptors 2, 4, 6, 7, 8, and 10, which include the lowest and highest monitors. Thus, again, a spatially related variability is seen in S_1. The GMG model shows three clusters, with receptors 11 and 12—generally upwind—included in the cluster with the largest S_1. MROAD2 shows extreme variability in S_1, and again monitors 11 and 12 (this time accompanied by receptor 5) show the largest S_1. The divergence of LAMB from the other models is apparent in the figure, as is the fact that there is considerable spatial variability in S_1.

Figure 14 shows that different models have disparate patterns of spatial variability in S_1. Several models seem to find it difficult to handle upwind monitors. CALINE3 has the smallest spatial variability in S_1, and does not have serious difficulties with the upwind monitors. Further investigation of the concentration distribution at each receptor and of wind patterns is called for to elucidate the reasons for the spatial fluctuations in S_1.

Figure 15 shows a graph for S_2 in which the same conventions used in Figure 12 have been followed. This time fairly tight clusters are seen for most of the receptors for all the models except LAMB, which diverges so much from the others that it requires its own scale. The most interesting feature of Figure 15 is the location of monitors 1 and 11; both show the largest negative value for HIWAY-II, CALINE3, GMG, and MROAD2. As with S_1, it will be necessary to investigate further the CO distribution at the various receptors to uncover the cause of this behavior.

Figure 16 shows that all the models exhibit considerable spatial inhomogeneity in S_3. CALINE3 is the only model with values of S_3 greater than 0.5; these occur at receptors 2, 3, 5, 6, 7, 8, and 12. HIWAY-II has one cluster near $S_3 = 0$ for receptor 1, 4, 11, and 12, which (interestingly) include one of the receptors closest to the roadway (receptor 1) and the two that are farthest (receptors 11 and 12). ROADMAP shows two well-separated clusters, each with five receptors. GMG and MROAD2 display the greatest variability in S_3; both models show the highest S_3 at receptor 2. All the models have difficulty matching the CO fluctuations at monitors 1, 4, 9, and 11. Plots of time series at these and the other monitors would be helpful in finding the reasons for the discrepancy. The diagnostic analysis program described in Appendix F can be used to generate such plots.

Figure 16 also provides insight about the value of S_4, the temporal correlation coefficient. The symbol "◇" shows the value of S_4 for each model, and the figure allows one to relate S_4 to its components because S_4 is an average of the coefficients of the various receptors. (S_4 is not the arithmetic mean; see App. B.) Thus, for HIWAY-II, the cluster of sites 1, 4, 11, 12 can be seen to be lowering the value of S_4. Similar influences related to the spatial variability of the individual coefficients can be observed for the other models.

Figure 17 shows a graph for S_6. The divergence of the LAMB model from the others is readily apparent. For the other five models, receptors 1 and 11 exhibit the largest values of S_6. Excluding 1 and 11, all the other receptors form a fairly tight cluster for each model, a cluster covering a range of about a factor of 2 in the value of S_6. In view of these and earlier results, it is clear that something special is happening at receptor 1 and 11 that would merit additional investigation if one were trying to select a model for a real application, or if one were developing a model.

Another useful type of display is the scatterplot of observation and prediction. Figure 18 shows two scatterplots for the HIWAY-II model: one shows the log-transformed (Fig. 18(a)) and the other the untransformed concentrations (Fig. 18(b)). Comparing these two graphs shows that the logarithmic transformation emphasizes the effect of the low concentrations (large negative values on the graph), whereas the linear plot emphasizes the impact of high concentrations. Thus, Figure 18(a) shows that the low value of S_3 for HIWAY-II ($S_3 = 0.324$, see Table 11) is heavily influenced by a number of very low predictions with logarithms ranging from -11.51 to -4.72, which correspond to CO levels of approximately 10^{-5} and 10^{-2} ppm. Few, if any, CO instruments will measure such low levels, but of course models can and do predict such low values. Because the measurements are accurate to ± 1 ppm, consideration should be given to rounding off the predictions to the nearest ppm or tenth of a ppm to avoid this problem.

The tendency of HIWAY-II to underestimate the observations is apparent in Figure 18(a), because although the log of the observations ranges from -1.61 to 2.95, that of the predictions goes from -11.51 to 2.08. The underprediction of high concentrations is more obvious in the linear plot of Figure 18(b), which shows that observations greater than about 10 ppm are underestimated by a wide margin. The

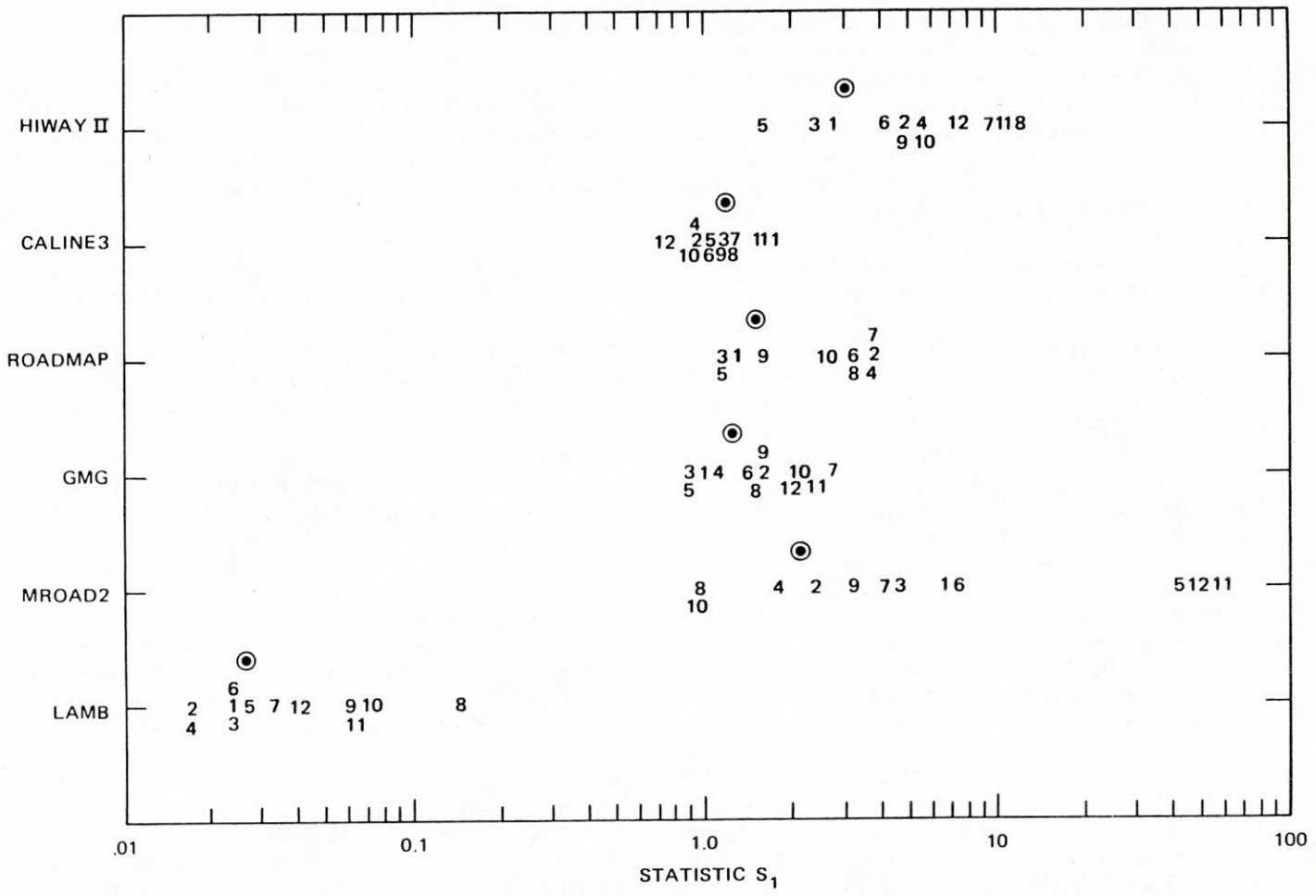


Figure 14. Graph of statistic S_1 for individual receptors and models.

correlation coefficient for Figure 18(b) is r = 0.353, which is slightly higher than that for Figure 18(a). Especially noteworthy is the point in the upper left corner of Figure 18(b), whose coordinates are (0.24, 19.2), because it is associated with the highest observed CO level, which was 19.2 ppm. This point has a very large residual that contributes about 4 percent to the sum of squared residuals that are used to calculate S_6; without this point S_6 would be equal to 3.16 ppm instead of 3.23, a 2 percent reduction. Thus, one of 802 points contributes about 2 percent to S_6. Consequently, this point merits additional study to determine the validity of the measurement and, if the data are valid, to see whether it provides any clues about the model's behavior in this concentration range. As will be shown, this point exerts a similar influence on S_6 for all the other models.

Figure 19 shows log-transform and linear scatterplots for CALINE3. Figure 19(a) exhibits the same feature seen in Figure 18(a), namely, a scatter of points on the left half of the figure that are associated with very low predictions. Apart from these, the main sequence of points shows somewhat less scatter than Figure 18(a), and this is reflected in the higher correlation coefficient. Figure 19(b) shows a more obvious linear relationship between observation and prediction, and the tendency to underpredict is also evident. The correlation coefficient for Figure 19(b) is r = 0.535, and is higher than that for Figure 19(a), which is r = 0.415. Figure 19(b) shows that the highest observed CO (19.2 ppm) is still underestimated, but not as much as in Figure 18(b). The coordinates of this point in Figure 19(b) are (4.26, 19.2), and its contribution to S_6 is about 3 percent.

Scatterplots for the remaining four models are shown in Figures 20 through 23; each figure includes log-transformed and linear plots of observation and prediction. The plots showing the log-transformed CO display a swarm of points with little linearity, and thus low correlation coefficients. However, the correlation is statistically significant only for GMG. The linear scatterplots (part (b) of Figures 20 through 22) show that the correlation is higher than for the log-transformed CO, but the improvement is small. The linear plots also show the highest observed CO in its familiar position in the upper left corner of the graph. Its contribution to the value of S_6 for these other models is similar to those discussed previously. Figure 23(b) allows one to appreciate the magnitude of the discrepancy between observation and

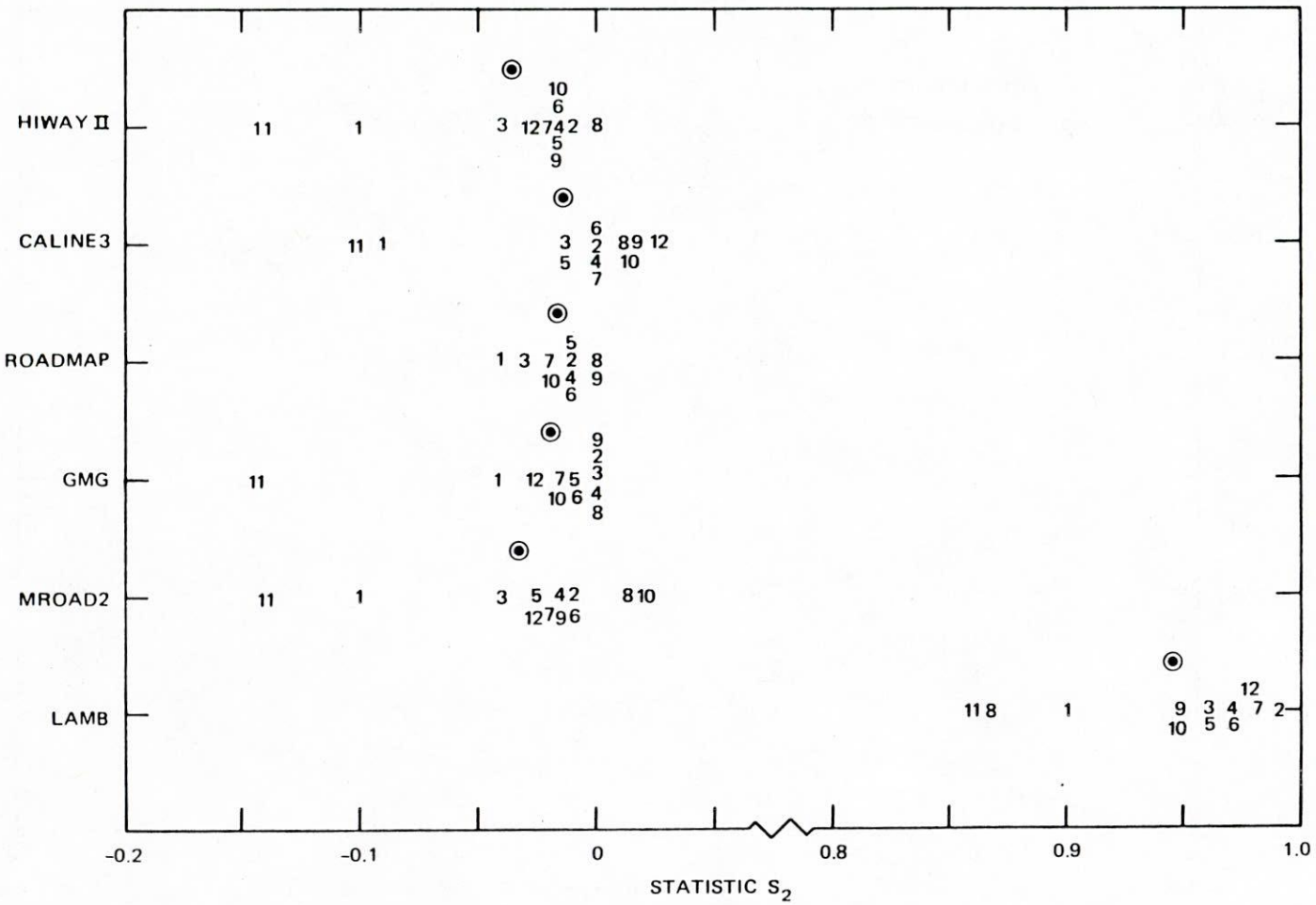


Figure 15. Graph of statistic S_2 for individual receptors and models.

prediction for the LAMB model—the predictions range as high as 681 ppm, whereas the observations only reach 19.2 ppm.

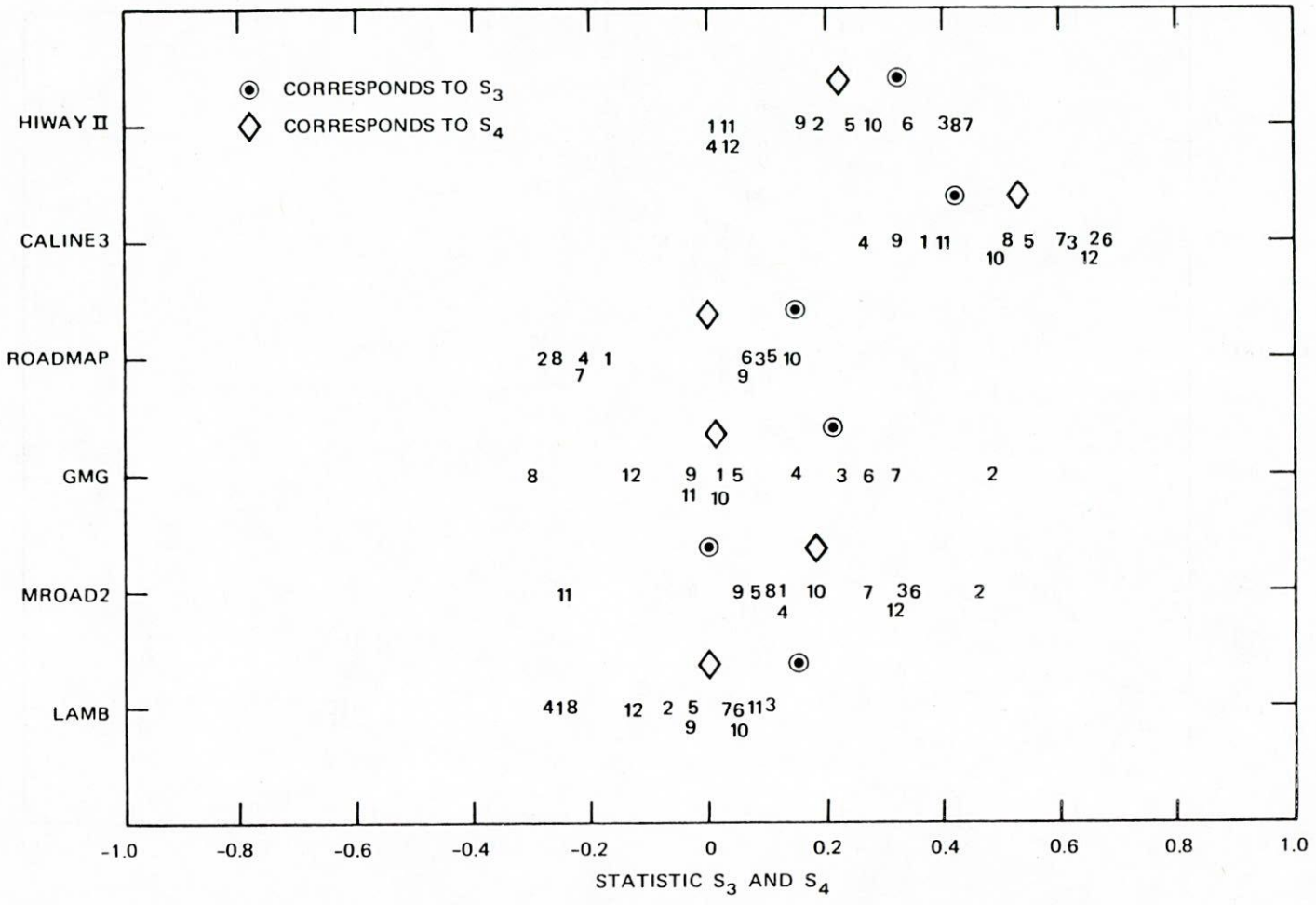
The diagnostic displays previously discussed are a small but important component of a complete diagnostic analysis. The plots have raised a number of questions that merit further examination. The diagnostic analysis package developed in this project (see App. F) provides various tools that can be used to obtain the desired answers. For example, it would be fruitful to compare the entire distribution of observations and predictions, and this can be done using the quantile-quantile plot option of the analysis package. In addition, to examine the behavior of the various statistics, it would be helpful to regress the residuals as a function of wind speed or direction, which can be done using the multiple regression procedure in the package. The possibilities are many, and the diagnostic tools provided are flexible enough to accommodate most of them.

Confidence Intervals and Sample Size

The computer program developed in this project for calculating S_1 through S_6 also computes a 95 percent confidence interval for each statistic. Appendix D of this report contains the mathematical derivation of the confidence intervals, and Appendix E describes the computer program. This section discusses the interpretation of the confidence intervals and their use in guiding the estimation of the size of the working data base.

The value of S_i computed by the program is an estimate of the unknown true value of the statistic. The value of S_i is called a point estimate because it is a single number. The confidence interval describes the precision of the point estimate (i.e., how close the point estimate may be to its true value). The precision is low if the interval is wide and high if the interval is narrow. In general, a small sample size leads to a wide interval and low precision, and a large sample size to narrow interval and high precision.

Specifically, a 95 percent confidence interval contains the true value 95 percent of the time. It is inaccurate to state that there is a 95 percent probability that the true value is inside some particular interval because the true value either is or is not inside the interval. The correct interpretation is that the 95 percent confidence is related to the expectation that if many intervals were calculated, on the average 95 percent of them would contain the true value.



Numbers correspond to the individual receptors shown in Figure 11; ◉ and ◇ denote the values of S_3 and S_4 for all receptors combined.

Figure 16. Graph of statistic S_3 and S_4 for individual receptors and models.

Figure 24 displays the 95 percent confidence intervals for S_1 for all models, as well as the point estimates of S_1, which are represented by the inverted triangle. The figure indicates that CALINE3, ROADMAP, and GMG have overlapping confidence intervals that include a value of one. Hence, their respective performances relative to S_1 are statistically indistinguishable at the 95 percent confidence level. The differences in precision are apparent from the width of the intervals, although the visual impression is somewhat distorted by the logarithmic scale. HIWAY-II has the widest interval and the interval width decreases as follows: ROADMAP, GMG, MROAD2, CALINE3, and LAMB. The intervals for HIWAY-II, MROAD2, and LAMB are significantly different from a unit value, as evidenced by the fact that their respective confidence intervals do not overlap one.

The confidence intervals for S_2 are displayed in Figure 25, in which the LAMB model requires its own scale in order to fit into the figure. Except for LAMB, all the other models show considerable overlap in their confidence intervals, indicating substantial similarity in their behavior with respect to S_2. CALINE3, ROADMAP, and GMG include zero in their respective confidence intervals, which shows that their respective point estimates of S_2 do not differ significantly from zero, a very good result.

Figure 26 shows the confidence intervals for S_3. The wide intervals for five of the models indicate a low precision for the point estimate. The intervals for ROADMAP, MROAD2, and LAMB straddle zero, indicating that S_3 is not statistically significant for these models. CALINE3 exhibits the narrowest interval and, thus, the highest precision. The intervals for HIWAY-II and CALINE3 overlap, suggesting that for these two models, S_3 does not differ significantly at the 95 percent confidence level.

For S_4, the intervals for all the models except CALINE3 are so wide, as can be seen in Figure 27, that the point estimates are not very credible. In fact, the intervals for ROADMAP, GMG, MROAD2, and LAMB all contain zero; hence, their point estimates of S_3 are not statistically significant at the 95 percent level. CALINE3, by contrast, has a relatively high precision. The general lack of precision for

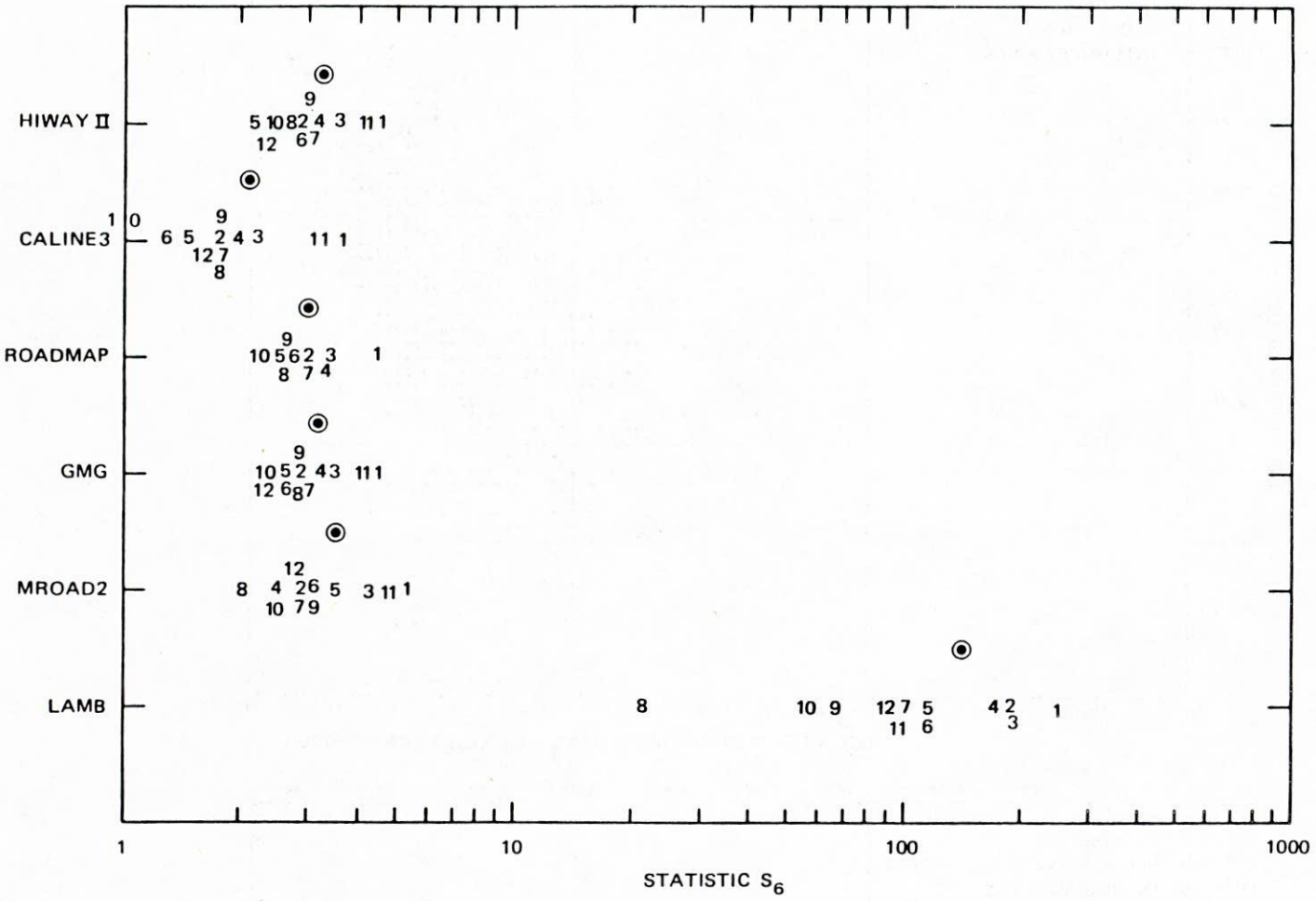


Figure 17. Graph of statistic S_6 for individual receptors and models.

five of the models is a consequence of the spatial inhomogeneity of the correlation coefficients for the individual receptors, which was depicted in Figure 16.

Figure 28 shows the confidence intervals for S_5. In contrast to S_4, S_5 generally exhibits relatively good precision in the form of moderately narrow intervals. Only MROAD2 has a wide confidence interval. The confidence interval for LAMB is contained in that for HIWAY-II, implying statistically similar behavior with respect to S_5. Indeed, the figure shows considerable overlap of the intervals for all the models except MROAD2. Because none of the intervals contains zero, all the S_5 coefficients are statistically significant at the 95 percent confidence level.

Statistic S_6 exhibits the highest precision of the six statistics, as can be seen in Figure 29. The figure shows that the confidence interval for CALINE3 is the narrowest and that it does not overlap any of the others. Hence, as measured by S_6, CALINE3's performance is significantly better than that of the other models. The graph also shows that there is no statistically significant difference in the values of S_6 for HIWAY-II, ROADMAP, GMG, and MROAD2. S_6 for the LAMB model looks like an outlier, which it is.

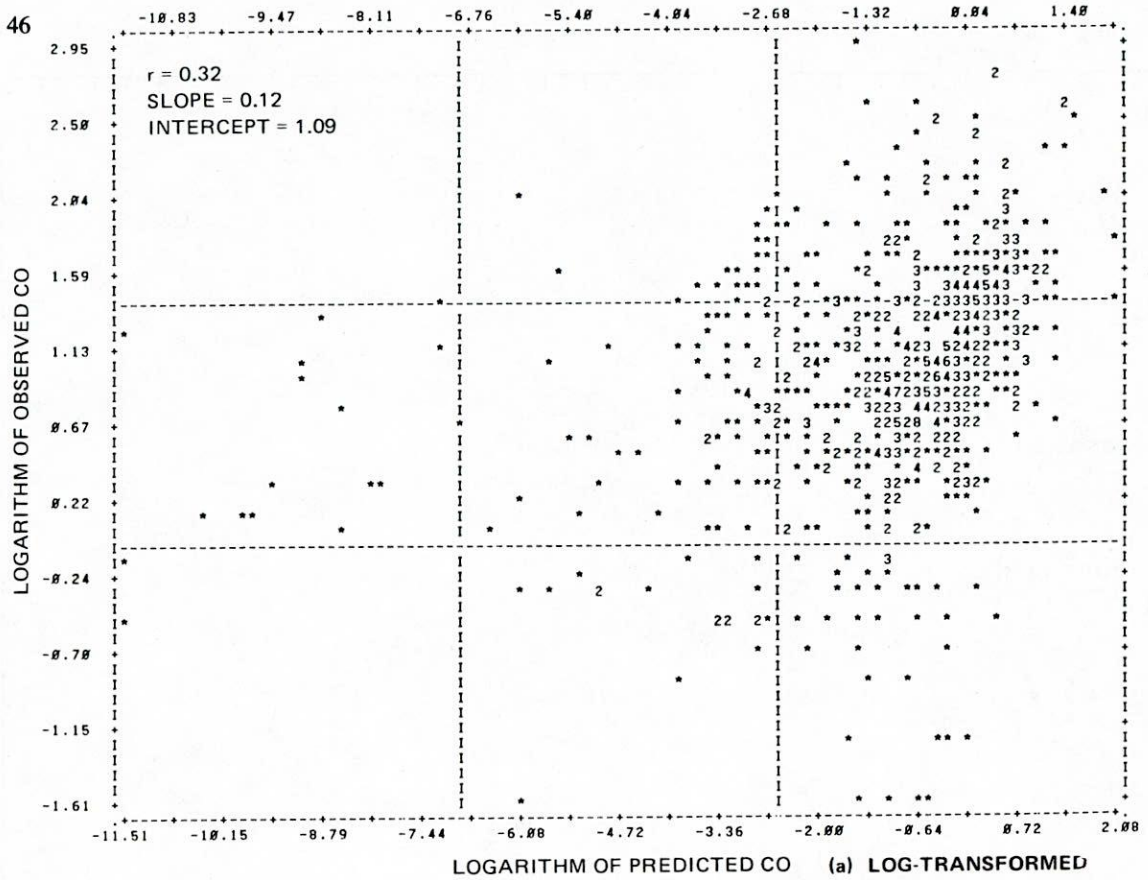
In general, Figures 24 through 29 indicate that the precision of S_1, S_2, S_5, and S_6 is acceptable, but that of S_3 and S_4 is not. One way to attempt to increase the precision is by increasing the size of the working data base. To estimate the new sample size, the following formula is recommended:

$$N_1 = N_o \frac{1}{6} \sum_{i=1}^{6} (W_i/w_i)^2 \qquad (15)$$

where

N_1 = new sample size;
N_o = present sample size;
W_1 = current width of the 95 percent confidence interval for the ith statistic, S_i; and
w_i = desired width of the 95 percent confidence interval for S_i.

In the example, $N_o = 802$ points. The precision of S_1, S_2, S_5, and S_6 seems acceptable, hence $w_i = W_i$ for $i = 1, 2, 5$, and 6. For illustration purposes, it is assumed that we want



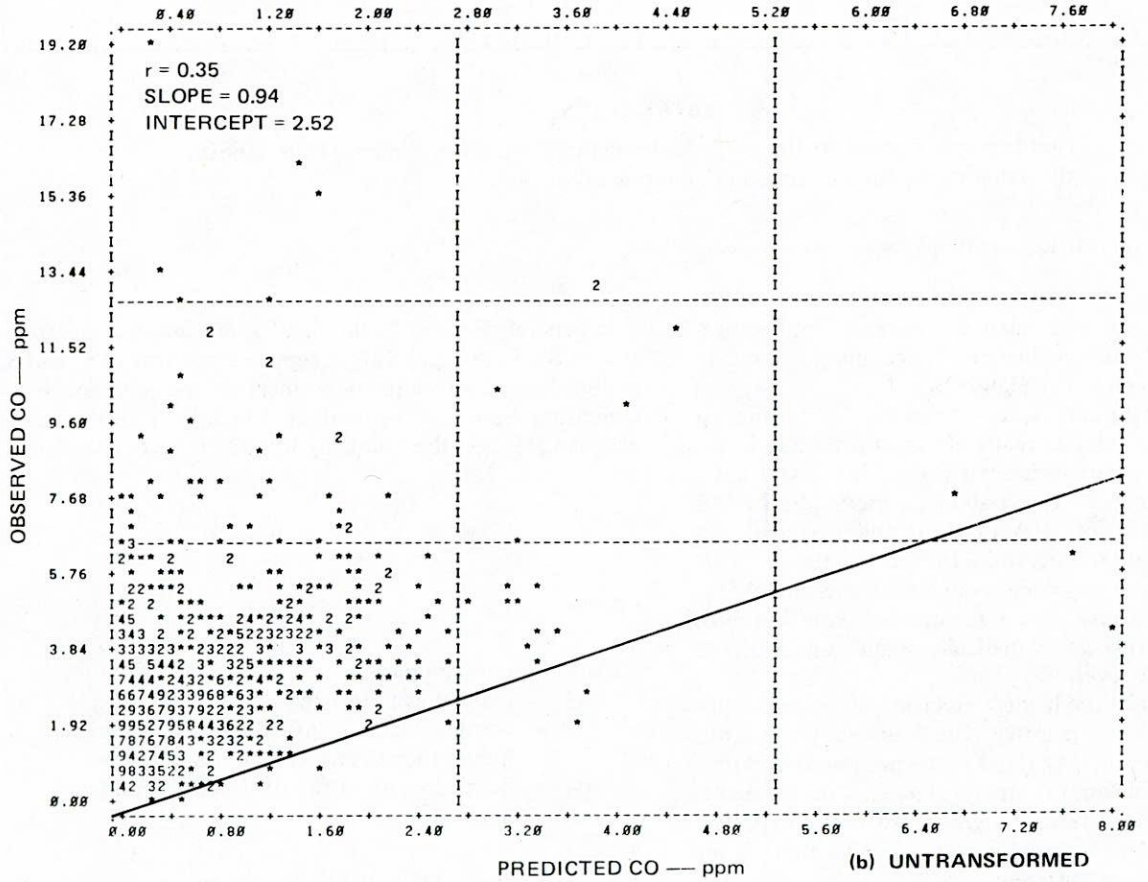
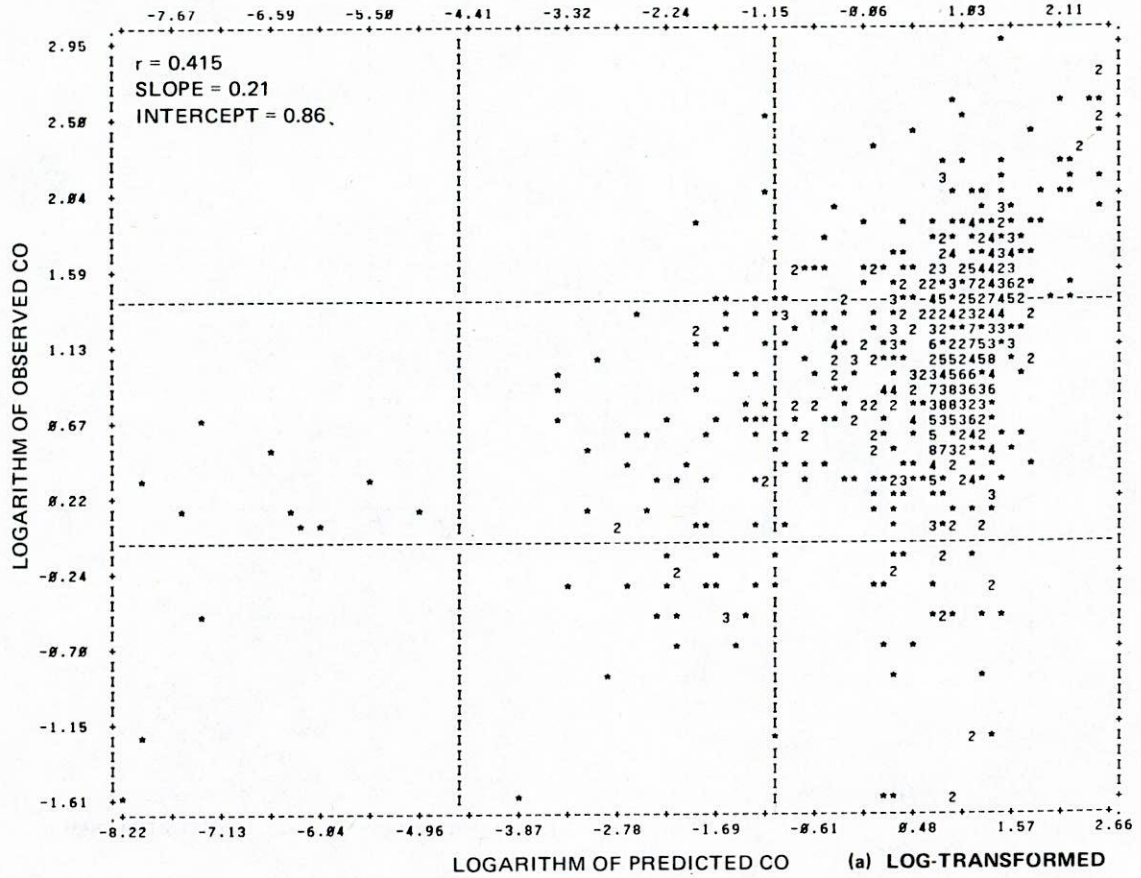


Figure 18. Scatterplot of observed and predicted CO for HIWAY-II (plotted numbers indicate number of points with same coordinates; plotted line represents line of equality).



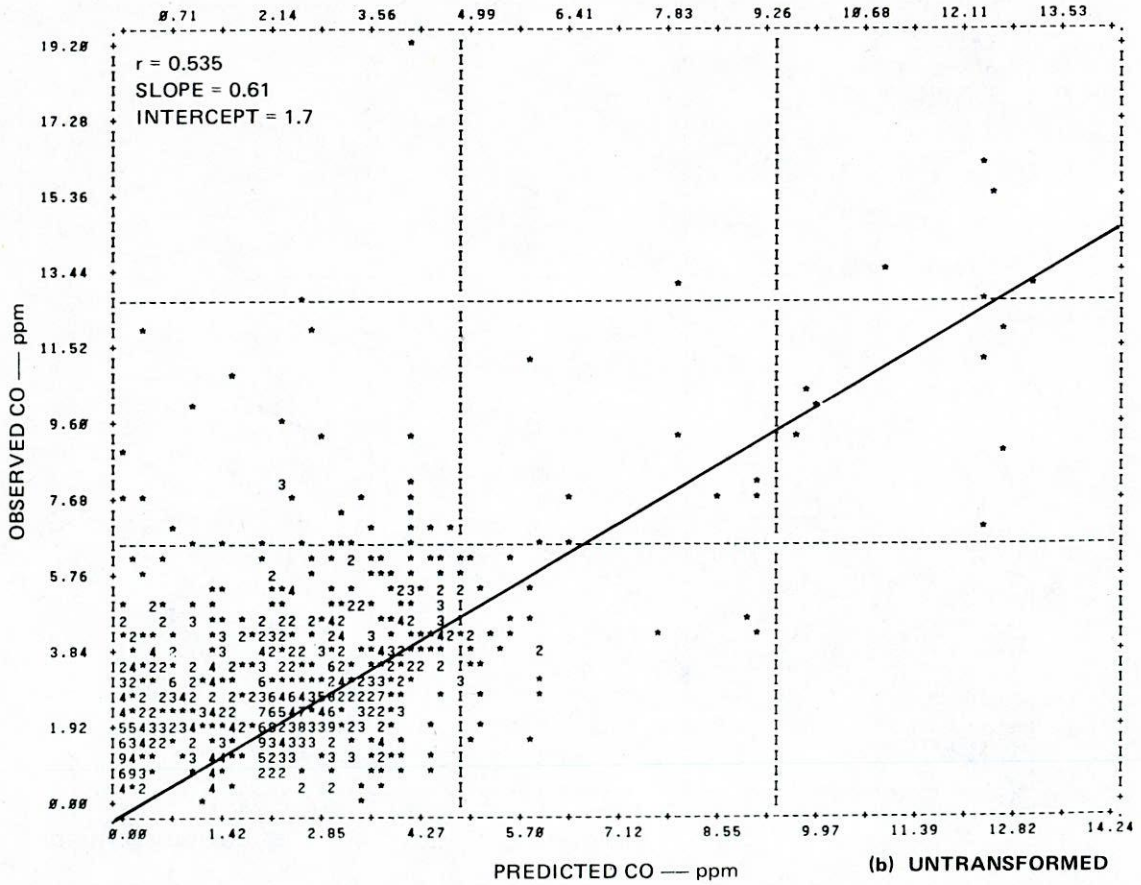
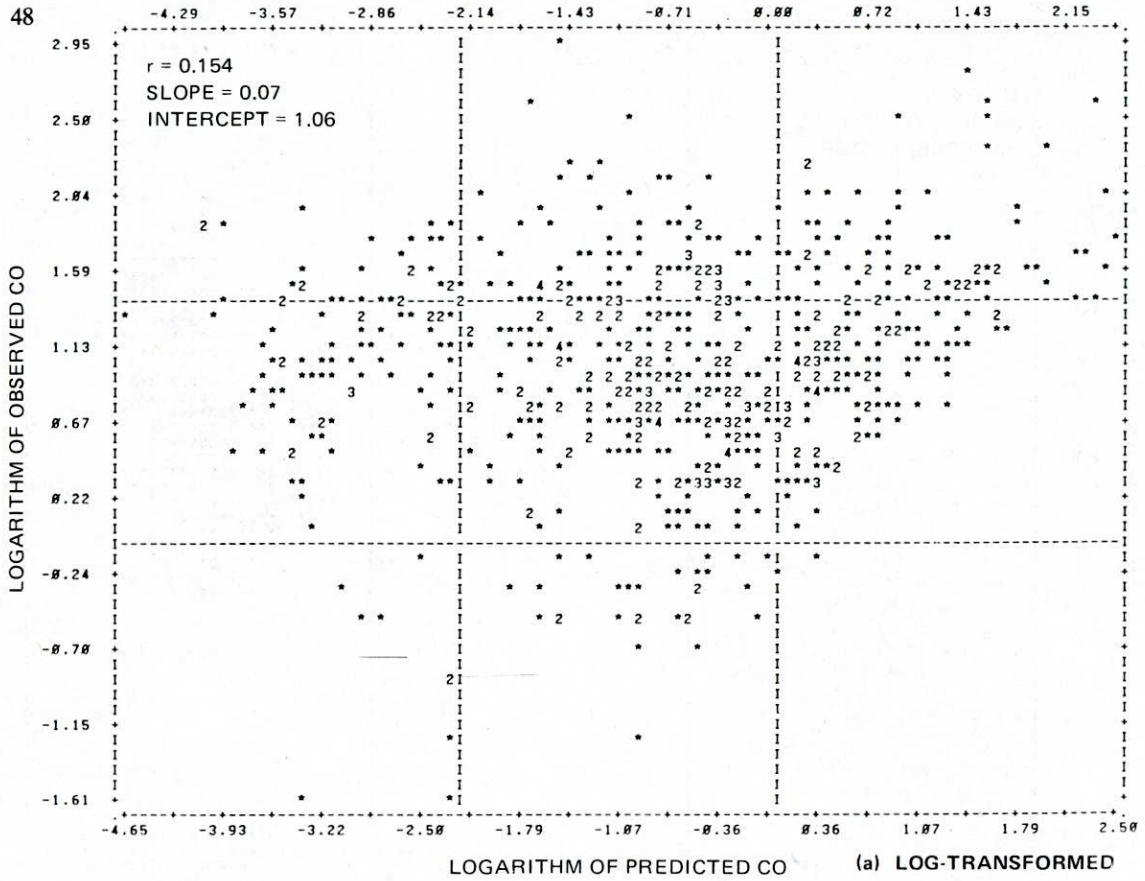


Figure 19. Scatterplot of observed and predicted CO for CALINE3 (plotted numbers indicate number of points with same coordinates; plotted line represents line of equality).



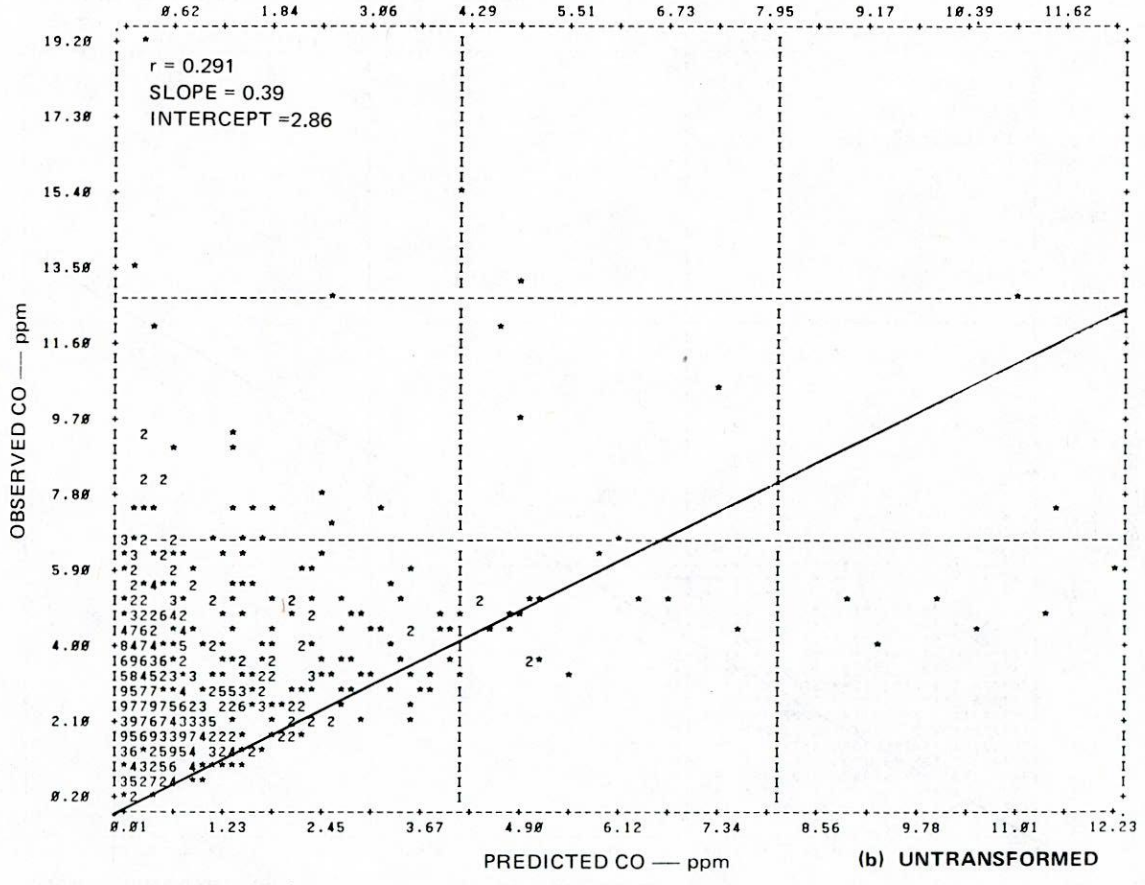


Figure 20. Scatterplot of observed and predicted CO for ROADMAP (plotted numbers indicate number of points with same coordinates; plotted line represents line of equality).

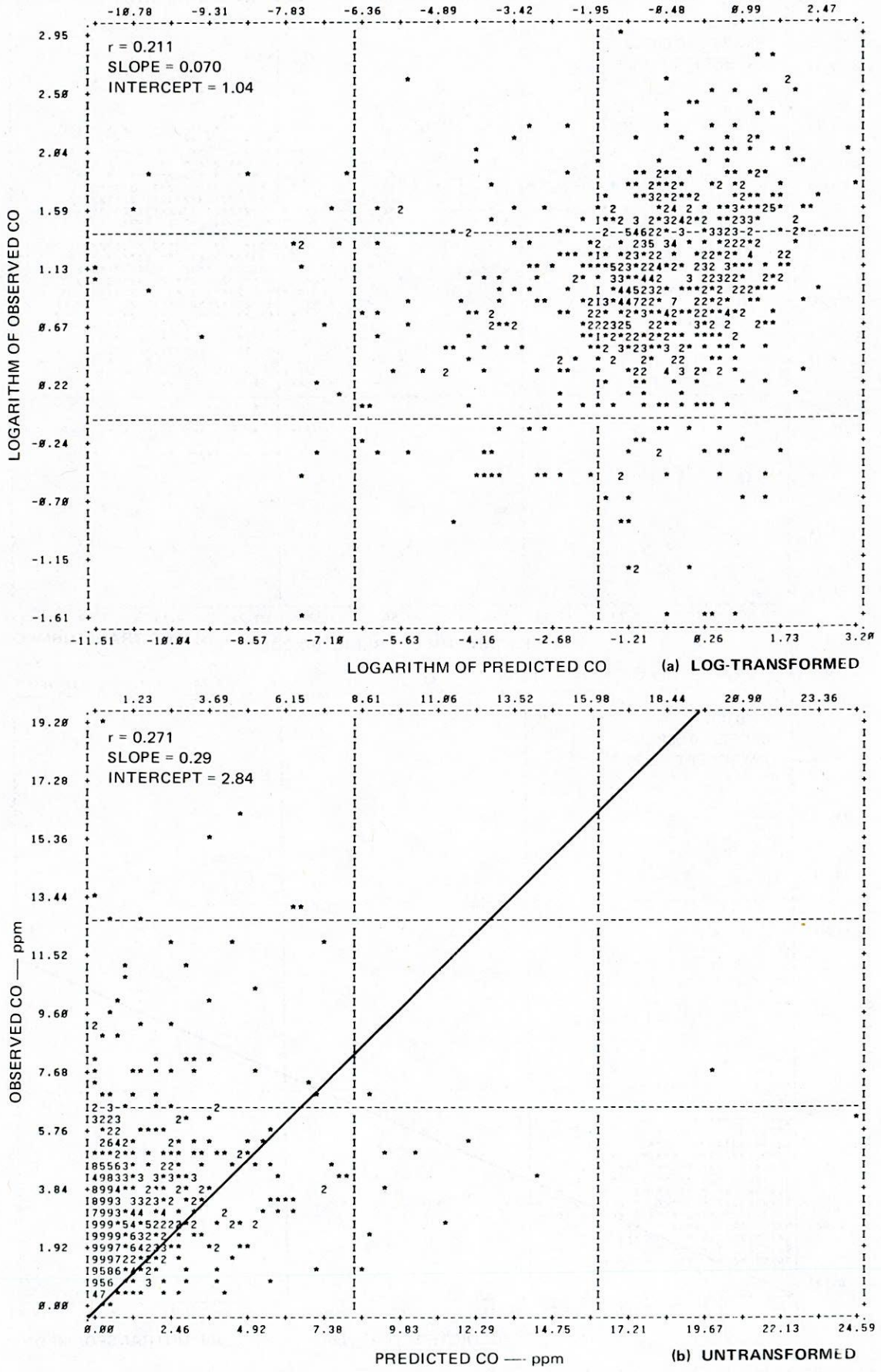


Figure 21. Scatterplot of observed and predicted CO for GMG (plotted numbers indicate number of points with same coordinates; plotted line represents line of equality).

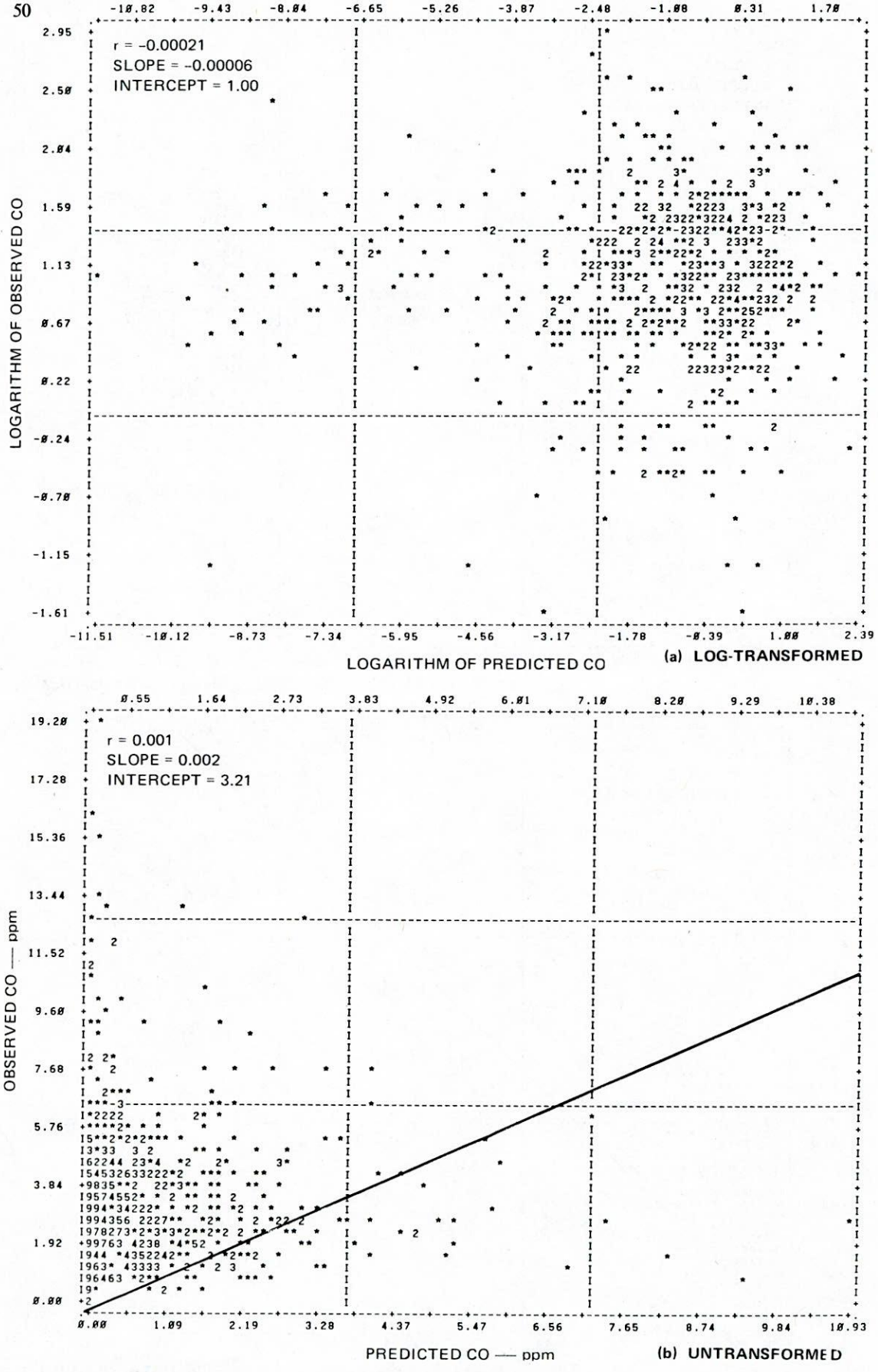


Figure 22. Scatterplot of observed and predicted CO for MROAD2 (plotted numbers indicate number of points with same coordinates; plotted line represents line of equality).

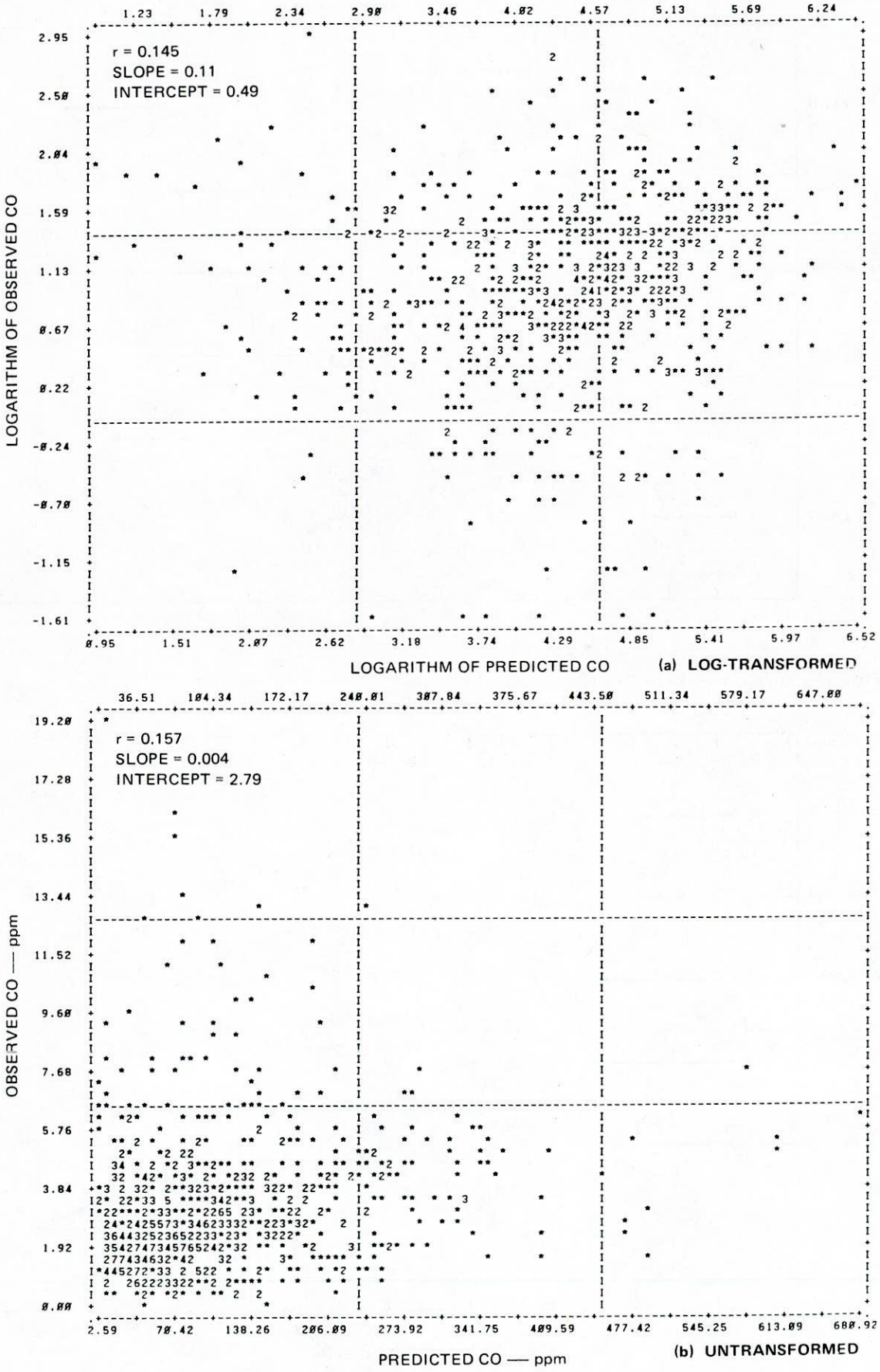


Figure 23. Scatterplot of observed and predicted CO for LAMB (plotted numbers indicate number of points with same coordinates).

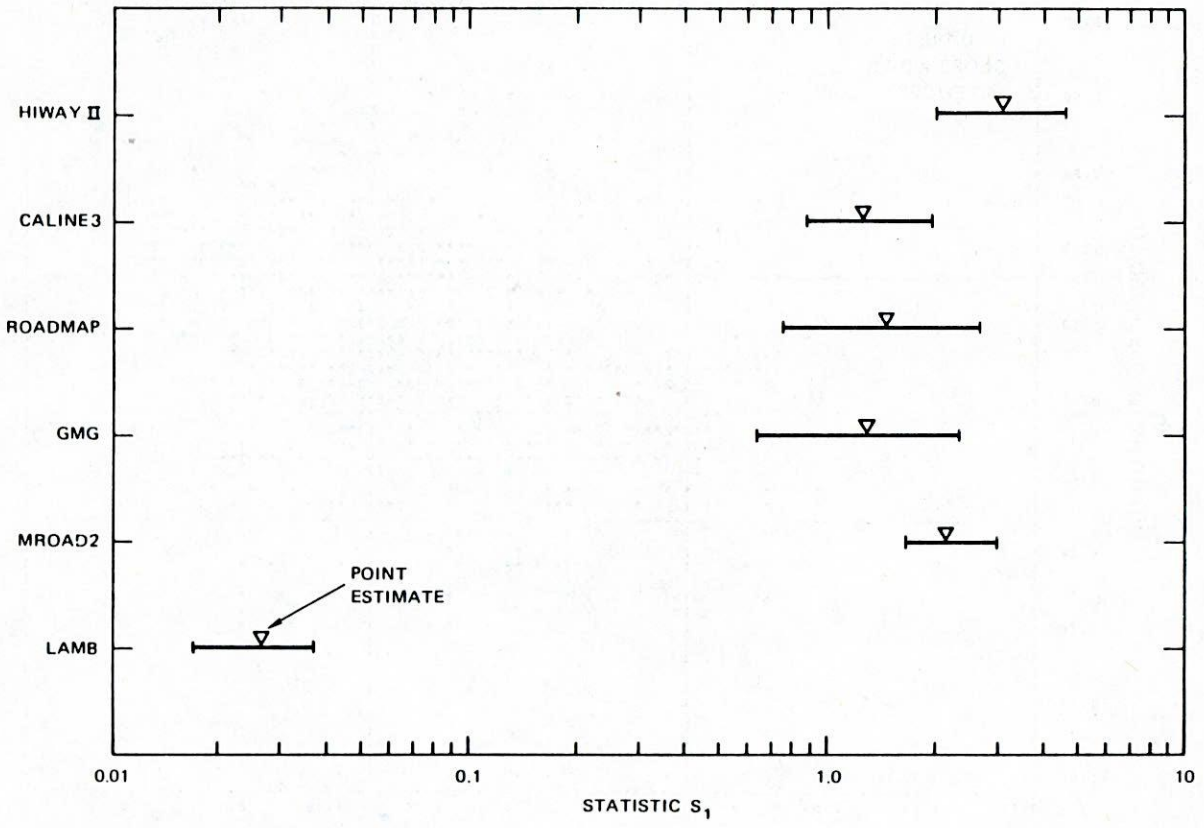


Figure 24. Statistic S_1—95 Percent confidence intervals.

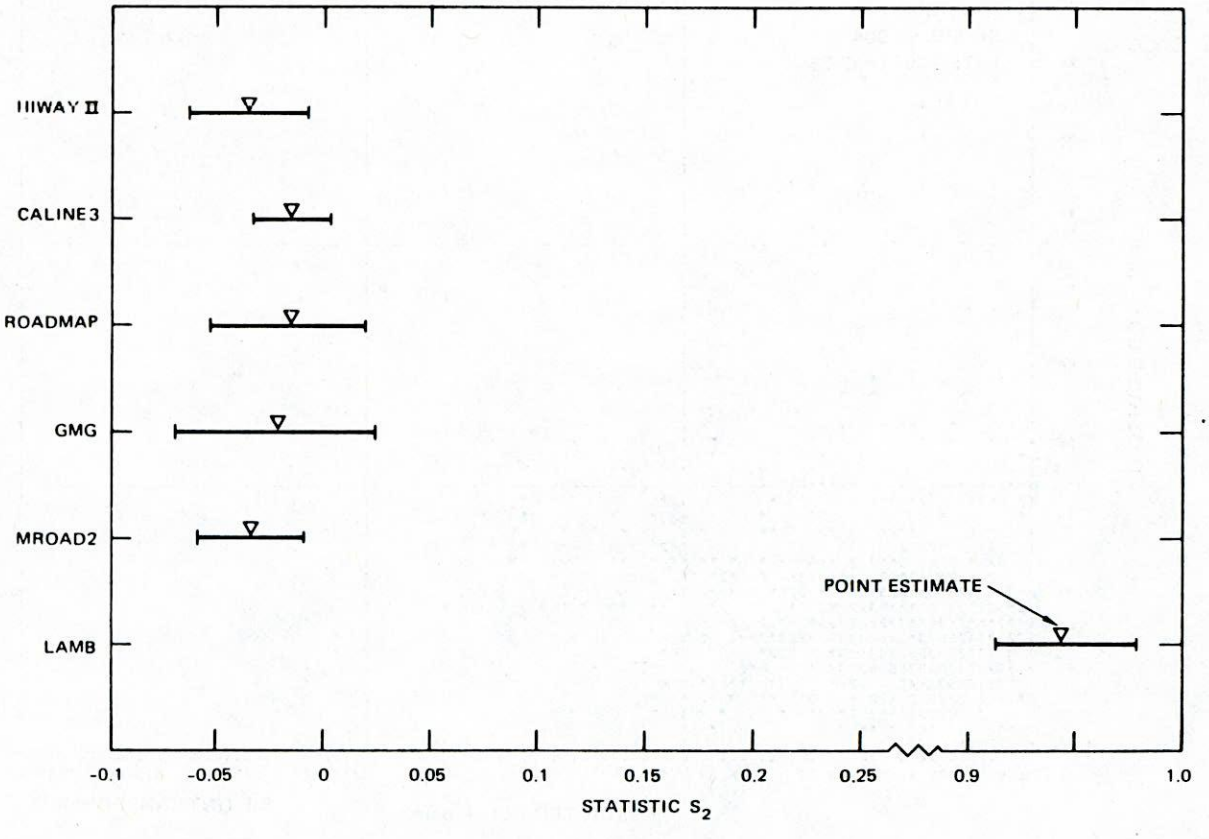


Figure 25. Statistic S_2—95 Percent confidence intervals.

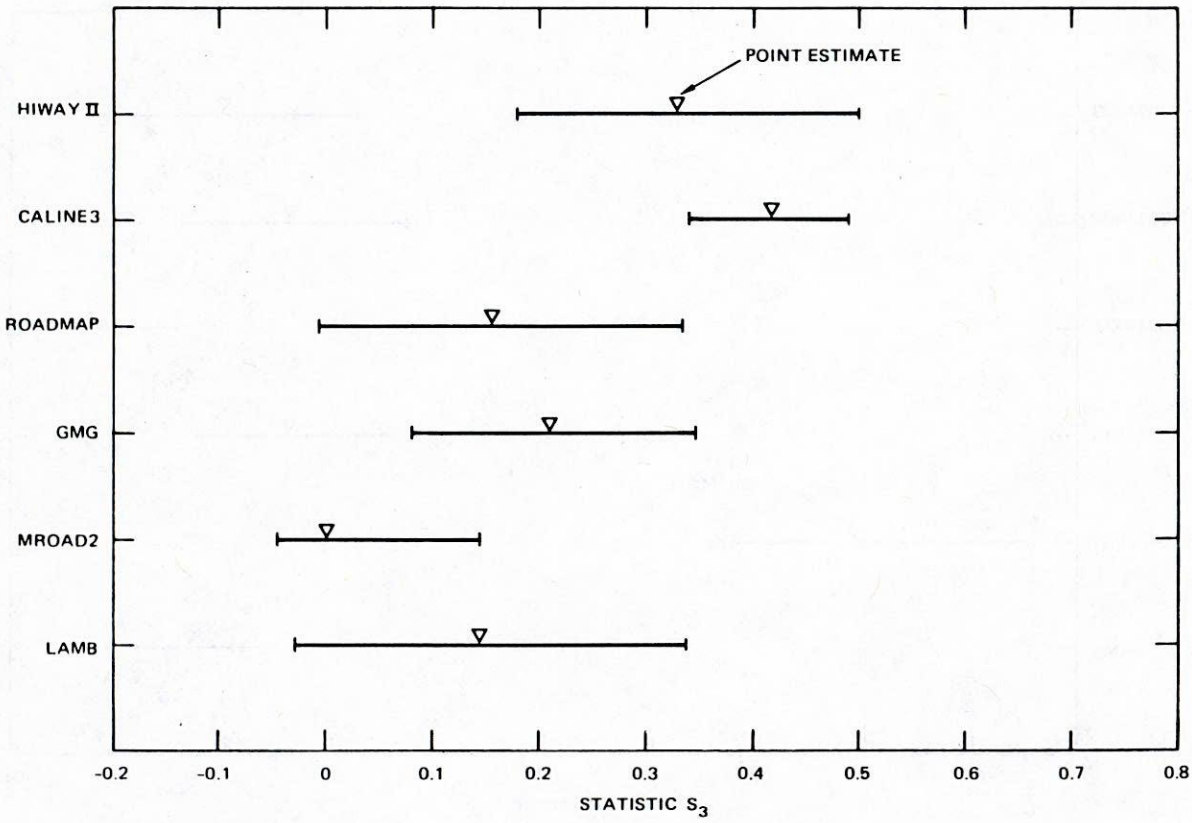


Figure 26. Statistic S_3—95 Percent confidence intervals.

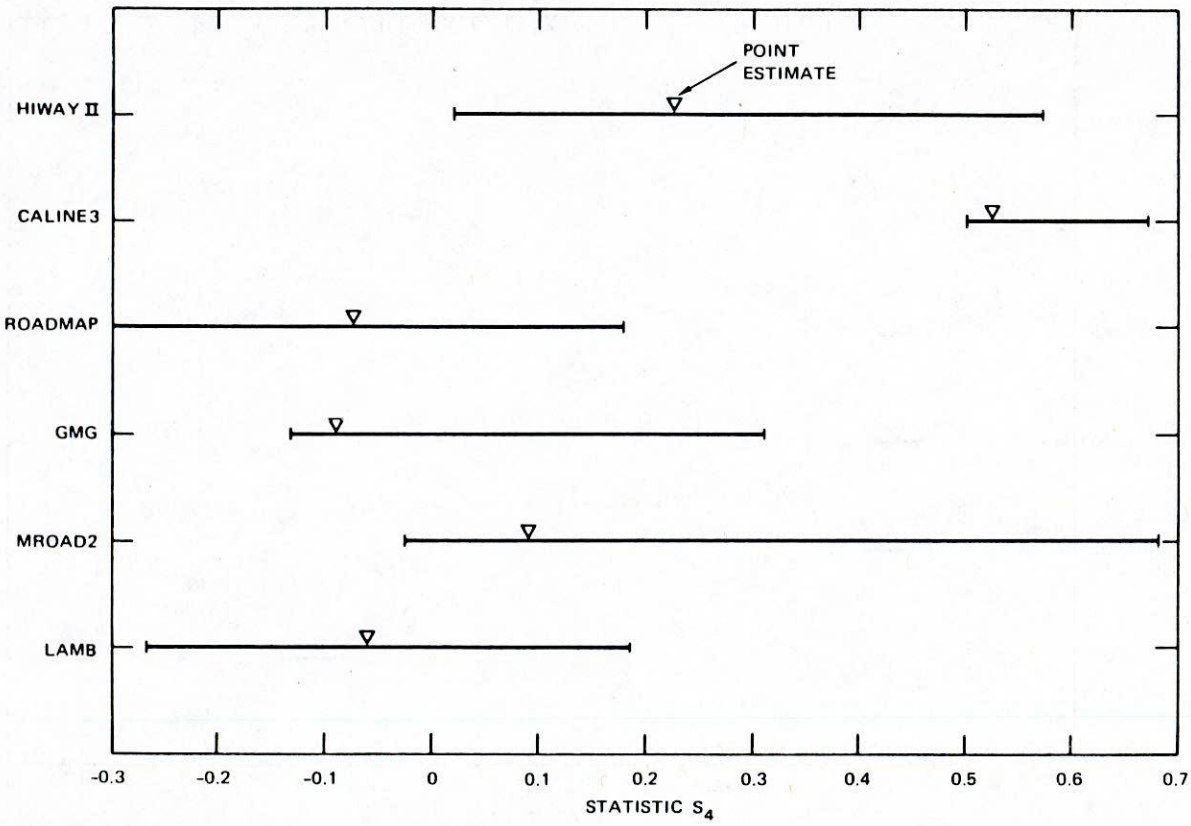


Figure 27. Statistic S_4—95 Percent confidence intervals.

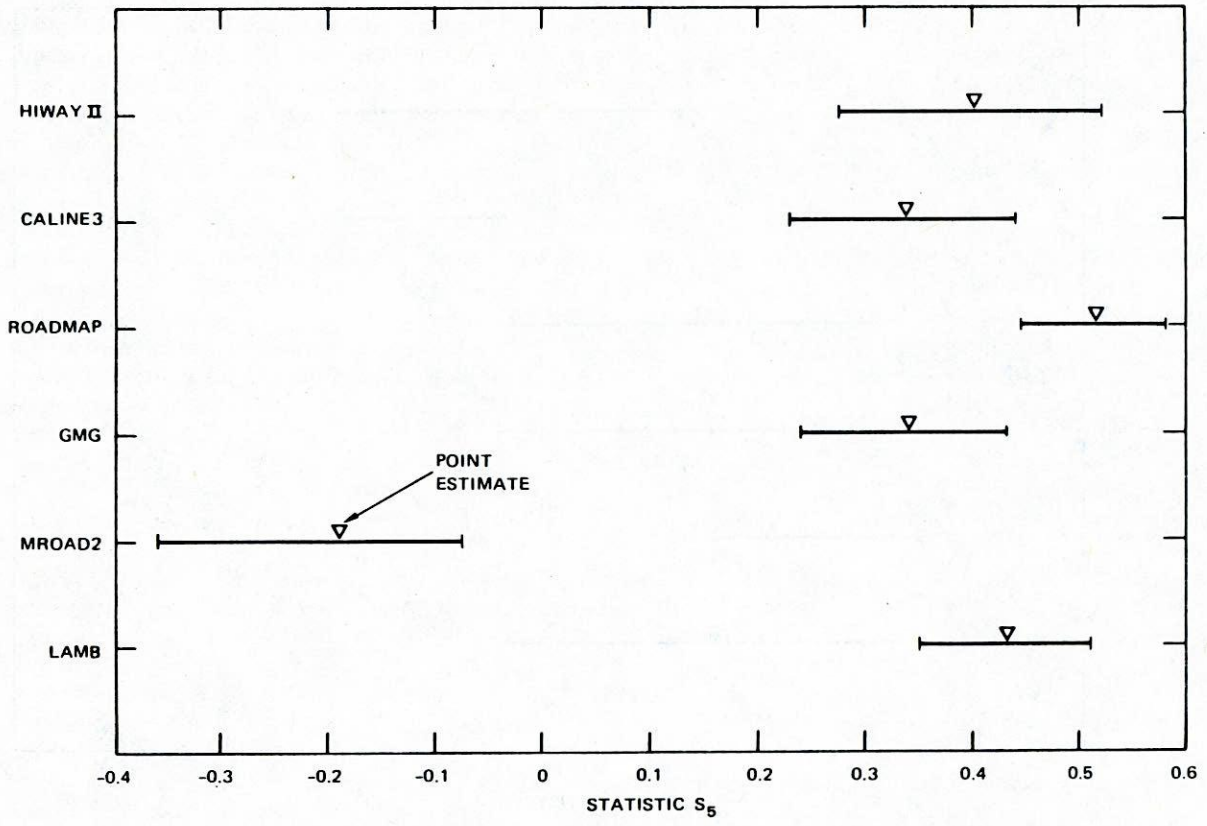


Figure 28. Statistic S_5—95 Percent confidence intervals.

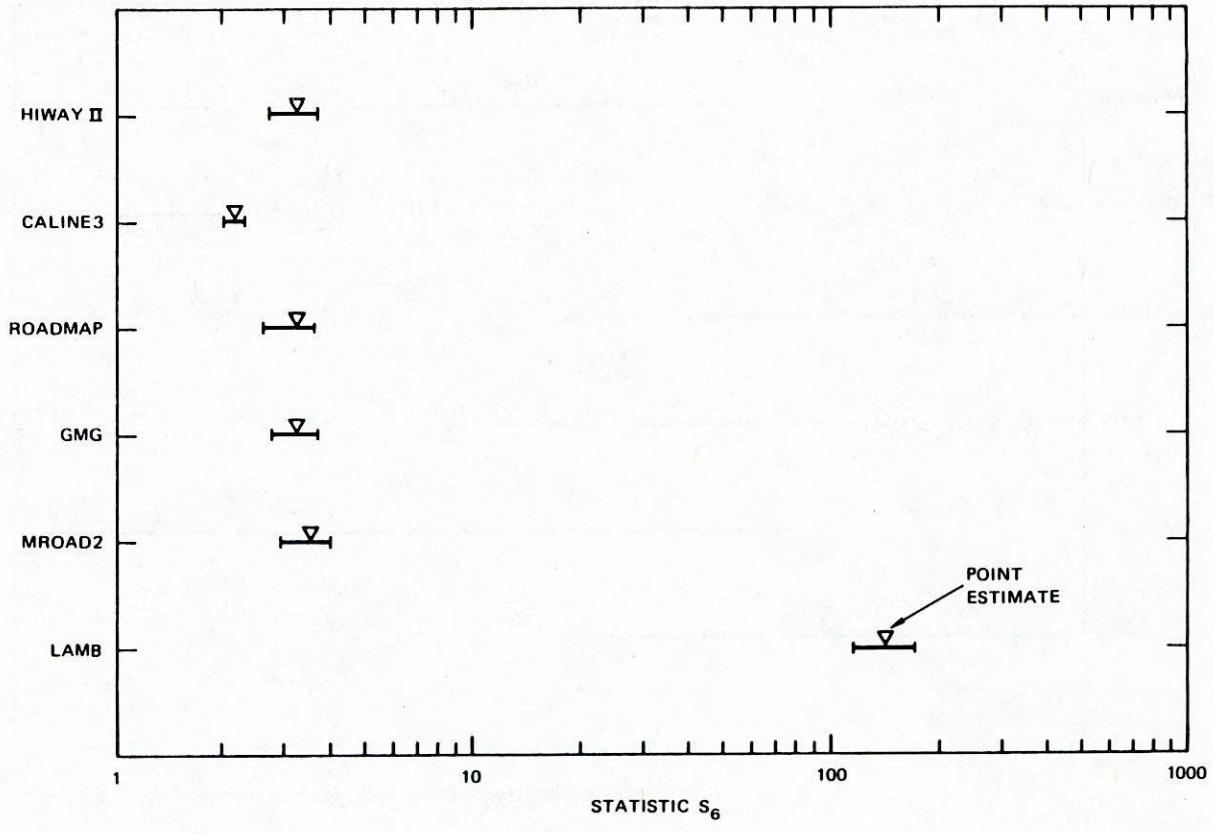


Figure 29. Statistic S_6—95 Percent confidence intervals.

to cut in half the confidence intervals for S_3 and S_4. Then $w_i = \frac{1}{2} W_i$ for i = 3, 4. Substituting in Eq. 15 yields

$$N_1 = 802 \times 2 \doteq 1604$$

Hence the original sample size must be doubled to improve the precision of S_3 and S_4 by a factor of 2. However, there is no guarantee that the desired result will be achieved. Every time a new sample size is defined the dispersion models must be rerun with the new data, the performance statistics and their confidence intervals recomputed, and the precision of each S_i checked to see whether the precision is satisfactory. If it is not, a new sample size is recalculated and the cycle is repeated. Thus, achieving a desired precision is an iterative process that continues until one of the following occurs:

- The desired precision is achieved.
- One runs out of data without achieving the desired precision.
- It is decided that it is too costly to continue, and whatever precision has been attained is accepted.

The example has shown that the sample size can escalate rapidly in the face of stringent precision requirements. Indeed, simply increasing the sample size to improve precision is a brute-force approach to the problem. Instead, before attempting to increase the sample size, it is recommended that a careful investigation be conducted of the reasons for the low precision in the statistics of interest. Such an investigation may well reveal some characteristic of the data base or of the model or of both that could be advantageously used to achieve the desired precision. Alternatively, some basic limitation could be uncovered that could prevent altogether any improvement in precision regardless of sample size. There is no "sure-fire" solution to the problem of improving precision, and the investigator must tread carefully to avoid unpleasant collisions with time and financial constraints.

RANKING MODELS

Approaches to Ranking

Once the performance statistics and associated FOMs have been computed, the next step is to select the model best suited for one's purpose. Such selection can take many forms, all being guided by the intended application and all involving some form of model ranking. The following discussion demonstrates the use of the two alternative ranking methods that were described in Chapter Two—the weighted-average liability (WAL) method and the minimum liability (ML) method. In addition, the WAL rankings are compared with those obtained by what is termed the sum-of-ranks (SOR) method. Although these ranking methods are useful for many applications, no claim is made that they will fit all the problems faced by a highway planner. Ultimately, the model ranking and selection will depend on the investigator's judgment. The methods described below are offered to assist the investigator to make choices; they are not meant to replace the investigator's judgment.

Application of Ranking Methods

The overall FOM shown in Tables 11 and 12 was computed using the WAL method. Given the overall FOM, the ranking of the models consists simply in ordering the models according to the value of the FOM. Thus, the overall FOM in Table 11 indicated the following ranking: CALINE3, GMG, ROADMAP, HIWAY-II, MROAD2, and LAMB. Table 12 yielded essentially the same ranking, but with GMG and ROADMAP tied for second.

The ML method assigns the smallest F_i as the overall FOM for the model. Based on Table 11, the ML method produces the following ranking: CALINE3 (FOM = 3.4), HIWAY-II (FOM = 2.1), GMG (FOM = 0.9), and ROADMAP, MROAD2, and LAMB tied for last place (FOM = 0). Clearly, a method that lumps ROADMAP with LAMB is insensitive, and for this reason was discarded. Nevertheless, it is interesting to note that CALINE3 emerged in first place using both WAL and ML methods, which is a consequence of its steady performance in all six categories.

It is useful to compare the WAL rankings against rankings derived from the SOR method, because the latter is insensitive to monotonic increasing changes in F_i. The SOR method works as follows. For each F_i, the SOR assigns a numeric value from 1 to 6 to each model, with 1 denoting the best performance and 6, the worst. For example, referring to Table 11, for F_1, CALINE3 would be assigned a 1, GMG a 2, ROADMAP a 3, MROAD2 a 4, HIWAY-II a 5, and LAMB a 6. This process is continued for all the F_i, and then the numeric values for each model are combined. The model with the lowest total is ranked first, that with the second lowest total is ranked second, and so on. Two ways of combining the ranks were used. One simply sums the six ranks, and the other uses the weighted sum $(R_1 + R_2)/2 + (R_3 + R_4 + R_5)/3 + R_6$, where R_i denotes the rank associated with F_i. The latter scheme is similar to that used to compute the overall FOM.

Table 13 gives the table of ranks that corresponds to the F_i in Table 11. Those cases in Table 13 that have been assigned the fraction ½ were tied with the same value of F_i. Both the simple and the weighted sum of ranks produce the following overall ranking: CALINE3, ROADMAP, GMG, HIWAY-II, MROAD2, and LAMB. This ranking is almost the same as that produced by the WAL method. The difference is that ROADMAP and GMG have exchanged places. This is not surprising because the overall performances of these two models are almost identical. For example, Table 12 shows them with the same overall FOM. In fact, applying the SOR method to Table 12 yields the same ranking as that shown in Table 13, but with slightly different sums of ranks. Thus, the SOR and WAL methods produce consistent results.

Because the SOR method is easily applied, it is recommended that it always be used to check the rankings produced by the WAL method. In fact, the SOR can be used by itself as the method for ranking the models. Clearly, it awards the top ranks to the models that score most consistently near the top of each category.

SENSITIVITY ANALYSIS

Approaches to Solving the Sensitivity Equation

Two approaches were considered in solving the sensitivity equation

Table 13. Application of sum of ranks method[1].

Model	Rank Assignment for Each FOM						Sum of Ranks	Weighted Sum of Ranks
	F_1	F_2	F_3	F_4	F_5	F_6		
HIWAY-II	5	5	2	2	3	4	21	11.3
CALINE3	1	1	1	1	4.5	1	9.5	4.2
ROADMAP	3	2	4.5	5.5	1	2	18	8.2
GMG	2	3	3	4	4.5	3	19.5	9.3
MROAD2	4	4	5	1	5	5	24	12.7
LAMB	6	6	4.5	5.5	2	6	30	16

[1]Based on Table 11.

$$\text{Var}(x) = B \cdot \text{Var}(P) \cdot B^t \qquad (16)$$

The first approach is the rigorous method, whereby, for each simulation period, the B matrix is numerically solved as part of the simulation. The output sensitivity matrix Var (x) is then obtained by means of the matrix multiplication given above. (Only the diagonals of the sensitivity matrix are really needed to calculate error bars for the simulation.) Hence, the model simulation output includes both the predicted concentrations and the sensitivity matrices (one for each simulation period).

The foregoing method is the most thorough approach. For this work, it may constitute an "overkill" if one can show that acceptable accuracy can be achieved by fewer simulations. It is suspected that the accuracy is limited by errors inherent in the quantification of the P matrix, which requires specification of the model input errors. Accurate quantification of these errors is extremely difficult for certain parameters that are derived from measurements (e.g., atmospheric stability, emission rates) and moderately difficult for other parameters that are measured directly (e.g., winds). Because of these inherent inaccuracies, a second approach is considered that minimizes the computational burden by considering a coarser resolution in computing the B matrix and by computing an x matrix that is consistent with the accuracy of the model input errors. An example of the second approach is described as follows.

An Example of the Simplified Procedure

Estimating the Var (P) matrix primarily relies on analysis of available literature (e.g., previous studies, manufacturers specifications). The general procedure for estimating the B matrix is presented in Figure 30. The main feature of the procedure is a screening operation prior to model runs. The screening operation serves to identify the most influential input parameters. As a result, the model runs can be reduced to a manageable number.

The procedure is used for each model being evaluated. However, results of sensitivity analyses from similar models are used in the screening process (i.e., estimating the sensitivity range for each parameter) and in augmenting results from the model runs. The screening operation can differ for each generic type of model. In this report an example for HIWAY-II is considered which is a Gaussian model.

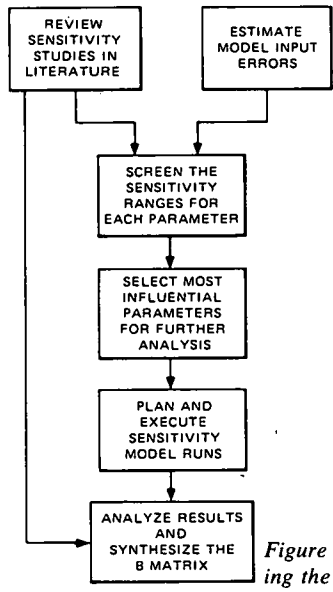


Figure 30. Procedure for estimating the sensitivity matrix.

ESTIMATING INPUT ERRORS AND SCREENING MODEL SENSITIVITY

In accordance with the procedure previously described typical errors and sensitivities for the model input parameters used with a Gaussian model were reviewed. Previous work by Texas A&M University (*31*) is the basis for estimating many typical input errors. Work by Dabberdt et al. (*33*) augments this study in quantifying errors in emission estimates. A study by the California Department of Transportation (*9*) provides a sound basis for screening model output errors given typical model input errors.

Table 14 contains the estimated upper bound in model output errors, given estimated model input errors. Because the functional form of HIWAY-II and most other Gaussian

Table 14. Screened Gaussian model sensitivities based on CALINE3 model results.

Parameter	Typical Error in Input	Error in Model Output[1] (percent)	Sensitivity Functional Form	Matrix Variable
Wind direction	8°	60	Numeric	C_2
Wind speed, u	0.25 u m/s	50	$\alpha = \frac{1}{u^2}$	C_1
Atmospheric stability	One stability category	60	Numeric	C_3
Mixing height	500 m	10	Numeric	Not used
Highway width	< 1 m	Negligible	---	Not used
Highway length	0.01 km	Negligible	---	Not used
Source height	< 1/2 m	Negligible	---	Not used
Surface roughness	100 percent	5	Numeric	Not used
Emissions	40 percent	40	Constant	Q_1

[1]Represents the approximate order of magnitude in expected error, given the indicated input error.

models is similar to that for CALINE3, it is expected the sensitivity matrices (B) to be about the same. The results in Table 14 can now be used to assist in efficiently planning sensitivity runs for the test case (HIWAY-II). For instance, sensitivity runs for source height probably will not be required because the expected output error is negligible. Similarly, one would plan more sensitivity runs for wind direction than for mixing height, because the output is six times more sensitive to errors in wind direction than it is to mixing height.

Sensitivity runs are not needed when the functional forms are known, because they are for emission input data; in this case, errors in the model output are equal to errors in the input emission estimates. Other parameters (e.g., atmospheric stability) require sensitivity runs because the functional forms are not easy to derive by hand and, hence, require a numerical solution. Table 14 shows that most of the expected model output error will result from errors in four input parameters—wind direction, wind speed, atmospheric stability, and emissions.

Thus, the model input has been quantified, and the four control parameters (as shown in Table 14) for the B matrix have been identified as follows:

Q_1 = emission rate;
G_1 = wind speed;
G_2 = wind direction; and
G_3 = atmospheric stability.

Mixing height has been neglected because the model output is not particularly sensitive to it, at least within the accuracy constraints of this analysis.

The Var (P) Matrix

The Var (P) matrix is essentially known at this point if it is assumed that errors among input parameters are uncorrelated. The Var (P) matrix then becomes:

$$P = \begin{bmatrix} \sigma_{11}^2 & 0 & 0 & 0 \\ 0 & \sigma_{22}^2 & 0 & 0 \\ 0 & 0 & \sigma_{33}^2 & 0 \\ 0 & 0 & 0 & \sigma_{44}^2 \end{bmatrix} = \begin{bmatrix} 0.16 & 0 & 0 & 0 \\ 0 & 0.06w^2 & 0 & 0 \\ 0 & 0 & 64 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where σ_{11}, σ_{22}, σ_{33}, and σ_{44} are estimated model input errors for emissions, wind speed, wind direction, and atmospheric stability class, respectively.

The B Matrix

Evaluation of the B matrix requires one to make computer runs to determine the sensitivities not already known. The only sensitivity known is that for emission rate (as given in Table 14, the sensitivity runs to determine model sensitivities for a number of receptor orientations (receptor angle with respect to roadway and receptor distance from the roadway). It was arbitrarily decided to define four B matrices, each as a function of a different downwind distance; these distances were set at 15, 30, 60, and 90 m, respectively. Further, each of the individual B matrices had four rows, with each row a function of a different receptor angle from the roadway; these receptor angles were set at 15°, 30°, 60°, and 90°, respectively.

The next step is to make computer runs which enable one to quantify the B matrix elements. As an example, consider the plots for atmospheric stability class as shown in Figure 31. Four plots are presented for each of the receptor distances. For a unit change in atmospheric stability class, the magnitude for the output change could be approximated as constant for a specific receptor orientation (angle and distance). Hence, each element in the B matrix is a different constant that represents the upper bound of expected sensitivity. The element value is a percentage fraction change of model output magnitude per unit change in the model input parameter.

Referring to Figure 31(a), one can observe the change in $\partial x/\partial G_3$ with receptor angle. The percentage difference between two adjacent curves varies from about 6 percent at a receptor angle of 15° to about 50 percent at a receptor angle of 90°.

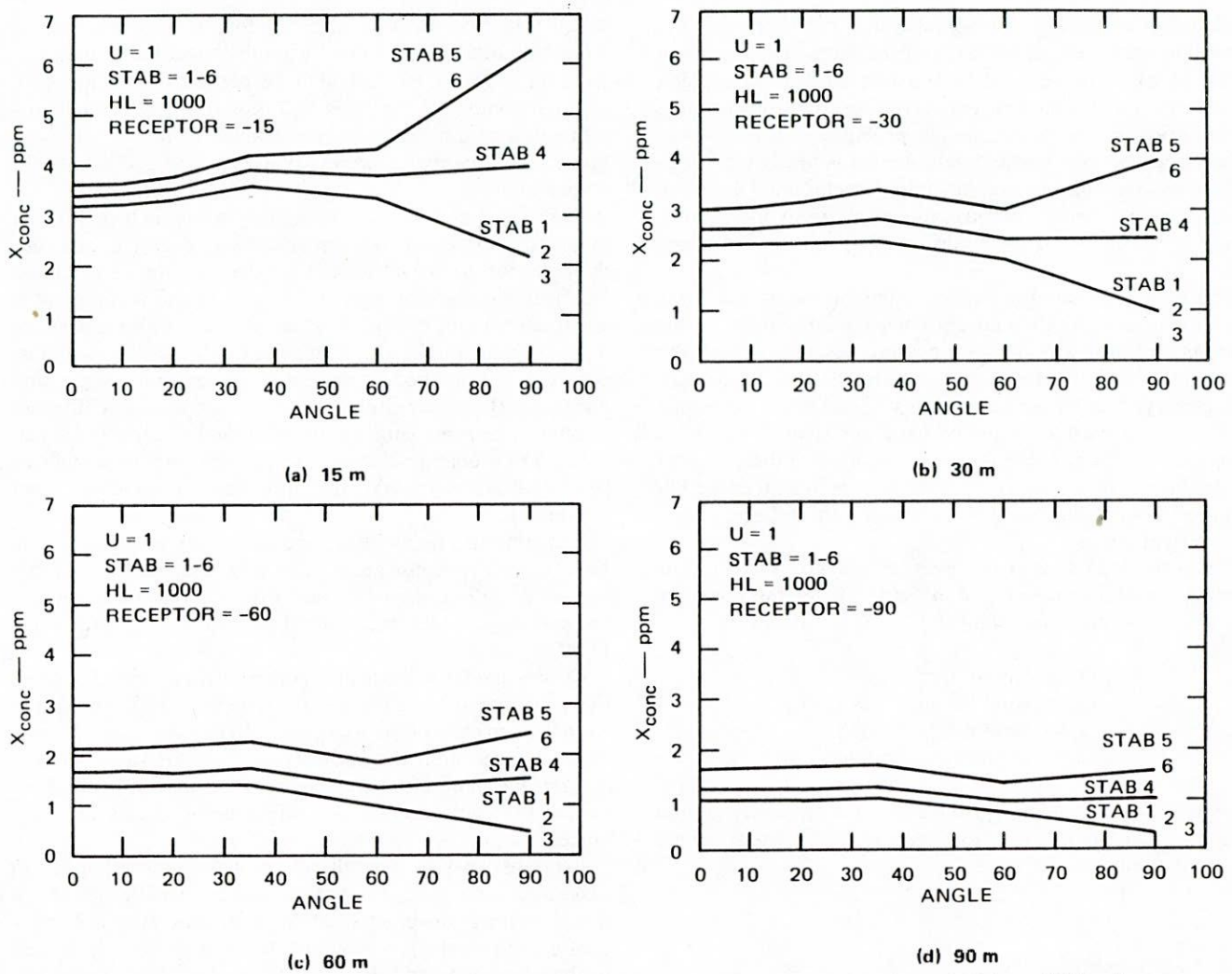
Curves similar to those of Figure 31 were prepared to show the wind speed variance. (Wind speeds of 1, 2, and 5 m/s were used in each of the four plots.) In this case, $\partial x/\partial G_1$ was defined in terms of the known function form (i.e., $-k/u^2$). A different value of k was determined from the wind speed curves for each combination of receptor angles and distances.

For sensitivity to wind direction, it was not necessary to make additional plots. Instead, estimates could be made from the plots for atmospheric stability and wind speed. The receptor angle to the roadway ordinate can be directly related to the estimated wind direction error of 8° (from Table 14). The sensitivity to wind direction variation is simply the slope of the curve at the receptor angle of interest.

Thus, the foregoing procedure provides enough information to allow one to define the B matrix. There is much subjectivity in the actual quantification process, although this can be reduced by opting for more computer runs. In the examples, a probable maximum value was estimated for each element, considering magnitudes only. (It is assumed that there are equal positive or negative model output errors resulting from input errors.)

The B matrix is shown, as follows, for the roadway-to-receptor distance of 15 m:

$$B_{15} = \begin{bmatrix} \frac{\partial \chi_1}{\partial Q} & \frac{\partial \chi_1}{\partial G_1} & \frac{\partial \chi_1}{\partial G_2} & \frac{\partial \chi_1}{\partial G_3} \\ \frac{\partial \chi_2}{\partial Q} & \frac{\partial \chi_2}{\partial G_1} & \frac{\partial \chi_2}{\partial G_2} & \frac{\partial \chi_2}{\partial G_3} \\ \frac{\partial \chi_3}{\partial Q} & \frac{\partial \chi_3}{\partial G_1} & \frac{\partial \chi_3}{\partial G_2} & \frac{\partial \chi_3}{\partial G_3} \\ \frac{\partial \chi_4}{\partial Q} & \frac{\partial \chi_4}{\partial G_1} & \frac{\partial \chi_4}{\partial G_2} & \frac{\partial \chi_4}{\partial G_3} \end{bmatrix}$$



(Each plot is for designated downwind distance; windspeed = 1 m/s, mixing height = 1000 m)

Figure 31. HIWAY-II model output as a function of receptor angle and atmospheric stability classes.

$$B_{15} = \begin{bmatrix} 1 & \frac{3}{u^2} & 0 & 0.06 \\ 1 & \frac{3.5}{u^2} & 0.03 & 0.07 \\ 1 & \frac{4}{u^2} & 0.03 & 0.20 \\ 1 & \frac{6}{u^2} & 0.04 & 0.50 \end{bmatrix}$$

where x_1, x_2, x_3, and x_4 are values at receptor angles of 15°, 30°, 60°, and 90°, respectively.

The sensitivities to emission rates are one in the first column; the second column reflects the dependence on the inverse square of the wind speed. The third column shows that the model is insensitive to wind direction errors for the receptor angle (15°) closest to the line normal to the roadway. The fourth column shows values of sensitivity to atmospheric stability classification errors. Two values (0.06 and 0.05) were discussed earlier.

Matrices similar to the one above were also prepared for distances of 30, 60, and 90 m, but are not presented herein.

The Var (x) Matrix

The Var (x) matrix is calculated from Eq. 5 in Chapter Three. Results for the receptor highway to receptor distance of 15 m are given as follows:

$$\text{Var (x)} = \begin{bmatrix} 0.16 + \frac{0.54}{u^2} & 0.16 + \frac{0.61}{u^2} & 0.17 + \frac{0.72}{u^2} & 0.19 + \frac{1.08}{u^2} \\ 0.16 + \frac{0.61}{u^2} & 0.22 + \frac{0.74}{u^2} & 0.24 + \frac{0.84}{u^2} & 0.27 + \frac{1.26}{u^2} \\ 0.17 + \frac{0.72}{u^2} & 0.24 + \frac{0.84}{u^2} & 0.28 + \frac{0.96}{u^2} & 0.38 + \frac{1.44}{u^2} \\ 0.19 + \frac{1.08}{u^2} & 0.27 + \frac{1.26}{u^2} & 0.38 + \frac{1.44}{u^2} & 0.65 + \frac{2.16}{u^2} \end{bmatrix}$$

The diagonal of Var (x) is model output variances at receptor sites at 15°, 30°, 60°, and 90°, respectively; the square root of these elements is the magnitude of the model output error resulting from the model sensitivity to input errors described in Var (P). The nondiagonal elements are covariances of two different receptor locations which, as stated in Chapter Three, are measures of correlations of the output variations.

Analysis of the Diagonal Elements

The work here focuses on the diagonal elements, because they can be used to describe an error bar in the prediction. For example, tolerances are placed on a prediction for x_1, x_2,

$$x_1 \pm S_{11}$$

$$x_2 \pm S_{22}$$

$$x_3 \pm S_{33}$$

$$\vdots$$

$$x_q \pm S_{qq}$$

Each of the foregoing gives ranges to the prediction that allow for variations induced by errors in the model input data.

The values of the diagonal elements were calculated for each of the four matrices. In the process, four wind speed values were substituted for u. Table 15 contains the result for each of the matrices. Corresponding curves for expected errors are given in Figure 32. Results show that the expected

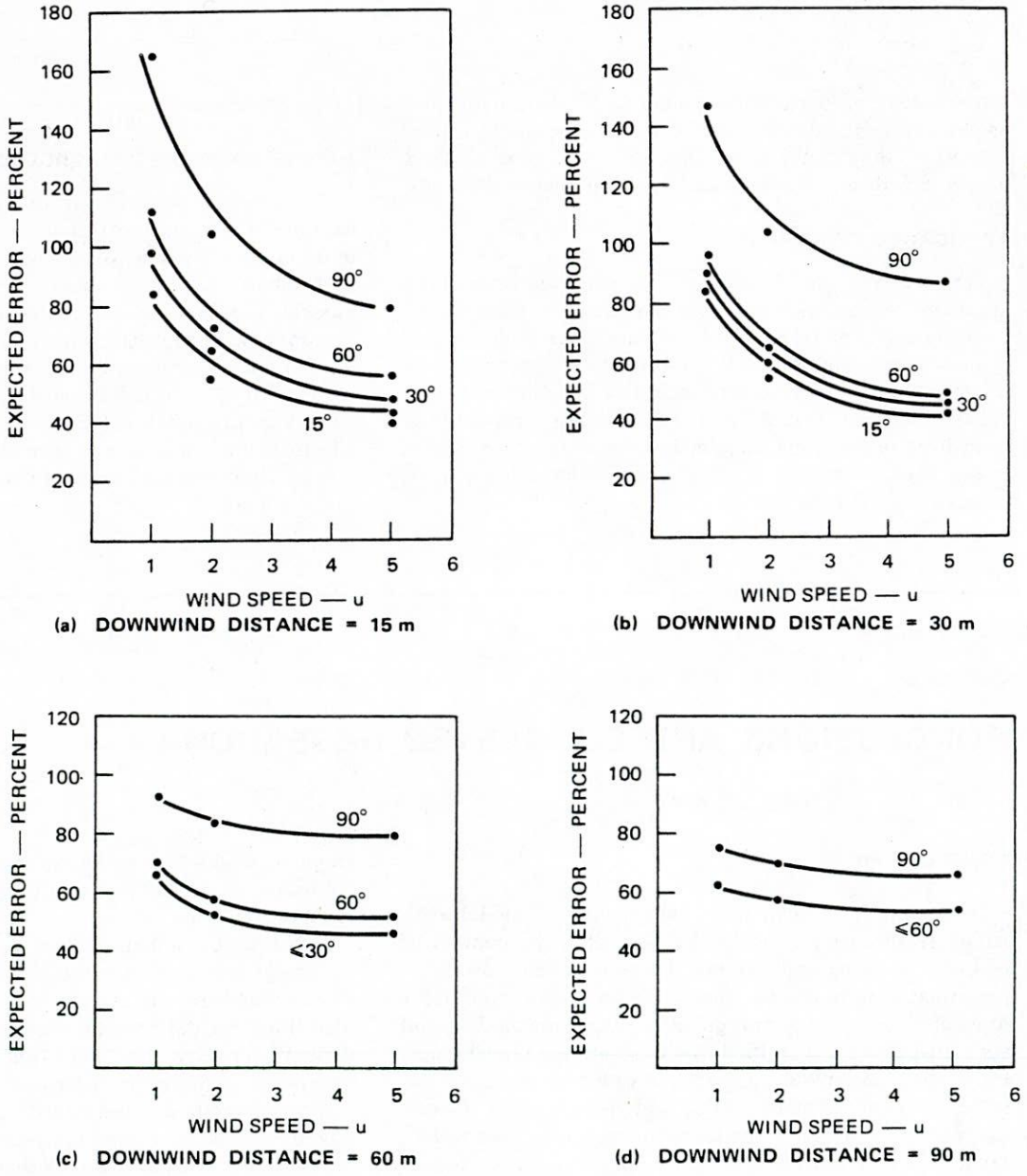


Figure 32. Expected error as a function of downwind distance, roadway angle, and wind speed.

Table 15. Expected error (percent) as a function of distance, angle (θ), and wind speed (u).

Distance (m)	u	θ deg 15	30	60	90
15	1	84	98	111	165
	2	54	64	72	105
	5	43	50	56	80
	∞	40	47	53	75
30	1	84	91	97	146
	2	54	60	63	104
	5	43	48	49	88
	∞	40	46	46	85
60	1	66	66	70	93
	2	51	51	56	82
	5	46	46	51	79
	∞	45	45	50	79
90	1	62	62	63	75
	2	53	53	54	68
	5	51	51	52	66
	∞	50	50	51	66

error could be as low as 40 percent of the prediction for small angles, downwind distances of 30 m and less, and high wind speeds; or they could be as high as 165 percent for large angles, downwind distances of 15 m, and low wind speeds.

APPLICATION OF RESULTS

Table 15 and Figure 32 show that the expected error can be approximated by knowing three parameters—wind speed, roadway angle, and downwind distance. These values could be incorporated into the HIWAY-II model so that the appropriate value for expected error could be included with each model prediction. Treatment of this prediction error could be formalized in much the same manner as were observational errors. (See the section in Chapter Two addressing "Treatment of Data Errors.")

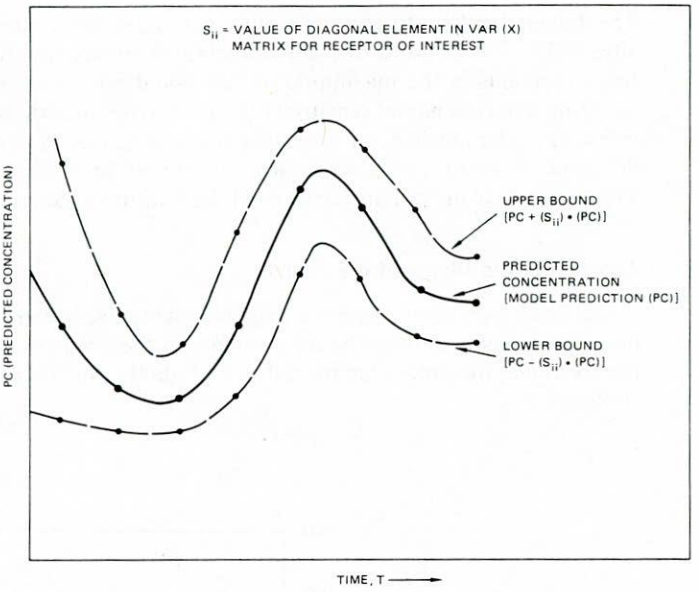


Figure 33. Simulation showing expected error bounds.

EXTENSION OF THE METHODOLOGY

The example presented in this section entailed a simple exercise of the sensitivity analysis methodology. Regardless of the complexity, one can always apply the bare formulation that solves the Var (x) matrix. For numerical simulation models, the solution is time variant and must be treated accordingly. In such cases quantification of the B matrix can be extremely complex, even if one can use crude approximations to most of the B matrix elements. In the complex case, Var (x) must be solved at each simulation time. Figure 33 shows this concept. For each simulation time, the model output will provide a predicted concentration and upper and lower bounds.

CHAPTER SIX

CONCLUSIONS AND SUGGESTED RESEARCH

CONCLUSIONS

In this study, a systematic examination of model evaluation requirements was conducted within the context of highway planning applications. The components of model performance were defined that would be most important for users of highway-related air pollution dispersion models, and six statistics were identified that describe the model's performance quantitatively. Associated with each of the six statistics is a figure of merit (FOM), which is a numerical index that serves to describe model performance in an absolute sense. The individual FOMs were combined in various ways to obtain an overall FOM. These quantitative measures were supplemented with a variety of semiquantitative and graphical diagnostic techniques that provide further insight about the model's behavior.

Concurrently, a methodology was developed for sensitivity analysis that describes the response of a model in the presence of errors in the model's inputs. Such input errors place fundamental limitations on the accuracy of model predictions; sensitivity techniques allow the user to estimate the magnitude of these limitations. The performance statistics, diagnostic methods, and sensitivity techniques were integrated in a model evaluation procedure (MEP). The MEP provides the logical framework that links these three components of the performance assessment with other key ele-

ments such as the composition and size of the data base used to test the model. To assist the user, computer programs were developed for computing the performance statistics, for conducting diagnostic analyses, and for estimating the size of the data base needed to satisfy user-specified requirements on the accuracy of the value of the performance statistics. Finally, a comprehensive data base was assembled and documented for the purpose of conducting full-fledged model evaluations in subsequent studies. A subset of this data base was used in the present study to demonstrate the application of the statistical methodology.

The six statistics used to measure model performance have certain attributes that require some elaboration. First, although these statistics provide a reasonably complete quantitative description of model performance, they are considered to be the minimum number capable of providing such a description. Thus, no claim is made or is intended that the possibilities have been exhausted. Rather, recognizing the special nature of many model verification problems, users are encouraged to supplement these statistics with others that may be especially suitable for a particular problem. By using the six statistics defined in this report, evaluations of model performance conducted by different researchers can be compared on a common basis. Moreover, for these six statistics one can readily compute adjustments that account for observational error, and can calculate confidence intervals using computer programs developed in this study.

Because a quantitative description of model performance was of interest, the six statistics are, of necessity, summary statistics. They summarize in the sense that a single number (i.e., the value of each statistic) measures model response in a particular performance category. By their nature, summary statistics describe complex situations in a simplified manner, and thus can hide or mask patterns in the performance data—knowledge of which may be very valuable to the researcher in gaining a better understanding of model behavior. Such patterns can be discerned by the judicious application of diagnostic analysis. Thus, users are urged to always supplement the summary statistics with diagnostic analyses, which should include at the very least scatterplots of observations and predictions. For this reason considerable effort was devoted to reviewing and applying diagnostic techniques, and to the development of algorithms for diagnostic analysis.

Using several statistics to measure model performance poses the problem of how to compare the performance of different models. For example, is a model that has a low rms error (S_6) and a low correlation coefficient (S_3) preferable to one with a higher rms error but also a higher correlation? Such judgments are inherently subjective and depend on the experience of the researcher, as well as on the nature of the intended application. An attempt has been made to deal with this problem by devising the figure of merit (FOM) numerical scale, which allows one to express the various statistics on a common basis. There is, of course, a large degree of subjectivity involved in constructing the FOM scale, but this is the nature of the problem. The FOM was found to be convenient, helpful, and intuitively appealing. However, no claim is made that the FOM scale defined in this study is unique or is the best. Although this particular FOM scale can be useful in many cases, the concept, rather than the FOM scale itself, is considered to be the more important contribution. Thus, the specific FOM scale is offered as a technique to be investigated.

Regarding the preliminary evaluation of the six selected models, it is emphasized that it is inappropriate to consider the results of the evaluations to be general indicators of the capabilities of these models. The tests were not meant to be definitive, but were to serve only to illustrate the application of the methodology. Thus, the evaluations are limited in scope, and such questions as the accuracy of the emissions inputs could not be addressed. Such matters are the province of model evaluation studies *per se,* for example (*5, 9, 32*). Nevertheless, the evaluation provided some interesting lessons. First, one of the models turned out to be data-base dependent; the performance statistics clearly showed it to be out-of-step with the other models. This experience shows that before an application is undertaken, a model developed using a particular data set should be tested on at least one other data set that is independent of the one originally used. Second, the preliminary analysis did point out some potential deficiencies in the models that require further investigation. The diagnostic analysis also indicated the need for closely scrutinizing the data base to see whether it contains some particular feature that the models find difficult to handle. Finally, as a group, the Gaussian models performed as well as or better than the numerical model that was not data-base dependent. This result is consistent with the findings of Rao et al. (*5*).

SUGGESTED RESEARCH

It is recommended that future research on the topics covered in this study include the following:

- Perform exhaustive evaluations of several dispersion models using the methods of this report and, in particular, using a variety of data sets from the archive assembled in this study.
- Develop a computer program for sensitivity analysis.
- Investigate the construction of other figure-of-merit scales, and devise a general approach for constructing such scales.

In selecting models for evaluation, the choices should be restricted to those models that are extensively documented. Inadequate documentation can result in too much time being spent in getting the numerical models operational, and this situation should be avoided in the future.

CHAPTER SEVEN

BIBLIOGRAPHY ON MODEL EVALUATION

Bornstein, R.D., and Anderson, S.F., "A Survey of Statistical Techniques Used in Validation Studies of Air Pollution Prediction Models." *Technical Report No. 23,* SIAM Institute for Mathematics and Society (SIMS), Department of Statistics, Stanford University (Mar. 1979).

Brier, G.W., "Validity of the Air Quality Display Model Calibration Procedure." *EPA-R4-73-017,* Final Report for National Environmental Research Center (1973).

Burr, J.C., and Clymer, A.B., "Air Modeling in Ohio EPA." *Proc.* Conference on Environmental Modeling and Simulation, Cincinnati, Ohio (Apr. 19–22, 1976).

Calder, K.L., "A Climatological Model for Multiple Source Urban Air Pollution." First Meeting of NATO/CCMS Panel on Modeling, Frankfurt, Germany (Oct. 9, 1970).

Chock, D.P., "General Motors Sulfate Dispersion Experiment: Assessment of the EPA HIWAY Model." *J. Air Poll. Control Assoc.,* Vol. 27, No. 1 (1977) pp. 39–45.

Chock, D.P., "An Advection-Diffusion Model for Pollutant Dispersion near Roadways." *J. Appl. Meteorol.,* Vol. 17 (1978) pp. 926–989.

Chock, D.P., "A Simple Line-Source Model for Dispersion near Roadways." *Atmos. Environ.,* Vol. 12 (1978) pp. 823–829.

Christiansen, J.G., "Design and Application of the Texas Episodic Model." *Proc.,* Conference on Environmental Modeling and Simulation, Cincinnati, Ohio (Apr. 19–22, 1976).

Clarke, J.F., "A Single Diffusion Model for Calculating Point Concentrations for Multiple Sources." *J. Air Poll. Control Assoc.,* Vol. 14, No. 9 (1964) pp. 347–352.

Darling, E.M., Prerau, D.S., Downey, P.J., and Mengert, P.H., *Highway Air Pollution Dispersion Modeling: Preliminary Evaluation of Thriteen Models.* U.S. Department of Transportation, Transportation Systems Center, Cambridge, Mass. (June 1977).

Dowell, K.E., et al., "Diffusion Model Validation from Ambient Air Measurements Around 5 Coal-Fired Electrical Power Plants in Indiana." Fourth Symposium on Turbulence, Diffusion and Air Pollution, Reno, Nev. (Jan. 15, 1979).

Duewer, W.H., MacCracken, M.C., and Walton, J.J., "The Livermore Regional Air Quality Model: II Verification and Sample Application in the San Francisco Bay Area." *J. Appl. Meteorol.* Vol. 17 (1978) pp. 273–311.

Egan, B.A., and Lavery, T.F., "Highway Designs and Air Pollution Potential." Third Urban Technological Conference and Technical Display, Boston, Mass. (Sept. 25–28, 1973).

Elderkin, C.E., et al., "Diffusion-Deposition Measurements and Modeling." Symposium on Atmospheric Diffusion and Air Pollution, Santa Barbara, Calif. (Sept. 9–13, 1974).

Eskridge, R.E., and Demerjian, K.L., "A Highway Model for the Advection, Diffusion and Chemical Reaction of Pollutants Released by Automobiles: Part I—Advection and Diffusion of SF_6 Tracer Gas." Joint Conference on Application of Air Pollution Meteorology, Salt Lake City, Utah (Nov. 29, 1977).

Fortak, H.G., "Numerical Simulation of the Temporal and Spatial Distributions of Urban Air Pollution Concentrations." Symposium on Multiple Source Urban Diffusion Models, Chapel Hill, N.C. (Oct. 27–30, 1969).

Fuggle, R.F., and Dutkiewicz, R.K., "An Air Pollution Survey of Capetown and its Comparison with Selected Computer Models." *Proc.,* Fourth International Clean Air Conference, Tokyo, Japan (May 16–20, 1977).

Gifford, F.A., "The Simple ATDL Urban Air Pollution Model." *Proc.,* Fourth Meeting of the Expert Panel on Air Pollution Modeling. Oberursel, Germany (May 28–30, 1973).

Goodin, W.R., McRae, G.J., and Seinfeld, J.H., "Validity and Accuracy of Atmospheric Air Quality Models." Third Symposium on Atmospheric Turbulence, Diffusion, and Air Quality, Raleigh, N.C. (Oct. 19–22, 1976).

Groff, G.K., "Empirical Comparison of Models for Short Range Forecasting." *Management Sci.,* Vol. 20, No. 1 (1973) pp. 22–31.

Guldberg, P.H., Wright, T.E., and McAllister, A.R., "Use of Climatological Dispersion Model for Air Quality Maintenance Planning in the State of Rhode Island." *Proc.,* Conference on Environmental Modeling and Simulation, Cincinnati, Ohio (Apr. 19–22, 1976).

Habegger, L.J., et al., "Dispersion Simulation Techniques for Assessing the Air Pollution Impacts of Ground Transportation Systems." Argonne National Laboratory, Ill. (Apr. 1974).

Hayes, S.R., "Performance Measures and Standards for Air Quality Simulation Models." *Final Report EPA Contract 68-02-253,* Systems Applications, Inc., San Rafael, Calif. (1979).

Koch, R.C., Pelton, D.J., and Hwang, P.H., "Sampled Chronological Input Model (SCIM) Applied to Air Quality Planning in Two Large Metropolitan Areas." *Proc.,* Conference on Environmental Modeling and Simulation, Cincinnati, Ohio (Apr. 19–22, 1976).

Koch, R.C., and Thayer, S.D., "Validation and Sensitivity Analysis of the Gaussian Plume Multiple-Source Urban Diffusion Model." *Final Report Contract CPA 70-94,* Geomet, Inc. (1971).

Koogler, J.E., et al., "A Multivariate Model for Atmospheric Dispersion Predictions." *J. Air Poll. Control Assoc.,* Vol. 17, No. 4 (1967) pp. 211–214.

Liu, M.K., and Seinfeld, J.H., "A Comparison of the Grid and Trajectory Models of Urban Air Pollution." Symposium on Atmospheric Diffusion and Air Pollution, Santa Barbara, Calif. (Sept. 9–13, 1974).

Maldonado, C., and Bullin, J.A., "Modeling Carbon Monoxide Dispersion from Roadways." *Environ. Sci., and Tech.* Vol. 11, No. 12 (1977) pp. 1071–1076.

McCollister, G.M., and Wilson, K.R., "Linear Stochastic Models for Forecasting Daily Maxima and Hourly Concentrations of Air Pollutants." *Atmos. Environ.*, Vol. 9 (1975) pp. 417–423.

McNider, R.T., "Variability Analysis of Long-Term Dispersion Models." Joint Conference on Application of Air Pollution Meteorology, Salt Lake City, Utah (Nov. 29–Dec. 2, 1977).

Miller, D.R., Butler, G., and Bramall, L., "Validation of Ecological Systems Models." *J. Environ. Mgmt.*, Vol. 4 (1976) pp. 404–407.

Miller, M.E., and Holzworth, G.C., "An Atmospheric Diffusion Model for Metropolitan Areas." *J Air Poll. Control Assoc.*, Vol. 17, No. 1, (1967) pp. 46–50.

Mukherji, S.K., et al., "Modified Dispersion Modeling Procedures for Indiana Power Plants." *Proc.*, Conference on Environmental Modeling and Simulation, Cincinnati, Ohio (Apr. 19–22, 1976).

Mullen, J.B., et al., "Development and Validation of a Model for Diffusion in Complex Terrain." Joint Conference on Applications of Air Pollution Meteorology, Salt Lake City, Utah (Nov. 29, 1977).

Nappo, C.J., "A Method for Evaluating the Accuracy of Air Pollution Prediction Models." *Proc.*, Symposium on Atmospheric Diffusion and Air Pollution, Santa Barbara, Calif. (Sept. 9–13, 1974).

Newman, E., and Spiegler, D.B., "Operational Experience with an Air Quality Control System—AIRMAP." *Proc.*, Symposium on Atmospheric Diffusion and Air Pollution, Santa Barbara, Calif. (Sept. 9–13, 1974).

Noll, K.E., Miller, T.L., and Claggett, M., "A Comparison of Three Highway Line Source Dispersion Models." *Atmos. Environ.*, Vol. 12 (1978) pp. 1323–1329.

Okamoto, S., and Shiozawa, K., "Validation of an Air Pollution Model for the Keihin Area." *Atmos. Environ.*, Vol. 12 (1978) pp. 2139–2149.

Porter, R.A., and Christiansen, J.H., "Modeling of Particulate and Sulfur Dioxide in Support of Ten-Year Planning." *Proc.*, Conference on Environmental Modeling and Simulation, Cincinnati, Ohio (Apr. 19–22, 1976).

Prahm, L.P., and Christensen, M., "Validation of a Multiple Source Gaussian Air Quality Model." *Atmos. Environ.*, Vol. 11 (1977) pp. 791–795.

Prasad, C., "Improvements in Air Quality Display Model." *Proc.*, Conference on Environmental Modeling and Simulation, Cincinnati, Ohio (Apr. 19–22, 1976).

Rao, S.T., Kennan, M., Sistala, G., and Samson, P., "Dispersion of Pollutants Near Highways: Data Analysis and Model Evaluation." EPA-600/4-79-011 (Feb. 1979).

Rao, S.T., and Visalli, J.R., "On the Comparative Assessment of the Performance of Air Quality Models," *J. Am. Poll. Control Assoc.*, Vol. 31 (1981) pp. 851–860.

Reynolds, S.D., et al., "Mathematical Modeling of Photochemical Air Pollution: Vol. III Evaluation of the Model." *Atmos. Environ.*, Vol. 8 (1974) pp. 563–593.

Ruff, R.E., and Javitz, H.S., "Evaluation of Real-Time Air Quality Model (RAM) Using the RAPS Data Base." *Final Report EPA Contract 68-02-2770,* SRI International, Project 6868 (Apr. 1979).

Ruff, R.E., and Simmon, P.B., "Evaluation of Emission Inventory Methodologies for the RAPS Program." Project 4331, Stanford Research Institute, Menlo Park, Calif. (1977).

Shieh, L.J., and Shir, C.C., "Analysis of Input Parameters and Results of Urban Air Pollution Computation." *Proc.*, Third Symposium on Atmospheric Turbulence, Diffusion, and Air Quality, Raleigh, N.C. (Oct. 19–22, 1976).

Smith, D.B., and Ruch, R.B., "Comparative Performance on Complex Terrain of Several Air Quality Impact Assessment Models Based on Aerometric Program Data." Fourth Symposium on Turbulence, Diffusion, and Air Pollution, Reno, Nev. (Jan. 15–18, 1979).

Snee, R.D., "Validation of Regression Models: Methods and Examples." *Technometrics,* Vol. 19, No. 4 (1977) pp. 415–427.

Tesche,T.W., Burton, C.S., and Mirabella, V.A., "Recent Verification Studies with the SAI Urban Airshed Model in the South Coast Air Basin." Fourth Symposium on Turbulence, Diffusion, and Air Pollution, Reno, Nev. (Jan. 15–18, 1979).

Thayer, S.D., "The Development and Validation of an Airport Air Quality Model." Symposium on Atmospheric Diffusion and Air Pollution, Santa Barbara, Calif. (Sept. 9–13, 1974).

Tikvart, J.A., and Mears, C.E., "Application of the Single Source (CRSTER) Model to Power Plants: A Summary." *Proc.*, Conference on Environmental Modeling and Simulation, Cincinnati, Ohio (Apr. 19–22, 1976).

Turner, B.D., Zimmerman, J.R., and Busse, A.D., "An Evaluation of Some Climatological Dispersion Models." Third Meeting of Panel on Modeling of NATO Committee on the Challenges of Modern Society, Paris, France (Oct. 2–4, 1972).

Turner, D.B., "A Diffusion Model for an Urban Area." *J. Appl. Meteorol.*, Vol. 3, No. 1 (1964) p. 83.

Wang, I.T., "Airport Vicinity Air Pollution Study, Model Application and Validation and Air Quality Impact Analysis at Washington National Airport." *Final Report,* Argonne National Laboratory, Energy and Environmental Systems Division (1974).

Zannetti, P., and Switzer, P., "Some Problems of Validation and Testing of Numerical Air Pollution Models." Fourth Symposium on Turbulence, Diffusion, and Air Pollution Reno, Nev. (Jan. 15–18, 1979).

REFERENCES

1. *Federal Register,* Vol. 44, No. 92 (May 10, 1979) pp. 27558–27604.
2. Burns, J.R., "Error Analysis of Nonlinear Simulations: Applications to World Dynamics." *IEEE Trans. Systems, Man and Cybernetics,* Vol. SMC-5, No. 3 (May 1975) pp. 331–340.
3. U.S. Environmental Protection Agency, "SAROAD Users Manual." *APTD-0663* (July 1971).
4. Dabberdt, W.F., "Air Quality on and Near Roadways." SRI International Project 2761, Draft Final Report (Aug. 1977).
5. Rao, S.T., Keenan, M., Sistala, G., and Samson, P., "Dispersion of Pollutants Near Highways—Data Analysis and Model Evaluation." *EPA 600/4-79-011* (Feb. 1979).
6. Downey, P.J., Garlitz, J.D., and Murphy, K.H., "Comparison of Six Highway Air Pollution Dispersion Models Using Synthetic Data." Report *DOT-TSC-OST-76-58* (Sept. 1977).
7. Zimmerman, J.R., and Thompson, R.S., "User's Guide for HIWAY, a Highway Air Pollution Model." *EPA 650/4-74-008* (1974).
8. Rao, S.T., and Keenan, M.T., "Suggestions for Improvement of the EPA-HIWAY Model." New York State Department of Environmental Conservation, Albany, N.Y. (1979).
9. Benson, P.E., "CALINE3—A Versatile Dispersion Model for Predicting Air Pollutant Levels Near Highways and Arterial Streets." *FHWA/CA/TL-79/23,* Interim Report (1979).
10. Chock, D.P., "A Simple Line-Source Model for Dispersion Near Raodways." *GMR-2407,* General Motors Research Publication (1977).
11. Bullin, J.A., Polasek, J.C., and Green, N.J., "Analytical and Experimental Assessment of Highway Impact on Air Quality." *RHWATX79-218-4,* Texas A&M University (Oct. 1978).
12. Bullin, J.A., and Polasek, J.C., "TRAPS II User's Guide." Report *FHWATX 78-218-2* (Mar. 1978).
13. Carpenter, W.A., and Clemaña, G.G., "Analysis and Comparative Evaluation of AIRPOL-4." *VHRTC75-R55,* Virginia Highway and Transportation Research Council, Charlottesville, Va. (1975).
14. Carpenter, W.A., and Clemaña, G.G., "The Theory and Mathematical Development of AIRPOL-4." *VHTRC75-R49,* Virginia Highway and Transportation Research Council, Charlottesville, Va. (1975).
15. Darling, E.M., Prerau, D.S., Downey, P.J., and Mengert, P.H., "Highway Air Pollution Dispersion Modeling: Preliminary Evaluation of Thirteen Models." *DOT-TSC-OST-77-33,* PB271 049 (June 1977).
16. Chock, D.P., Personal communication (Oct. 8, 1979).
17. Kirsch, J.W., and Mason, B.F., "I-205 Highway Impact Study: A Final Report." Oregon State Highway Department (Dec. 1973).
18. Kirsch, J.W., and Mason, B.F., "Mathematical Models for Air Pollution Studies Involving the Oregon I-205 Highway Project." *555-R-76-2744,* Systems, Science and Software, Inc., La Jolla, Calif. (1975).
19. Lamb, R.G., Hogo, H., and Reid, L.E., "A Lagrangian Approach to Modeling Air Pollutant Dispersion." *EPA 600/4-79-023* (Apr. 1979).
20. Danard, M.B., "Numerical Modeling of CO Concentration Near a Highway." *J. Appl. Meteorol.,* Vol. 11 (1972) pp. 947–957.
21. Ragland, K.W., and Peirce, J.J., "Boundary Layer Model for Air Pollutant Concentrations Due to Highway Traffic." *J. Air Poll. Control Assoc.,* Vol. 25 (1975) pp. 48–51.
22. Eskridge, R.E., Binkowski, F.S., Hunt, J.C.R., Clark, T.L., and Demerjian, K.L., "Highway Modeling. Part II: Advection and Diffusion of SF_6 Tracer Gas," *J. Appl. Meteorol.,* Vol. 18, No. 4 (1979) pp. 401–412.
23. Eskridge, R.E., and Hunt, J.C.R., "Highway Modeling, Part I: Prediction of Velocity and Turbulence Fields in the Wake of Vehicles." *J. Appl. Meteorol.,* Vol. 18, No. 4 (Apr. 1979) pp. 387–400.
24. Pitter, R.L., "User's Manual ROADS, PSMOG, VIS1." Oregon Graduate Center, Beaverton, Ore. (1976).
25. Chock, D.P., "An Advection-Diffusion Model for Pollutant Dispersion Near Roadways." *GMR 2590,* No. 39, General Motors Research Publication (Nov. 1977).
26. Lissaman, P.B.S., "A Simple, Unsteady Concentration Model Explicitly Incorporating Ground Roughness and Heat Flux." *Paper No. 73-129,* Air Pollution Control Assoc. Meeting, Chicago, Ill. (June 1973).
27. Egan, B.A., and Lavery, T.F., "Highway Designs and Air Pollution Potential." AIAA 3rd Urban Technology Conference, Boston, Mass. (Sept. 26–28, 1973).
28. U.S. Environmental Protection Agency, "Mobile Source Emission Factors, Final Document." *EPA-400/9-78-005,* Office of Transportation and Land Use Policy, Washington, D.C. (Mar. 1978).
29. U.S. Environmental Protection Agency, "User's Guide to MOBILE1: Mobile Source Emissions Model." Office of Air, Noise, and Radiation, Washington, D.C. (1978).
30. U.S. Environmental Protection Agency, "Modal Program Guide." An update to "Automobile Exhaust Emission Modal Analysis Model." *EPA-460/3-74-005* (1974).
31. Bullin, J.A., Green, N.J., and Polasek, J.C., "Determination of Vehicle Emission Rates from Roadways by Mass Balance Techniques." *Environmental Science and Technology,* Vol. 14 (June 1980) pp. 700–705.
32. Dabberdt, W.F., Shelar, E., Marimont, D., and Skinner, G., "Analyses, Experimental Studies and Evaluations of Control Measures for Air Flow and Air Quality On and Near Highways, Vol. I: Experimental Studies, Analyses, and Model Development." *FHWA-RD-78-179,* SRI International (Dec. 1979).

33. KISH, L., and FRANKEL, M.R., "Balanced Repeated Replications for Standard Errors." *J. Am. Statistical Assoc.*, Vol. 65, No. 331 (Sept. 1970) pp. 1071–1094.
34. MILLER, R.G., "The Jackknife—A Review." *Biometrika*, Vol. 61, No. 1 (1974) pp. 1–15.
35. EFRON, B., "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics*, Vol. 7, No. 1 (1979) pp. 1–26.
36. VISALLI, J.R., "Effects of Variable Vehicular Age and Classification Distributions in Mobile Source Modeling." *J. Air Poll. Control Assoc.*, Vol. 31, No. 1 (1981) pp. 68–71.

APPENDIX A
MATHEMATICAL NOTATION

The definitions and derivations in Appendices B, C, and D require the use of a standardized mathematical notation. This appendix defines that mathematical notation.

CONCENTRATIONS

OC(i,j)	the observed concentration at the ith monitor on the jth day
LOC(i,j)	the natural logarithm of OC(i,j)
OC(i,[j])	the jth largest observation at the ith monitor
OC(•,[j])	the jth largest observed concentration among all monitors.

The notation for the predicted concentration is obtained by substituting PC for OC.

The notation for the true concentrations (e.g., the observed concentrations less observational error) is obtained by substituting TC for OC.

The notation for the observational error is obtained by substituting ERR for OC.

The letters M or A may be appended to the symbols OC, TC, and ERR in order to denote the observed concentration, true concentration, and observational error under the multiplicative and additive models, respectively.

CONSTANTS

AQS	the air quality standard, or any threshold value used in the computation of S_2.

DATA BASE SIZE

I	the number of monitors
J	the number of time periods
N(i,•)	the number of observation-prediction pairs at the ith monitor
N(•,•)	the total number of observation-prediction pairs
n(i,•)	the integer portion of {0.95 x N(i,•) + 1}
n(•,•)	the integer portion of {0.95 x N(•,•) + 1}
N(•,j)	the number of observation-prediction pairs in the jth time period.

FUNCTIONS

IND(x)	the indicator function of x, equal to 1.0 if x is true and 0.0 otherwise
PHI(x)	the probability that a standard normal deviate is less than or equal to x, also denoted $\Phi(x)$.

SAMPLE MOMENTS

SMEAN{OC(i,•)}	the sample mean of the observed concentrations at the ith monitor $= \sum_{j=1}^{N(i,\cdot)} OC(i,j)/N(i,\cdot)$
SVAR{OC(i,•)}	the sample variance of the observed concentrations at the ith monitor $= \sum_{j=1}^{N(i,\cdot)} [OC(i,j) - SMEAN\{OC(i,\cdot)\}]^2 \times 1/[N(i,\cdot) - 1]$
SSTD{OC(i,•)}	the sample standard deviation at the ith monitor, obtained as the square root of SVAR {OC(i,•)}

SCOVAR{OC(i,·), PC(i,·)}

the sample covariance of the observation-prediction pairs at the ith monitor

$$= \sum_{j=1}^{N(i,\cdot)} ([OC(i,j) - SMEAN\{OC(i,\cdot)\}] \times [PC(i,j) - SMEAN\{PC(i,\cdot)\}] \times 1/[N(i,\cdot) - 1])$$

The sample moments may also be computed for the entire monitoring network. In this case, the symbol "·" replaces the symbol "i" in the definition.

The sample moments may be computed for the jth time period over all monitors by substituting "·" and "j" for "i" and "·" respectively in the definition.

TRUE MOMENTS

E{OC(i,j)} the expected value of OC(i,j)

STD{OC(i,j)} the standard deviation of OC(i,j)

VAR{OC(i,j)} the variance of OC(i,j)

COVAR{OC(i,j), PC(i,j)}

the covariance of OC(i,j) and PC(i,j)

CORR{OC(i,j), PC(i,j)}

the correlation coefficient of OC(i,j) and PC(i,j)

DISTRIBUTIONS

NORM[E{OC(i,j), VAR{OC(i,j)}]

a normal distribution with mean E{OC(i,j)} and variance VAR{OC(i,j)}

LNORM[E{LOC(i,j), VAR{LOC(i,j)}]

a lognormal distribution; the natural logarithm of any variable with this distribution follows the previously defined normal distribution

CHISQ[N(·,·)] a chi-square distribution with N(·,·) degrees of freedom.

OMISSION OF CERTAIN SYMBOLS

The symbols "(·,·)" and "(i,j)" may be omitted to simplify the notation. For example, OC denotes OC(i,j) and SVAR{OC} denotes SVAR{OC(·,·)}. Note that confusion as to which symbol is being omitted cannot arise--neither OC(·,·) nor SVAR{OC(i,j)} is defined.

APPENDIX B
DEFINITION OF THE COMPONENT STATISTICS

In this appendix we define the six component statistics, denoted by S_1 through S_6, that measure the ability of the air quality model to predict the observed concentrations in the monitoring network. We also define more localized versions of S_1 through S_6 that are specific for an individual monitoring instrument or for a given time period. These localized statistics are not used in the FOM but may be useful for diagnostic analysis.

THE RATIO OF THE LARGEST OBSERVATIONS AND PREDICTIONS (S_1)

The first component statistic of the FOM is the ratio of the largest 5 percent of the observed concentrations to the largest 5 percent of the predicted concentrations. For the ith monitoring station this statistic is defined as*

$$S_1(i,\cdot) = \sum_{j=n(i,\cdot)}^{N(i,\cdot)} OC(i,[j]) \; / \; \sum_{j=n(i,\cdot)}^{N(i,\cdot)} PC(i,[j])$$

*The mathematical notation used herein was defined in Appendix A.

For the entire monitoring network this statistic is defined as

$S_1 = S_1(\cdot,\cdot)$ = The above equation with the symbol "·" substituted everywhere for "i".

Note that in these definitions it is not required that the largest observations and predictions occur on the same day.

THE DIFFERENCE BETWEEN THE PREDICTED AND OBSERVED PROPORTIONS OF EXCEEDANCES (S_2)

The second component of the FOM measures the difference between the predicted and observed proportion of exceedances of a concentration threshold, which may be the air quality standard. For the ith monitor, this statistic is defined as

$$S_2(i,\cdot) = \frac{1}{N(i,\cdot)} \sum_{j=1}^{N(i,\cdot)} IND[PC(i,j) > AQS] - \frac{1}{N(i,\cdot)} \sum_{j=1}^{N(i,\cdot)} IND[OC(i,j) > AQS]$$

For the entire monitoring network this statistic is defined as

$S_2 = S_2(\cdot,\cdot)$ = The above equation with the symbol "·" substituted everywhere for "i".

PEARSON'S CORRELATION COEFFICIENT (S_3)

The third component is Pearson's correlation coefficient. For the ith monitor, this statistic is defined as

$$S_3(i,\cdot) = \frac{\text{SCOVAR}\{LOC(i,\cdot)\ ,\ LPC(i,\cdot)\}}{\text{SSTD}\{LOC(i,\cdot)\}\ \text{SSTD}\{LPC(i,\cdot)\}} \quad ,$$

and for the entire network this statistic is defined as

$S_3 = S_3(\cdot,\cdot)$ = The above equation with the symbol "." substituted every where for "i" .

THE TEMPORAL CORRELATION COEFFICIENT (S_4)

The temporal correlation coefficient is the fourth component statistic of the FOM. For the ith monitor, this statistic is defined as

$$S_4(i,\cdot) = S_3(i,\cdot) \quad .$$

For the entire network this statistic is defined as

$$S_4 = \tanh\left[\frac{1}{I}\sum_{i=1}^{I} \tanh^{-1}\{S_3(i,\cdot)\}\right]$$

where "tanh" is the hyperbolic tangent function and "$\tanh^{-1}$" is the inverse hyperbolic tangent function.

THE SPATIAL CORRELATION COEFFICIENT (S_5)

The spatial correlation coefficient is the fifth component of the FOM. For the jth time period, this statistic is defined as

$$S_5(\cdot,j) = S_3(\cdot,j) \quad ,$$

and for the entire time period this statistic is defined as

$$S_5 = \tanh\left[\frac{1}{J}\sum_{j=1}^{J} \tanh^{-1}\{S_3(\cdot,j)\}\right] \quad .$$

THE ROOT-MEAN-SQUARE ERROR (S_6)

The root mean square error is the sixth component of the FOM. For the ith monitor, this statistic is defined as

$$S_6(i,\cdot) = \left\{\sum_{j=1}^{N(i,\cdot)} [OC(i,j) - PC(i,j)]^2 \,/\, N(i,\cdot)\right\}^{1/2} \quad .$$

For the entire network, this statistic is defined as

$S_6 = S_6(\cdot,\cdot)$ The above equation with "." substituted for "i" .

APPENDIX C
DERIVATION OF THE CONFIDENCE INTERVALS FOR THE COMPONENT STATISTIC

As discussed in Chapter Two of this report, both parametric and nonparametric confidence intervals were computed for the component statistics. This appendix presents the mathematical derivation of those confidence intervals.* A computer program that computes confidence intervals is described in Appendix E.

THE PARAMETRIC CONFIDENCE INTERVALS

The parametric confidence intervals (PCI) depend on assumptions concerning the form of the joint distribution of OC and PC. First, the approximately I x J pairs [OC(i,j), PC(i,j)] are assumed to be mutually independent. For the statistics S_1 through S_5 we assume that LOC(i,j) and LPC(i,j) follow a bivariate normal distribution with parameters ({LOC}, E{LPC}, VAR{LOC}, VAR{LPC} and CORR{LOC,LPC}). For the statistic S_6 we assume that OC(i,j) - PC(i,j) follows a normal distribution with mean zero and variance VAR {OC-PC}.

*The mathematical notation used herein was defined in Appendices A and B.

The PCI for S_1

If the numerator and denominator of S_1 are divided by N-n+1, we can recognize that S_1 is an estimate of the mean of the distribution formed from the largest 5 percent of an LNORM (E{LOC}, VAR{LOC}) distribution divided by the mean of the distribution formed from the largest 5 percent of an LNORM (E{LPC}, VAR{LPC}) distribution. Note that these means, denoted M_1S_1 and M_2S_1 respectively, are functions only of the unknown parameters of the distribution and the sample size.

We desire to estimate the variability of S_1 around M_1S_1/M_2S_2. We denote the second moment of S_1 around M_1S_1/M_2S_2 by VS_1. Under the assumption that $S_1/(VS_1)^{1/2}$ is approximately normally distributed with mean M_1S_1/M_2S_2 and variance one, a 95 percent confidence interval for M_1S_1/M_2S_2 is $S_1 \pm 1.96(VS_1)^{1/2}$. The difficulty in computing this confidence interval is that VS_1 is itself a complicated function of the parameters of the joint bivariate distribution of LOC and LPC. We hypothesize, however, that VS_1 is a relatively stable function of the unknown parameters, so that VS_1 may be approximated using a priori estimates of the unknown parameters. A technique to approximate VS_1 is discussed later in this appendix.

The PCI for S_2

The S_2 statistic is an estimate of the difference between the probability that the PC exceeds the threshold concentration (denoted AQS) and the probability that the OC exceeds the threshold concentration. We

shall denote these probabilities as P_1S_2 and P_2S_2 respectively. To determine a 95 percent confidence interval for P_1S_2/P_2S_2 we must compute the variance of S_2, which we shall denote by VS_2. Because $S_2/(VS_2)^{1/2}$ is approximately normally distributed with mean P_1S_2/P_2S_2 and variance 1.0, a 95 percent confidence interval for P_1S_2/P_2S_2 is $S_2 \pm 1.96 (VS_2)^{1/2}$. As before, VS_2 is itself a function of the unknown parameters of the joint distribution of OC and PC. The approximation of VS_2, given a priori estimates of these parameters, is discussed later in this appendix.

The PCI for S_3

Under the assumption that LOC and LPC follow a bivariate normal distribution, a 95 percent confidence interval for the correlation coefficient, denoted CORR{LOC,LPC} is given by

$$\left[\tanh\left(\tanh^{-1} S_3 - \frac{1.96}{\sqrt{N-2}}\right), \tanh\left(\tanh^{-1} S_3 + \frac{1.96}{\sqrt{N-2}}\right)\right]$$

Consequently, it is only necessary for the investigator to provide an a priori estimate of S_3 (e.g., of CORR{LOC,LPC}) to compute the width of the interval for any given sample size. By varying N, the investigator can determine the initial sample size.

The PCI for S_4

Under the assumption that LOC and LPC follow a bivariate normal distribution, with a common correlation coefficient CORR{LOC,LPC} for all sites, a 95-percent confidence interval for this parameter is given by [tanh(AI - 1.96BI), tanh(AI + 1.96BI)], where

$$AI = \sum_{i=1}^{I} \tanh^{-1} S_3(i,\cdot) \;/\; I$$

$$BI = \left[\sum_{i=1}^{I} \frac{1}{N(i,\cdot) - 2}\right]^{1/2} / \; I$$

Consequently, it is only necessary for the investigator to provide an a priori estimate of AI [e.g., of $\tanh^{-1}$ (CORR{LOC,LPC})] to calculate the width of the interval for any given sample size. If we assume that the $N(i,\cdot)$ are approximately equal to N/I, then BI is equal to $(N - 2I)^{-1/2}$.

The PCI for S_5

Under the assumption that LOC and LPC follow a bivariate normal distribution, with a common correlation coefficient CORR{LOC,LPC} for all sites, a 95-percent confidence interval for this parameter is given by [tanh (AJ - 1.96 x BJ), tanh (AJ + 1.96 x BJ)], where

$$AJ = \sum_{j=1}^{J} \tanh^{-1} S_3(\cdot,j) \;/\; J$$

$$BJ = \frac{1}{J}\left[\sum_{j=1}^{J} \frac{1}{N(\cdot,j) - 2}\right]^{1/2}$$

If the N(·,j) are all approximately equal to N/J, then BJ simplifies to $(N - 2J)^{-1/2}$. As before, when the investigator provides an a priori estimate of AJ [e.g., of $\tanh^{-1}$ (CORR{LOC,LPC})], the width of the confidence interval can be calculated for any given sample size.

The PCI for S_6

Under the assumption that OC and PC follow a normal distribution with mean zero and standard deviation STD{OC-PC}, it can be shown that S_6 is distributed as

$$\text{STD\{OC-PC\}}\ [\text{CHISQ}(N)/N]^{1/2}$$

We note that for N larger than 30, the lower and upper 2.5 percent cutoff values of a $(\text{CHISQ}(N)/N)^{1/2}$ distribution are approximately $1 - 1.96/(2N)^{1/2}$ and $1 + 1.96/(2N)^{1/2}$. This allows us to write the 95 percent confidence interval for STD{OC-PC} as

$$\left[S_6 / \left(1 + \frac{1.96}{\sqrt{2N}}\right),\ S_6 / \left(1 - \frac{1.96}{\sqrt{2N}}\right) \right]$$

Once the investigator provides an a priori estimate of S_6, (e.g., of STD{OC-PC}), the above formula may be used to compute the width of the confidence interval for any given N.

A nonparametric confidence interval (NPCI) may be constructed for each of the component statistics of the FOM. Because these intervals do not require the imposition of a parametric density function, they are more likely to yield correct answers than the parametric confidence intervals discussed earlier. In addition, the NPCI are computed after the data set is chosen and therefore avoid the inaccuracies associated with a priori estimates used with the PCI.

There are four related methods for deriving nonparametric confidence intervals that have gained acceptance in the last decade. These methods are: true replication, pseudo-replication (33), jackknifing (34), and bootstrapping (35). Each of these methods estimates the variability in a component statistic attributable to the finiteness of the data base by computing values for the component statistic on subsets of the data base. There are no practically important restrictions on the form of the underlying probability distributions (e.g., they need not be lognormal). However, in the theoretical derivation of these methods the independence of successive observation-prediction pairs is assumed. The validity of two of the methods--jackknifing and bootstrapping--appears to be sensitive to violation of the independence assumption. The remaining two methods--true and pseudo-replication--appear to be less sensitive to violations of independence assumptions. Consequently, the latter two methods seem preferable to either jackknifing or bootstrapping. We have selected true replication because it is easier to

implement and to explain than is pseudo-replication. We illustrate the true replication method using the S_2 statistic. The S_2 statistic is the proportion of predicted concentrations above a threshold level less the proportion of observed concentrations above the threshold. If our data base of observation-prediction pairs has 20 percent of the predictions and 15 percent of the observations exceeding the threshold concentration, then S_2 is computed to be 5 percent. The true replication method for computing the variability of S_2 requires that the data base be divided into subsets that are similar to each other and to the total data base. Although it is customary to divide the data base into tenths, we believe that because the independence assumption may be violated it is advisable to use fifths. The data base will be divided into fifths by taking groups of five days at a time and randomly assigning these days to the data subsets so that each subset receives one of the days. Using this method, the sequential correlation between hourly observation-prediction pairs in each data subset will probably be only slightly smaller than the correlation in the overall data base, and the composition of each data subset will be approximately equal in terms of the number of hourly pairs, number of days, number of Mondays through Sundays, and the distribution of those days over the months in the data base. Let the S_2 statistic computed over the Kth data subset be denoted by $S_2S(K)$, K = 1, ..., 5. Intuitively, if the five $S_2S(K)$ statistics are close to one another, then the statistic S_2 which is based on a sample size five times as large as any $S_2S(K)$ should have a small variance. In fact, the estimate of the variance of S_2 is $SVAR\{S_2S\}/5$, where

$SVAR\{S_2S\}$ is the sample variance of the $S_2S(K)$ statistics. Because S_2 follows approximately a normal distribution, a 95 percent confidence interval for the expectation of S_2 (the difference in the probabilities of an exceedance) can be computed as

$$S_2 \pm t_{0.975,4} \frac{SSTD\{S_2S\}}{\sqrt{5}}$$

where $t_{0.975,4} = 2.776$ is the upper 0.975-percentile point of the Student's t distribution with 4 degrees of freedom.

COMPUTATION OF CONSTANTS REQUIRED IN THE PARAMETRIC CONFIDENCE INTERVALS FOR S_1 AND S_2

Under the model that LOC(i,j) and LPC(i,j) are bivariate normal with means E{LOC} and E{LPC}, variances VAR{LOC} and VAR{LPC} and correlation CORR{LOC,LPC}, the confidence intervals for S_1 and S_2 require the computation of constants denoted M_1S_1, M_2S_1, VS_1, and VS_2. These constants are defined as follows:

M_1S_1 = the mean of the distribution formed from the largest 5 percent of an LNORM(E{LOC}, VAR{LOC}) distribution.

M_2S_1 = the mean of the distribution formed from the largest 5 percent of an LNORM (E{LPC}, VAR{LPC}) distribution.

VS_1 = the second moment of S_1 around M_1S_1/M_2S_1.

VS_2 = the variance of S_2.

The computation of the above constants begins with the a priori specification of E{LOC}, E{LPC}, VAR{LOC}, VAR{LPC}, and CORR{LOC,LPC}. If we let X denote a variable with a NORM (E{LOC}, VAR{LOC}) distribution then we can write

$$M_1S_1 = E\{\exp(X)\ IND\ [X > E\{LOC\} + 1.96\ STD\{LOC\}]\} \ /$$
$$P\ [X > E\{LOC\} + E\{LOC\} + 1.96\ STD\{LOC\}]$$

$$A = E\{LOC\} + 1.96\ STD\{LOC\}$$

$$B = E\{LOC\} + 1.96\ STD\{LPC\}$$

$$M_1S_1 = 20 \times (2\pi)^{-1/2}\ (STD\{LOC\})^{-1} \int_A^{\infty} \exp\left[\frac{(x - E\{LOC\})^2}{2\ VAR\{LOC\}} + x\right] dx$$

and similarly

$$M_2S_1 = 20 \times (2\pi)^{-1/2}\ (STD\{LPC\})^{-1} \int_B^{\infty} \exp\left[\frac{(x - E\{LPC\})^2}{2\ VAR\{LPC\}} + x\right] dx$$

Consequently M_1S_1 and M_2S_2 can be determined by numerical integration.

The constant VS_1 may be computed using a Monte Carlo program. On the Kth iteration of the program we generate the random sample

$$X_1(K,j)\ ,\ Y_1(K,j) \qquad 1 \le j \le N$$

following a bivariate normal distribution with parameters E{LOC}, E{LPC}, VAR{LOC}, VAR{LPC}, and CORR{LOC,LPC}.

We compute

$$Z_1(K) = \frac{\sum_{j=n}^{N} \exp\ [X_1(K,[j])]}{\sum_{j=n}^{N} \exp\ [Y_1(K,[j])]} \ .$$

The estimate of VS_1 is

$$VS_1 = \frac{1}{19} \sum_{K=1}^{20} \left(Z_1(K) - \frac{M_1S_1}{M_2S_1}\right)^2 \ .$$

The constant VS_2 was defined as

$$VS_2 = VAR\left\{N^{-1} \sum_{j=1}^{N} IND(PC(\cdot,j) > AQS) - N^{-1} \sum_{j=1}^{N} IND(OC(\cdot,j) > AQS)\right\}$$

The above expression simplifies into the equation

$$VS_2 = N^{-1}\ VAR\{INC(PC > AQS)\} + N^{-1}\ VAR\{IND(OC > AQS)\} - 2N^{-1}\ COV\{IND(PC > AQS),\ IND(OC > AQS)\}$$

If we let

$$P_1 = PROB\{PC > AQS\} = 1 - PHI\left(\frac{\ln AQS - E\{LPC\}}{STD\{LPC\}}\right)$$

$$P_2 = PROB\{OC > AQS\} = 1 - PHI\left(\frac{\ln AQS - E\{LOC\}}{STD\{LOC\}}\right)$$

$$Q1 = 1 - P_1$$

$$Q2 = 1 - P_2$$

then we may write

$$VS_2 = \frac{P_1 \times Q1}{N} + \frac{P_2 \times Q2}{N} - \frac{2}{N}\ COV\{IND(PC > AQS)\ ,\ IND(OC > AQS)\}$$

The covariance term may be decomposed as follows

$$COV\{IND(PC > AQS)\ ,\ IND(OC > AQS)\}$$

$$= E\ \{[IND(PC > AQS) - P_1]\ [IND(OC > AQS) - P_2]\}$$

$$= PROB\{PC < AQS\ ,\ OC < AQS\} \times Q1 \times Q2 + PROB\{PC < AQS\ ,\ OC > AQS\} \times Q1 \times (-P_2) + PROB\{PC > AQS\ ,\ OC < AQS\} \times (-P_1) \times Q2 + PROB\{PC > AQS\ ,\ OC > AQS\} \times (-P_1) \times (-P_2)$$

Letting

$$P_3 = PROB\{PC > AQS\ ,\ OC > AQS\}$$

$$Q3 = 1 - P_3$$

we can show that

$$PROB\{PC < AQS\ ,\ OC < AQS\} = 1 - P_1 - P_2 + P_3$$

$$PROB\{PC < AQS\ ,\ OC > AQS\} = P_1 - P_3$$

$$PROB\{PC > AQS\ ,\ OC < AQS\} = P_2 - P_3$$

Substituting backwards and simplifying we find

$$COVAR\{IND(PC > AQS)\ ,\ IND(OC > AQS)\} = P_3 - P_1 \times P_2$$

$$VS_2 = \frac{P_1 \times Q1}{N} + \frac{P_2 \times Q2}{N} - 2\ \frac{(P_3 - P_1 \times P_2)}{N}$$

To calculate VS_2 we only require a numerical value for P_3, which can be obtained by numerical integration of the bivariate normal density with means of zero, variances of 1 and a correlation of CORR{LOC,LPC} over the region where the x coordinate is greater than (ln AQS - E{LOC})/STD{LOC} and the y coordinate is greater than (ln AQS - E{LPC})/STD{LPC}.

APPENDIX D
DERIVATION OF THE ADJUSTMENTS FOR OBSERVATIONAL ERROR IN THE COMPONENT STATISTICS

As discussed in Chapter Two of this report, adjustments to the component statistics for observational error were computed under a multiplicative and an additive error model. This appendix presents the derivation of those adjustments.*

THE MODEL SPECIFICATIONS

The multiplicative model may be written as

$$OCM = TCM[1 + ERRM]$$

where

TCM ~ LNORM(E{LTCM} VAR{LTCM})

ERRM ~ NORM(0, VAR {ERRM})

*The mathematical notation used herein was defined in Appendices A and B.

and these two variables are independent. In this model, STD{ERRM} is expressed as a percentage, e.g., 4 percent. If we take natural logarithms in this model we obtain

$$LOCM = LTCM + \ln[1 + ERRM]$$

Because the standard deviation of ERRM is small, this expression may be simplified. Taking a first-order Taylor expansion of the last term, we find that the model may be approximated as

$$LOCM = LTCM + ERRM$$

Under this latter specification

$$LOCM \sim NORM(E\{LTCM\}, VAR\{LTCM\} + VAR\{ERRM\})$$

The additive model may be written as

$$OCA = TCA + ERRA$$

Here,

$$TCA \sim LNORM(E\{LTCA\}, VAR\{LTCA\})$$

$$ERRA \sim NORM(0, VAR\{ERRA\})$$

and these two variables are independent. The STD{ERRA} is expressed in ppm, e.g., ± 2 ppm. Under this model

$$OCA \sim LNORM(E\{LTCA\}, VAR\{LTCA\}) + NORM(0, VAR\{ERRA\})$$

THE RATIO OF THE LARGEST OBSERVATIONS AND PREDICTIONS

The bias in S_1 is the amount by which S_1 differs from the value that would be obtained by substituting TC for OC in the definition of S_1. Without this substitution, the expected value of the numerator of S_1 is the expected value of the sum of the largest N-n+1 order statistics out of N samples from either an LNORM(E{LTCM}, VAR{LTCM} + VAR{ERRM}) distribution or from an LNORM (E{LTCM}, VAR{LTCA}) + NORM (0, VAR[ERRA}) distribution, depending on whether the multiplicative or additive model is used. The expectation will be denoted CS_1MTE under the multiplicative model and CS_1ATE under the additive model. With the substitution of TC for OC, the expected value of the numerator of S_1 is the expectation of the sum of the largest N-n+1 order statistics out of N samples from either an LNORM(E{LTCM}, VAR{LTCM}) distribution or from an LNORM(E{LTCA}, VAR{LTCA}) distribution depending, on whether the multiplicative or additive model is used. The expectation will be denoted CS_1MT under the multiplicative model and CS_1AT under the additive model. Because CS_1MTE is larger than CS_1MT, and CS_1ATE is larger than CS_1AT, S_1

is biased upward. We define the S_1 statistic adjusted for multiplicative and additive error by the formulae

$$S_1M = S_1 * CS_1MT/CS_1MTE$$

$$S_1A = S_1 * CS_1AT/CS_1ATE$$

Methods for computing CS_1MTE, CS_1ATE, CS_1MT, and CS_1AT are given later in this appendix.

THE DIFFERENCE BETWEEN THE PREDICTED AND OBSERVED PROPORTIONS OF EXCEEDANCES

The bias in S_2 is the amount that S_2 differs from the value that would be obtained by substituting TC for OC in the definition of S_2. Without this substitution, the expected value of the second term in S_2 is the probability that a variable following an LNORM(E{LTCM}, VAR{LTCM} + VAR{ERRM}) distribution is greater than the AQS, or the probability that a variable following an LNORM(E{LTCA}, VAR{LTCA}) + NORM(0, VAR{ERRA}) distribution is greater than the AQS, depending on whether the multiplicative or additive model is used. The probability will be denoted CS_2MTE under the multiplicative model and CS_2ATE under the additive model. With the substitution of TC for OC, the expectation of the second term in S_2 is either the probability that a variable following an LNORM(E{LTCM}, VAR{LTCM}) distribution is greater than the AQS, or the probability that a variable following an LNORM(E{LTCA}, VAR{LTCA}) distribution is greater than the AQS, depending on whether the

multiplicative or additive model is used. This probability will be denoted CS_2MT under the multiplicative model and CS_2AT under the additive model. Because CS_2MTE is larger than CS_2MT and CS_2ATE is larger than CS_2AT, S_2 is biased downward. We denote the S_2 statistic adjusted for multiplicative error by S_2M, and for additive error by S_2A. The statistic S_2M is obtained from S_2 by multiplying the second term in its definition by CS_2MT/CS_2MTE and the statistic S_2A is obtained from S_2 by multiplying the second term in its definition by CS_2AT/CS_2ATE. The calculation of CS_2MTE, CS_2ATE, CS_2MT, and CS_2AT are discussed later in this appendix.

PEARSON'S CORRELATION COEFFICIENT

The bias in S_3 is the amount that S_3 differs from the value that would be obtained by substituting TC for OC in the definition of S_3.

Under the multiplicative model, S_3 is an estimate of

$$\frac{\text{COVAR\{LOCM,LPC\}}}{\text{STD\{LOCM\} STD\{LPC\}}}$$

When TCM is substituted for OCM, S_3 is an estimate of

$$\frac{\text{COVAR\{LTCM,LPC\}}}{\text{STD\{LTCM\} STD\{LPC\}}}$$

Because ERRM is independent of LPC, it can be shown that

$$\text{COVAR\{LOCM,LPC\}} = \text{COVAR\{LTCM,LPC\}}$$

Consequently the adjustment for S_3 depends upon the relative magnitudes of STD{LOCM} and STD{LTCM}. Let CS_3M be a sample statistic that estimates STD{LTCM}/STD{LOCM}. The S_3 statistic adjusted for multiplicative error will be denoted S_3M and is defined by

$$S_3M = \frac{\text{SCOVAR\{LOC,LPC\}}}{[\text{SSTD\{LOC\}} \times CS_3M]\ \text{SSTD\{LPC\}}}$$

the computation of CS_3M is described later in this appendix.

Under the additive model, S_3 is an estimate of

$$\frac{\text{COVAR\{LOCA,LPC\}}}{\text{STD\{LOCA\}}\ \text{STD\{LPC\}}}$$

When TCA is substituted for OCA, S_3 is an estimate of

$$\frac{\text{COVAR\{LTCA,LPC\}}}{\text{STD\{LTCA\}}\ \text{STD\{LPC\}}}$$

To examine the numerators, we rewrite the additive model as

$$\text{OCA} = \text{TCA}\left(1 + \frac{\text{ERRA}}{\text{TCA}}\right)$$

Taking natural logarithms we obtain

$$\text{LOCA} = \text{LTCA} + \ln\left(1 + \frac{\text{ERRA}}{\text{TCA}}\right)$$

As long as ERRA/TCA is small, the above expression may be approximated as

$$\text{LOCA} = \text{LTCA} + \frac{\text{ERRA}}{\text{TCA}}$$

Using the above approximate expression for the additive model, and noting that ERRA is independent of LPC, we can establish that

$$\text{COVAR\{LOCA,LPC\}} = \text{COVAR\{LTCA,LPC\}}$$

Consequently the adjustment for S_3 depends on the relative magnitudes of STD{LOCA} and STD{LTCA}. Let CS_3A be a sample statistic that estimates STD{LTCA}/STD{LOCA}. The S_3 statistic adjusted for additive error will be denoted S_3A and is defined by

$$S_3A = \frac{\text{SCOVAR\{LOC,LPC\}}}{[\text{SSTD\{LOC\}}\times CS_3A]\ \text{SSTD\{LPC\}}}$$

The computation of CS_3A is described later in this appendix.

THE TEMPORAL CORRELATION COEFFICIENT

The temporal correlation coefficient S_4 was defined as the average over all sites of the Pearson's correlation coefficient $S_3(i,\cdot)$. Consequently the adjustments for multiplicative and additive error that were employed for S_3 may be used with S_4. We define

$$S_4M = \tanh\left\{\frac{1}{I}\sum_{i=1}^{I}\tanh^{-1}\left(\frac{\text{SCOVAR}\{\text{LOC}(i,\cdot),\ \text{LPC}(i,\cdot)\}}{[\text{SSTD}\{\text{LOC}(i,\cdot)\}\ \text{x}\ CS_3M]\ \text{SSTD}\{\text{LPC}(i,\cdot)\}}\right)\right\}$$

$$S_4A = \tanh\left\{\frac{1}{I}\sum_{i=1}^{I}\tanh\left(\frac{\text{SCOVAR}\{\text{LOC}(i,\cdot),\ \text{LPC}(i,\cdot)\}}{[\text{SSTD}\{\text{LOC}(i,\cdot)\}\ \text{x}\ CS_3A]\ \text{SSTD}\{\text{LPC}(i,\cdot)\}}\right)\right\}$$

SPATIAL CORRELATION COEFFICIENT

The spatial correlation coefficient S_5 is defined as the average over all time periods of the Pearson's correlation coefficient $S_3(\cdot,j)$. Consequently the adjustments for multiplicative and additive error that were employed for S_3 may also be used for S_5. We define

$$S_5M = \tanh\left\{\frac{1}{J}\sum_{j=1}^{J}\tanh^{-1}\left(\frac{\text{SCOVAR}\{\text{LOC}(\cdot,j),\ \text{LPC}(\cdot,j)\}}{[\text{SSTD}\{\text{LOC}(\cdot,j)\ \text{x}\ CS_3M]\ \text{SSTD}\{\text{LPC}(\cdot,j)\}}\right)\right\}$$

$$S_5A = \tanh\left\{\frac{1}{J}\sum_{j=1}^{J}\tanh^{-1}\left(\frac{\text{SCOVAR}\{\text{LOC}(\cdot,j),\ \text{LPC}(\cdot,j)\}}{[\text{SSTD}\{\text{LOC}(\cdot,j)\ \text{x}\ CS_3A]\ \text{SSTD}\{\text{LPC}(\cdot,j)\}}\right)\right\}$$

THE ROOT-MEAN-SQUARE ERROR

The bias in S_6 is the amount that S_6 differs from the value that would be obtained by substituting TC for OC in the definition of S_6.

Under the multiplicative model S_6 is an estimate of STD{OCM - PC}. When TCM is substituted for OCM, the statistic S_6 is an estimate of STD{TCM - PC}. Consequently the adjustment to S_6 depends on the relative magnitudes of STD{OCM - PC} and STD{TCM - PC}. For the multiplicative model, we write

$$\text{OCM} = \text{TCM} + \text{TCM x ERRM}$$

Subtracting PC from both sides of the above equation and taking variances we find

$$\text{VAR}\{\text{OCM} - \text{PC}\} = \text{VAR}\{\text{TCM} - \text{PC}\} + \text{VAR}\{\text{TCM x ERRM}\}$$

Note that the covariance term has disappeared because ERRM has a mean of zero and is independent of TCM and PC. The variance of TCM x ERRM may be written as

$$\text{VARM}\{\text{TCM x}\} = E\{(\text{TCM})^2\}\ \text{x VAR}\{\text{ERRM}\}$$

To further evaluate this expression, we note that

$$[\text{OCM}]^2 = [\text{TCM}]^2\ (1 + 2\ \text{x ERRM} + [\text{ERRM}]^2)$$

Taking expectations we obtain

$$E\{[OCM]^2\} = E\{[TCM]^2\} \times (1 + VAR\{ERRM\})$$

Combining the above equations we obtain

$$VARM\{OCM - PC\} = VAR\{TCM - PC\} + E\{[OCM]^2\} \frac{VAR\{ERRM\}}{1 + VAR\{ERRM\}}$$

Let CS_6M be a sample estimator of

$$E\{[OCM]^2\} \frac{VAR\{ERRM\}}{1 + VARM\{ERRM\}}$$

The S_6 statistic adjusted for multiplicative error is denoted by S_6M and defined as

$$S_6M = [(S_6)^2 - CS_6M]^{1/2}$$

The computation of CS_6M is discussed later in this appendix.

For the additive model we may write

$$[OCA - PC] = [TCA - PC] + ERRA$$

Taking variances we obtain

$$VAR[OCA - PC] = VAR[TCA - PC] + VAR\{ERRA\}$$

Letting CS_6A be an estimator of VAR{ERRA}, we can define S_6A (the S_6 statistic adjusted for additive error) by

$$S_6A = [(S_6)^2 - CS_6A]^{1/2}$$

The computation of CS_6A is discussed later in this appendix.

COMPUTATION OF THE CONSTANTS REQUIRED FOR THE ADJUSTMENT FOR OBSERVATIONAL ERROR

Previous subsections of this appendix presented adjustments to the component statistics of the FOM necessary to account for observational error under multiplicative and additive models. This subsection presents methods for calculating certain constants that were previously defined in general terms only (e.g., CS_1MTE, CS_1MT, CS_1ATE, and CS_1AT), and are necessary to effect those adjustments.

Two models for observational error were presented earlier. Under the multiplicative model OCM followed an LNORM(E{LTCM}, VAR{LTCM} + VAR{ERRM}) distribution; under the additive model OCA followed an LNORM(E{LTCA}, VAR{LTCA}) + NORM(0, VAR{ERRA}) distribution. The estimation of the parameters of these distributions begins with the specification by the investigator of the error variance, based on his knowledge of the type of monitoring equipment being used in the field. For the

multiplicative model, STD{ERRM} is specified as a percentage (e.g., 3 percent) and for the additive model STD{ERRA} is specified in ppm (e.g., 2 ppm). After the standard deviation of the observational error has been specified, the estimates of the remaining parameters may be calculated. For the multiplicative model, the estimate of E{LTCM} is SMEAN{LOC} and the estimate of VAR{LTCM} is SVAR{LOC} - VAR{ERRM}. For the additive model we note that

$$E\{OCA\} = \exp[E\{LTCA\} + VAR\{LTCA\}/2]$$

$$VAR\{OCA\} = \exp[2 \times E\{LTCA\} + VAR\{LTCA\} \times (\exp[VAR\{LTCA\}] - 1) + VAR\{ERRA\} \quad .$$

Substituting SMEAN{OC} for E{OCA} and SVAR{OC} for VAR{OCA} above, we obtain a pair of equations in two unknowns. Partial simplification of those equations yields

$$SMEAN\{OC\} \triangleq \exp[E\{LTCA\} + VAR\{LTCA\}/2]$$

$$SVAR\{OC\} \triangleq (SMEAN\{OC\})^2 \times (\exp[VAR\{LTCA\}] - 1) + VAR\{ERRA\}$$

Further simplification yields our final estimates:

$$VAR\{LTCA\} \triangleq \ln\left(\frac{SVAR\{OC\} - VAR\{ERRA\}}{(SMEAN\{OC\})^2} + 1\right)$$

$$E\{LTCA\} \triangleq \ln[SMEAN\{OC\}] - VAR\{LTCA\}/2$$

In the computation of the adjustments for observational error, the estimates derived above will be used as if they were the true parameters.

Computation of the Adjustments for S_1

The adjustment constants for S_1 were denoted CS_1MTE, CS_1MT, CS_1ATE, and CS_1AT. We recommend that these constants be computed using Monte Carlo simulation. Specifically on the Kth iteration of the Monte Carlo program, we generate the following independent and identically distributed (iid) samples:

$$X1(K,1), \ldots, X1(K,N) \sim NORM(E\{LTCM\}, VAR\{LTCM\})$$

$$X2(K,1), \ldots, X2(K,N) \sim NORM(E\{LTCM\}, VAR\{LTCM\} + VAR\{ERRM\})$$

$$X3(K,1), \ldots, X3(K,N) \sim NORM(E\{LTCA\}, VAR\{LTCA\})$$

$$X4(K,1), \ldots, X4(K,N) \sim NORM(0, VAR\{ERRA\})$$

Define the transformed variables for $1 \leq j \leq N$

$$Y1(K,j) = \exp\{X1(K,j)\}$$

$$Y2(K,j) = \exp\{X2(K,j)\}$$

$$Y3(K,j) = \exp\{X3(K,j)\}$$

$$Y4(K,j) = Y3(K,j) + X4(K,j)$$

and compute

$$Z1(K) = \sum_{j=n}^{N} Y1(K,[j])$$

$$Z2(K) = \sum_{j=n}^{N} Y2(K,[j])$$

$$Z3(K) = \sum_{j=n}^{N} Y3(K,[j])$$

$$Z4(K) = \sum_{j=n}^{N} Y4(K,[j])$$

After 100 iterations of the Monte Carlo program, compute the estimates

$$CS_1MT = \sum_{K=1}^{100} Z1(K)/100$$

$$CS_1MTE = \sum_{K=1}^{100} Z2(K)/100$$

$$CS_1AT = \sum_{K=1}^{100} Z3(K)/100$$

$$CS_1ATE = \sum_{K=1}^{100} Z4(K)/100$$

$$CS_1ATE = \sum_{K=1}^{100} Z4(K)/100$$

Computation of the Adjustments for S_2

The adjustment constants for S_2 were denoted CS_2MTE, CS_2MT, CS_2ATE, and CS_2AT. The constants CS_2MTE, CS_2MT, and CS_2AT are given by the usual formulas

$$CS_2MTE = 1 - \Phi\left(\frac{\ln AQS - E\{LTCM\}}{SSTD\{LOC\}}\right)$$

$$CS_2MT = 1 - \Phi\left(\frac{\ln AQS - E\{LTCM\}}{SSTD\{LTCM\}}\right)$$

$$CS_2AT = 1 - \Phi\left(\frac{\ln AQS - E\{LTCA\}}{SSTD\{LTCA\}}\right)$$

If we let W and Y denote variables following a NORM(E{LTCA}, VAR{LTCA}), and NORM(0, VAR{ERRA}) distribution respectively, then we may write

$$CS_2ATE = PROB\{\exp(W) + Y > AQS\}$$

$$= \int PROB\{Y > AQS - \exp(w)\}\ dPROB\{W=w\}$$

$$= \int \left\{ 1 - \Phi\left(\frac{AQS - \exp(w)}{STD\{ERRA\}} \right) \right\} dPROB\{W=w\}$$

Letting $Z = (W - E\{LTCA\})/STD\{LTCA\}$ and changing variables, we obtain

$$CS_2ATE = \int \left\{ 1 - \Phi\left(\frac{AQS - \exp[z \times STD\{LTCA\} + E\{LTCA\}}{STD\{ERRA\}} \right) \right\}$$

$$\times [2\pi]^{-1/2} \exp\left(\frac{z^2}{2} \right) dz$$

For practical purposes, the limits of integration can be assumed to be ± 4.0. The constant CS_2ATE should be determined using numerical integration.

Computation of the Adjustments for S_3, S_4, and S_5

The adjustment constants for S_3, S_4, and S_5 were denoted CS_3M and CS_3A, and were defined as

$$CS_3M = STD\{LTCM\}/STD\{LOCM\}$$

$$CS_3A = STD\{LTCA\}/STD\{LOCA\}.$$

These constants may be estimated using our estimates of VAR{LTCM}, VAR{LTCA}, and VAR{LOC}.

Computation of the Adjustments for S_6

The adjustment constants for S_6 were denoted CS_6M and CS_6A, and were defined as

$$CS_6M = E\{[OCM]^2\} \frac{VAR\{ERRM\}}{1 + VAR\{ERRM\}}$$

$$CS_6A = VAR\{ERRA\}$$

These constants can be computed using our estimates of VAR{ERRM}, VAR{ERRA}, and the estimate

$$E\{[OCM]^2\} \doteq \frac{1}{N} \sum_{j=1}^{N} [OC(\cdot,j)]^2$$

APPENDIXES E, F, AND G

Appendixes E, F, and G are not published herewith but are contained under a separate binding and titled, "Development and Application of Methodology for Evaluating Highway Air Pollution Dispersion Models—Volume II: Appendixes E, F, and G." The contents of these materials are listed here for the convenience of qualified researchers. A computer program package including the documentation is available by written request to the Director, Cooperative Research Programs, Transportation Research Board, 2101 Constitution Avenue, N.W., Washington, D.C. 20418, and supplying a blank 12-in. and 8-in. diameter 9-track, 1600-BPI computer tape.

APPENDIX E

User's Guide to Programs for the Automatic Calculation of the Performance Statistics and Estimation of Sample Size

Table of Contents

APPENDIX F

User's Manual for the Pollution Dispersion Model Diagnostic Evaluation Program

Table of Contents

APPENDIX G

User's Guide to the Data Base

Table of Contents

THE TRANSPORTATION RESEARCH BOARD is an agency of the National Research Council, which serves the National Academy of Sciences and the National Academy of Engineering. The Board's purpose is to stimulate research concerning the nature and performance of transportation systems, to disseminate information that the research produces, and to encourage the application of appropriate research findings. The Board's program is carried out by more than 250 committees, task forces, and panels composed of more than 3,100 administrators, engineers, social scientists, attorneys, educators, and others concerned with transportation; they serve without compensation. The program is supported by state transportation and highway departments, the modal administrations of the U.S. Department of Transportation, the Association of American Railroads, and other organizations and individuals interested in the development of transportation.

The Transportation Research Board operates within the Commission on Sociotechnical Systems of the National Research Council. The National Research Council was established by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and of advising the Federal Government. The Council operates in accordance with general policies determined by the Academy under the authority of its congressional charter of 1863, which establishes the Academy as a private, nonprofit, self-governing membership corporation. The Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in the conduct of their services to the government, the public, and the scientific and engineering communities. It is administered jointly by both Academies and the Institute of Medicine.

The National Academy of Sciences was established in 1863 by Act of Congress as a private, nonprofit, self-governing membership corporation for the furtherance of science and technology, required to advise the Federal Government upon request within its fields of competence. Under its corporate charter the Academy established the National Research Council in 1916, the National Academy of Engineering in 1964, and the Institute of Medicine in 1970.

TRANSPORTATION RESEARCH BOARD
National Research Council
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

ADDRESS CORRECTION REQUESTED