UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711

DATE: 7/30/81

SUBJECT: Interim Procedures for Evaluating Air Quality Models

FROM: Joseph A. Tikvart, Chief J. Tikvart
Source Receptor Analysis Branch (MD-14)

TO: Chief, Air Programs Branch, Regions I - X

Attached is a report entitled "Interim Procedures for Evaluating Air Quality Models." The purpose of the report is to provide a general framework for the quantitative evaluation and comparison of air quality models. It is intended to help you decide whether a proposed model, not specifically recommended in the Guideline on Air Quality Models, is acceptable on a case-by-case basis for specific regulatory application. The need for such a report is identified in Section 7 of "Regional Workshops on Air Quality Modeling: A Summary Report."

An earlier draft (Guideline for Evaluation of Air Quality Models) was provided to you for comment in January 1981. We received comments from four Regional Offices and have incorporated many of the suggestions. These comments reflected a diversity of opinion on how rigid the procedures and criteria should be for demonstrating the acceptability of a nonguideline model. One Region maintained that EPA should establish minimum acceptable requirements on data bases, decision rationale, etc. Others felt that we should be more flexible in our approach. This report defines the steps that should be followed in evaluating a model but leaves room for considerable flexibility in details for each step.

The procedures and criteria presented in this new report are considered interim. They are an extension of recommendations resulting from the Woods Hole Workshop in Dispersion Model Performance held in Setpember 1980. That workshop was sponsored under a cooperative agreement between EPA and the American Meteorological Society. Thus, while some of the performance evaluation procedures may be resource intensive, they reflect most of the requirements identified by an appropriate scientific peer group. However, since the concepts are relatively new and untested, problems may be encountered in their initial application. Thus, the report provides suggested procedures; it is not a "guideline."

We recommend that you begin using the procedures on actual situations within the context of the caveats expressed in the Preface and in Section 5.3. Where suggestions are inappropriate, the use of alternative techniques to accomplish the desired goals is encouraged. Feedback on your experience and problems are important to us. After a period of time during which experience is gained and problems are identified, the report will be updated and guidance will gradually evolve. Questions on the use of the procedures and feedback on your experiences with their application should be directed to the Model Clearinghouse (Dean Wilson, 629-5681). An example of the procedures applied to a real data base is being developed under contract and should be completed in early 1982.

Attachment

cc: Regional Modeling Contacts, Region I - X
W. Barber
D. Fox
T. Helms
W. Keith
M. Muirhead
L. Niemeyer
R. Smith
F. White

INTERIM PROCEDURES FOR EVALUATING

AIR QUALITY MODELS

August 1981

United States Environmental Protection Agency
Office of Air, Noise and Radiation
Office of Air Quality Planning and Standards
Source Receptor Analysis Branch
Research Triangle Park, North Carolina 27711

Preface

The quantitative evaluation and comparison of models for application to specific air pollution problems is a relatively new problem area for the modeling community. It is expected that initially there will be a number of problems in this evaluation and comparison. Also several projects are underway that will subsequently provide better insight to the model evaluation problem and its limitations. Thus, procedures discussed in this document are considered to be interim.

Where material presented is inappropriate, the use of alternative techniques to accomplish the desired goals is encouraged. EPA Regional Offices and State air pollution control agencies are encouraged to use this information to judge the appropriateness of a proposed model for a specific application, but still must exercise judgment where specific recommendations are not of practical value. After a period of time during which experience is gained, problem areas will be identified and addressed in revisions to this document.

The procedures described herein are specifically tailored to operational evaluation, as opposed to scientific evaluation. The main goal of operational evaluation is to determine whether a proposed model is appropriate for use in regulatory decision making. The ability of various sub-modules (plume rise, etc.) to accurately reproduce reality or to add basic knowledge assessed by scientific evaluation is not specifically addressed by these procedures.

An example illustrating the procedures described in this document is currently being prepared and should be available in early 1982.

TABLE OF CONTENTS

Page

Summary

This document describes interim procedures for use in accepting, for a specific application, a model that is not specifically identified in the Guideline on Air Quality Models[1]. The primary basis for the model evaluation assumes the existence of a reference model which has some pre-existing status and to which the proposed nonguideline model can be compared from a number of perspectives. However for some applications if may not be possible to identify an appropriate reference model, in which case specific standards for model acceptance must be identified. Figure 1 provides an outline of the procedures described in this document.

After analysis of the intended application or the problem to be modeled, a decision is made on the reference model to which the proposed model could be compared. If an appropriate reference model can be identified, then the relative acceptability of the two models is determined as follows. The model is first compared on a technical basis to the reference model to determine if it would be expected to more accurately estimate the true concentrations. Next a protocol for model performance comparison is written. This protocol describes how an appropriate set of field data will be used to judge the relative performance of the proposed and the reference model. Performance measures recommended by the American Meteorological Society[2] are used to describe the comparative performance of the two models in an objective scheme. That scheme considers the relative importance to the problem of various modeling objectives and the degree to which the individual performance measures

support those objectives. Once the plan for performance evaluation is written and the data to be used are collected/assembled, the performance measure statistics are calculated and the weighting scheme described in the protocol is executed. Execution of the decision scheme will lead to a determination that the proposed model performs better, worse or about the same as the reference model for the given applications. The results of the technical and performance evaluations are considered together to determine the overall acceptability of the proposed model.

If no appropriate reference model is identified, the proposed model is evaluated as follows. First the proposed model is evaluated from a technical standpoint to determine if it is well founded in theory, and is applicable to the situation. This involves a careful analysis of the model features and intended usage in comparison with the source configuration, terrain and other aspects of the intended application. Secondly, if the model is considered applicable to the problem, it is examined to see if the basic formulations and assumptions are sound and appropriate for the problem. If the model is clearly not applicable or cannot be technically supported, it is recommended that no further evaluation of the model be conducted and that the exercise be terminated. Next, a performance protocol is prepared that specifies certain criteria that should be met. Data collection and execution of the performance protocol will lead to a determination that the model is acceptable or unacceptable. Finally results for the performance evaluation should be considered together with the results of the technical evaluation to determine the acceptability.

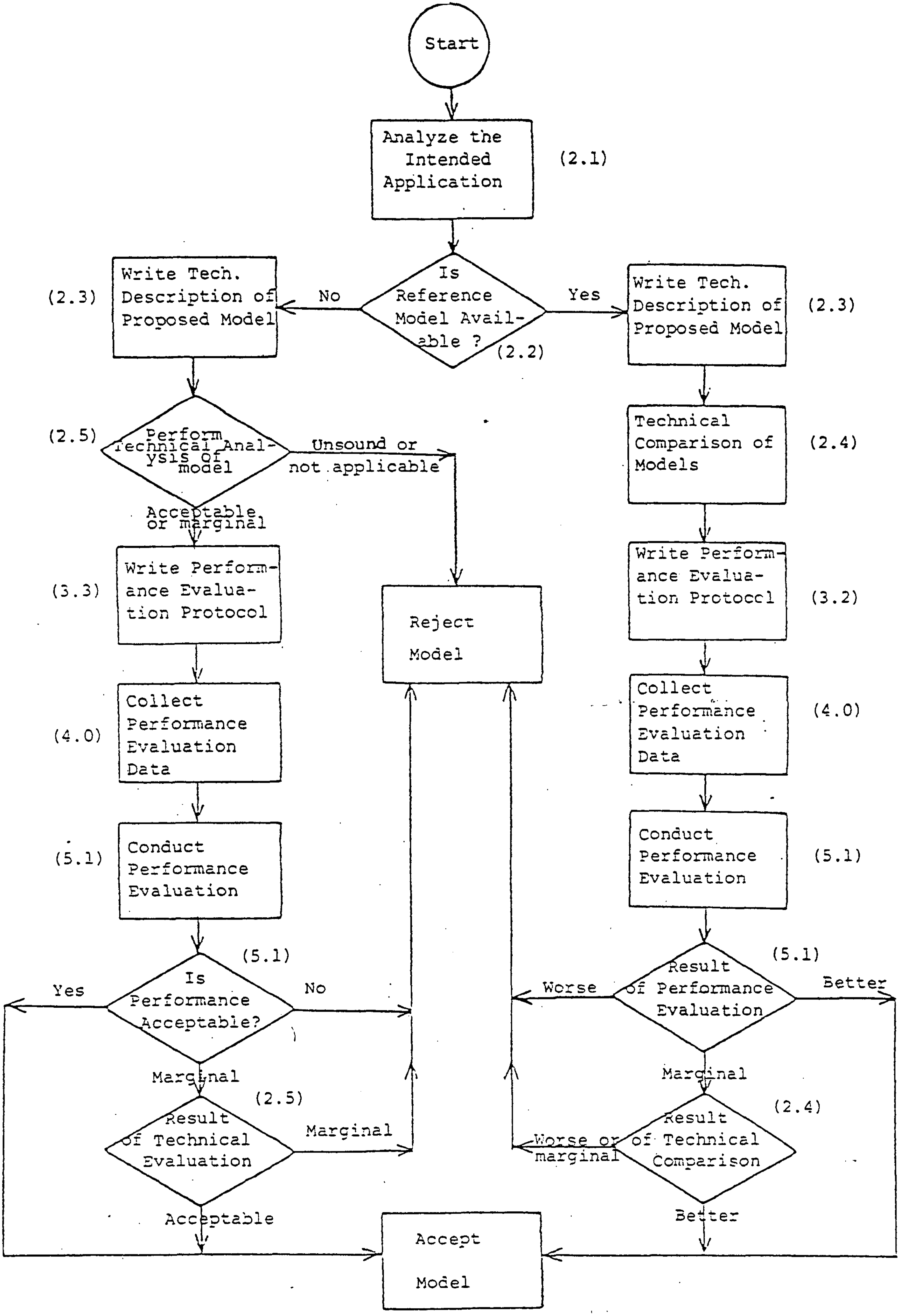
Start
Analyze the Intended Application
(2.1)
Is Reference Model Available ?
(2.2)
No
Yes
Write Tech. Description of Proposed Model
(2.3)
Write Tech. Description of Proposed Model
(2.3)
Perform Technical Analysis of model
(2.5)
Unsound or not applicable
Acceptable or marginal
Technical Comparison of Models
(2.4)
Write Performance Evaluation Protocol
(3.3)
Reject Model
Write Performance Evaluation Protocol
(3.2)
Collect Performance Evaluation Data
(4.0)
Collect Performance Evaluation Data
(4.0)
Conduct Performance Evaluation
(5.1)
Conduct Performance Evaluation
(5.1)
Is Performance Acceptable?
(5.1)
Yes
No
Result of Performance Evaluation
(5.1)
Worse
Better
Marginal
Result of Technical Evaluation
(2.5)
Marginal
Acceptable
Marginal
Result of Technical Comparison
(2.4)
Worse or marginal
Better
Accept Model

Figure 1. Decision Flow Diagram for Evaluating a Proposed Air Quality Model. (Applicable Sections of the Document are indicated in Parentheses.)

INTERIM PROCEDURES FOR EVALUATING AIR QUALITY MODELS

1.0 INTRODUCTION

This document describes interim procedures that can be used in judging whether a model, not specifically recommended for use in the Guideline on Air Quality Models[1] is acceptable for a given regulatory action. It identifies the documentation, model evaluation and data analyses desirable for establishing the appropriateness of a proposed model.

This document is only intended to assist in determining the acceptability of a proposed model for a specific application (on a case-by-case basis). It is not for use in determining whether a new model could be acceptable for general use and/or should be included in the Guideline on Air Quality Models. This document also does not address criteria for determining the adequacy of alternative data bases to be used in models, except in the case where a nonguideline model requires the use of a unique data base. The criteria or procedures generally applicable to the review of fluid modeling procedures are contained elsewhere. [3,4,5]

The remaining sections provide the following. Section 1.1 describes the history and the need for a consistent set of evaluation procedures, Section 1.2 provides the basis for performing the evaluation, and Section 1.3 suggests how the task of model evaluation should be coordinated between the applicant and the control agency. Section 2 describes the technical information needed to define the regulatory problem and the choice of the reference and proposed models. Section 2 also contains a suggested method of analysis to determine the applicability of the proposed model to the situation. Section 3 discusses the protocol to be used in judging the performance of the proposed model. Section 4 describes the design of the data base for the performance evaluation. Section 5 describes the execution of the performance evaluation and provides guidance for combining these results with other criteria to judge the overall acceptability of the proposed model. Appendix A provides a reviewer's checklist which can be used by the appropriate control agency in determining the acceptability of the applicant's evaluation. Appendix B describes the calculation of performance measures and related parameters. Appendix C is a summary of the Woods Hole Workshop on Dispersion Model Performance, sponsored by the American Meteorological Society.[2]

1.1 Need for Model Evaluation Procedures

The Guideline on Air Quality Models makes specific recommendations concerning air quality models and the data bases to be used with these models. The recommended models should be used in all evaluations relative to State Implementations Plans (SIPs) and Prevention of Significant Deterioration (PSD) unless it is found that the recommended model is inappropriate for a particular application and/or a more appropriate model or analytical procedure is available. However, for some applications the guideline does not recommend specific models and the appropriate model must be chosen on a case-by-case basis. Similarly, the recommended data bases should be used unless such data bases are unavailable or inappropriate. In these cases, the guideline states that other models and/or data bases deemed appropriate by the EPA Regional Administrator may be used.

Models are used to determine the air quality impact of both new and existing sources. The majority of cases where nonguideline models have been proposed in recent years have involved the review of new sources especially in connection with prevention of significant deterioration permit applications. However, most Regional Offices have received proposals to use nonguideline models for SIP relaxations and for general area-wide control strategies. Prior to 1977, many large scale control strategies involved the use of models not currently recommended in the Guideline on Air Quality Models. Such applications were frequently accepted. Nonguideline techniques have also been applied to large scale control strategies since 1977. In the Northeast and North Central U. S. where there are wide areas of nonattainment or marginal attainment of standards, nonguideline models are frequently proposed for use which would allow increased emissions from large point sources. In "cleaner" areas of the South and West, nonguideline models are also frequently proposed for new or modified point sources.

Many of the proposals to use nonguideline models have involved modeling of point sources in complex terrain and/or a shoreline environment. Other applications have included modeling point sources of photochemical pollutants, modeling in extreme environments (artic/tropics/deserts), modeling of fugitive emissions and modeling of open burning/field burning where smoke management (a form of intermittent control) is practiced. For these applications a refined approach is not recommended in the Guideline on Air Quality Models. Also a relatively small number of proposals involved applications where a recommended model was appropriate, but another model was judged preferable.

The types of nonguideline models proposed have included: (1) minor modification of computer codes to allow a different configuration/number of sources and receptors that essentially do not change the estimates from those of the basic model; (2) modifications of basic components in recommended models, e.g., different dispersion coefficients (measured or estimated), wind profiles, averaging times, etc; and (3) completely new models that frequently involve non-Gaussian approaches and/or phenomenological modeling (temporal/spatial modeling of the wind flow field or other meteorological inputs).

The Guideline on Air Quality Models, while allowing for the use of alternative models in specific situations, does not provide a technical basis for deciding on the acceptability of such techniques. To assure a more equitable approach in dealing with sources of pollution in all sections of the country it is important that both the regulatory agencies and the entire modeling community strive toward a consistent approach in judging the adequacy of techniques used to estimate concentrations in the ambient air. The Clean Air Act[6] recognized this goal and states that the "Administrator shall specify with reasonable particularity each air quality model or models to be used under specified sets of conditions . . . "

The use of a consistent set of procedures to determine the acceptability of nonguideline models should also serve to better ensure that the state-of-the-science is reflected. A properly constructed set of evaluation criteria should not only serve to promote consistency, but should better serve to ensure that the best technique is applied. It should be noted that a proposed model cannot be proprietary since it may be subject to public examination and could be the focus of a public hearing or other legal proceeding.

1.2 Basis for Evaluation of Models

The primary basis for accepting a proposed model for a specific application, as described in this document, involves a technical comparison and a comparison of performance between the proposed model and an applicable reference model. Under this scheme the greatest emphasis is placed on the performance evaluation. The proposed model would be acceptable for regulatory application if its performance is clearly better than that of the reference model. It should not be applied to the problem if its performance were clearly inferior to that of the reference model. When the performance evaluation is inconclusive or marginal one could decide in favor of the proposed model if it were found to be technically better than the reference model.

A secondary basis for accepting or rejecting a proposed model could involve the use of performance criteria written specifically for the intended application. While this procedure is not encouraged because of lack of experience in writing such criteria and the necessity of considerable subjectivity, it is recognized that in some situations it may not be possible to specify an appropriate reference model. Such a scheme would insure that the proposed model is technically sound and applicable to the problem, or at least marginally so, and that it pass certain performance requirements that are acceptable to all parties involved. Marginal performance together with a marginal determination on technical acceptability would suggest that the model should not be used.

At the present time one cannot set down a complete set of objective evaluation criteria and standards for acceptance of models using these concepts. Bases for such objective criteria are lacking in a number of areas, including a consistent set of standards for model performance, scientific consensus on the nature of certain flow phenomena such as interactions with complex terrain, etc. However, this document provides the framework for inclusion of future technical criteria as well as specifying currently available criteria.

1.3 Coordination with Control Agency

The general philosophy of this document is that the applicant or the developer of the model should perform the analysis. Depending on the complexity/sensitivity of the application and the level of uncertainty in the applicant's analysis, the reviewing agency should review this analysis and make a judgment on the findings, perform independent checks on certain aspects of the analysis, and/or perform an independent analysis. The reviewing agency must have access to all of the basic information that went into the analysis.

To avoid costly and time-consuming delays in execution of the model evaluation, the applicant is strongly urged to maintain close liaison with the reviewing agency(s) throughout the project. It is important that agreement be reached up-front on the choice of a reference model. It is especially important that meetings be held at the completion of the technical evaluation and before the initiation of the field data collection phase. At that time the reviewing agency can make a determination on the applicability of the proposed model (See Section 2) and the design of (or choice of) the data base network to be used in the performance evaluation. It is also important at that time to agree on the protocol and criteria for comparing the proposed and the reference models, including precise measures of model performance such as bias, precision, statistical significance levels, etc.

2.0 TECHNICAL EVALUATION

The technical evaluation consists of a determination of the appropriateness of the proposed model for the intended application, exclusive of the performance evaluation. To adequately address the technical evaluation requires a thorough understanding of the source-receptor relationships which must be addressed by the proposed model in the intended application, selection of an appropriate reference model and a technical comparison of the proposed model with the reference model. If no appropriate reference model can be identified, an in-depth technical investigation of the theory, operating characteristics and applicability of the proposed model should be undertaken. The following subsections describe these needs in more detail.

2.1 Intended Application

Information that needs to be assembled on the intended application includes a complete description of the source or sources to be modeled, e.g., the configuration of the sources, location and heights of stacks, stack parameters (flow rates and gas temperature) and location of any fugitive sources to be included. Appropriate* emission rates for each averaging time corresponding to ambient air quality standards for each pollutant should be used. In the case of complex industrial sources it is also generally necessary to obtain a plant layout including dimensions of plant buildings and other nearby buildings/obstacles. Mobile and area source emissions should be assembled in the format (i.e., line source segments, grid squares, etc.) to be used in the model.

* Section 4.1 in the Guideline on Air Quality Models discusses emission rates appropriate for use in regulatory modeling.

It it also generally necessary to have a topographic map or maps which cover the modeling area. If the topographic maps do not include the location of emission sources, monitors, instrumented towers, etc., a separate map with this information should be supplied. The areal coverage is sometimes predetermined by political jurisdiction boundaries, i.e., an air quality control region. More often, however, modeling is confined to the region where any significant threat to the standards or PSD increments is likely to exist. In these cases it is desirable to make crude determinations of the area to be considered and at the same time to tentatively determine the location of critical receptors for each pollutant where standards/increments are most likely to be threatened. The recommended approach for making these determinations is to make preliminary estimates of the concentration field using available models and available data. A preliminary estimate would utilize the appropriate emission rates for the regulatory problem and whatever representative meteorological data are available before the evaluation*.

It is recommended that two or three separate preliminary estimates of the concentration field be made. The first set of estimates could be made with the screening techniques mentioned or referenced in the Guideline on Air Quality Models. The second set of estimates would be done with the proposed model and the third set with the reference model (Section 2.2). Estimates for all averaging times should be calculated.

* A final set of model estimates, to be used in decision making, could utilize additional data collected during the performance evaluation as input to the appropriate model.

The three sets of estimates not only serve to define the modeling domain and critical receptors but also aid in determining the applicability of the proposed model (Sections 2.4 and 2.5) and the design of the performance evaluation data network (Section 4.0).

2.2 Reference Model

The primary approach used in this document to judge the acceptability of a proposed model relies on the philosophy that if the model is technically better and performs better than the recommended model or the model that has historically been applied to the situation, then the proposed model should be considered for use. In Section 2.4 procedures contained in the Workbook for Comparison of Air Quality Models[7] are used to the maximum extent possible, to make the technical comparison. Sections 3 and 4 describe procedures for comparing the performance of the "reference" model with that of the proposed model.

The first choice for a reference model should be the refined models recommended in the Guideline on Air Quality models and listed in Appendix A to that Guideline. However, not all modeling situations are covered by recommended models. For example, models for point sources of reactive pollutants or shoreline fumigation problems are not included. In these cases the applicant and the reviewing agency should attempt to agree on an appropriate and technically defensible reference model, based on the current technical literature and on past experience. Major considerations in the selection of the reference model under these circumstances are that it is applicable to the type of problem in question, has been described in published reports or the open literature, and is capable of producing concentration estimates for all averaging times for which a performance measure statistic must be calculated (usually one hour and the averaging times associated with the standards/increments). This latter requirement precludes the use of screening techniques which rely on assumed meteorological conditions for a worst case.

Where it is clearly not possible to specify a reference model, the proposed model must "stand alone" in the evaluation. In such cases the technical justification and the performance evaluation necessary to determine acceptability would have to be more substantial. Section 2.5 discusses a suggested rationale for determining if the model is technically justified for use in the application. Section 3.3 discusses some considerations in designing the performance evaluation protocol when no reference model comparison is involved.

2.3 Proposed Model

The model proposed for use in the intended application must be capable of estimating concentrations corresponding to the regulatory requirements of the problem as identified in Section 2.1. In order to conduct the performance evaluation the model should be capable of sequentially estimating hourly concentrations, and concentrations for all averaging times within the area of interest based on meteorological and emission inputs.

A complete technical description of the model is needed for the analysis in Section 2.4 or Section 2.5. This technical description should include a discussion of the features of the proposed model, the types of modeling problems for which the model would be applicable, the mathematical relationships involved and their bases, and the assumptions and limitations of the model. The model description should take the form of a report or user manual that completely describes its operation. Published articles which describe the model are useful. If the model has been applied to other problems, a review of these applications should also be undertaken. For models designed to handle complex terrain, land/water interfaces and/or other special situations, the technical description should focus on how the model treats these special factors. To the maximum extent possible, evidence for the validity of the methodologies should be included.

2.4 Comparison with the Reference Model

When an appropriate reference model can be identified it should be determined whether the proposed model is better to use than the reference model. The goal is to determine if the model can be *expected* to more accurately reproduce the actual concentrations caused by the subject source(s), with emphasis on dispersion conditions and subareas of the modeling domain that are most germane to the regulatory aspects of the problem (Section 2.1). The procedures described in the Workbook for Comparison of Air Quality Models are appropriate for this determination. This Workbook contains a procedure whereby a proposed model is qualitatively compared, on technical grounds to the reference model, taking into account the intended use of the two models and the specific application.

The Workbook procedure is application-specific; that is, the results depend upon the specific situation to be modeled. The reference model serves as a standard of comparison against which the user gages the proposed model being evaluated. The way in which the proposed model treats twelve aspects of atmospheric dispersion called "application elements," is determined. These application elements represent physical and chemical phenomena that govern atmospheric pollutant concentrations and include such aspects as horizontal and vertical dispersion, emission rate, and chemical reactions. The importance of each element to the application is defined in terms of an "importance rating." Tables giving the importance ratings for each element are provided in the Workbook, although they may be modified under some circumstances. The heart of the procedure involves an element-by-element comparison of the way in which each element is treated by the two models. These individual comparisons, together with the importance ratings for each element in the given application, form the basis upon which the final comparative evaluation of the two models is made.

It is especially important that the user understand the physical phenomena involved, because the comparison of two models with respect to the way that they treat these phenomena is basic to the procedure. Sufficient information is provided in the Workbook to permit these comparisons. Expert advice may be required in some circumstances. If alternate procedures are used to complete the technical comparison of models, they should be negotiated with the reviewing agency.

The results of the comparison of the proposed model with the reference model should indicate whether the proposed model is better, comparable or worse than the reference model. This information is used in the overall model evaluation in Section 5.

2.5 Technical Evaluation When No Reference Model Is Used

If it is not possible to identify an appropriate reference model (Section 2.2), then the procedures of Section 2.4 cannot be used and the proposed model must be technically evaluated on its own merits. The technical analysis of the proposed model should attempt to qualitatively answer the following questions:

1. Are the formulations and internal constructs of the model well founded in theory?

2. Does the theory fit the practical aspects and constraints of the problem?

To determine whether or not the underlying assumptions have been correctly and completely stated requires an examination of the basic theory employed by the model. The technical description of the model discussed in Section 2.3 should provide the primary basis for this examination. The examination of the model should be divided into several subparts that address various aspects of the formulation. For example, for some models it might be logical to separately examine the methodologies used to characterize the emissions, the transport, the diffusion, the plume rise, and the chemistry. For each of these model elements it should be determined whether the formulations are based on sound scientific, engineering and meteorological principles and whether all aspects of each element are considered. Unsound or incomplete specification of assumptions should be flagged for consideration of their importance to the actual modeling problem.

For some models, e.g., those that entail a modification to a model recommended in the Guideline on Air Quality Models or to the reference model, the entire model would not need to be examined for scientific credibility. In such cases only the submodel or modification should be examined. Where the phenomenological formulations are familiar and have been used before, support for their scientific credibility can be cited from the literature.

For models that are relatively new or utilize a novel approach to some of the phenomenological formulations, an in-depth examination of the theory should be undertaken. The scientific support for such models should be established and reviewed by those individuals who have broad expertise in the modeling science and who have some familiarity with the approach and phenomena to be modeled.

To determine how well the model fits the specific application, the model assumptions should be compared to the reality of the application. The assumptions involved in the methodologies proposed to handle each phenomenon should be examined to see if they are reasonable for the given situation. Particular attention should be paid to flagged assumptions which may either be only marginally valid from a basic standpoint or be implicit, and unstated to determine whether such assumptions are germane to the situation. For assumptions that are not met, it should be established that these deficiencies will not cause any significant differences in the estimated concentrations. The most desirable approach takes the form of sensitivity testing by the applicant where variations are made on the questionable assumptions within the model to determine whether or not these assumptions are indeed critical. Such an exercise should be conducted if possible and would involve obtaining model estimates before and after modification of formulas or data to reflect alternate assumptions. However, in many cases this exercise may be too resource-consumptive and the proof of model validity should still rest with the performance evalution described in Section 4.

Execution of the procedures in this section should lead to a judgment on whether the proposed model is applicable to the problem and can be scientifically supported. If these criteria are met, the model can be designated as appropriate and should be applied if its field performance (Section 4) is acceptable. When a model cannot be supported for use based on this technical evaluation, it should be rejected. When it is found that the model could be appropriate, but there are questionable assumptions, then the model can be designated as marginal and carried forward through the performance evaluation.

3.0 PROTOCOL FOR PERFORMANCE EVALUATION

The results of air quality simulation models are used in the process of setting emission limits, determining the suitability of proposed new source sites, etc. The goal of model performance evaluation is to determine the degree of confidence which should be placed in these results. To achieve this goal, model concentration estimates are compared with observed concentrations in a variety of ways. The primary methods of comparison produce statistical information and constitute statistical performance evaluation. However, statistical performance evaluation should be supplemented by additional qualitative analysis (case studies) and interpretation to ensure that the model realistically simulates the physical processes for which it was designed.

This section describes a process for evaluating the performance of the proposed model and determining whether that performance is adequate for the specific application. It describes specific statistical measures which should be used to characterize the performance of the model. The process requires that a protocol be prepared for comparing the performance of the reference and proposed model and describes a scheme to weigh the relative performance of each model according to the significance with which one model outperforms the other and in terms of the importance of each performance category. Some guidance is provided on how to evaluate model performance when comparison with a reference model is not possible.

Model performance should be evaluated for each of the averaging times specified in the appropriate regulations. In addition, performance for models whose basic averaging time is shorter than the regulatory averaging time must also be evaluated for that shorter period. Thus, for example, a model may calculate one-hour concentrations for SO_2 and determine concentrations for longer averaging periods from these one-hour averages. Performance of this model would then be evaluated separately for one, three, and 24-hour averages and, if appropriate, for the annual mean.

The performance evaluation measures and procedures result in part from the recommendations of the AMS Workshop on Dispersion Model Performance. Appendix C presents a summary of the Workshop recommendations.

3.1 Performance Measures

Performance measures may be classified as magnitude of difference measures and correlation or association measures. Magnitude of difference measures present a quantitative estimate of the discrepancy between measured concentrations and concentrations estimated by a model at the monitoring sites. Correlation measures quantitatively delineate the degree of association between estimations and observations. The quantitative measures should be supplemented by informative graphical techniques and interpretations such as histograms, isopleth analyses, scatter diagrams and the like. This subsection discusses the recommended performance measures and analyses.

Magnitude of difference performance measures compare estimated and observed concentrations through analysis of the model residual, d, defined as the difference between observed and estimated concentrations. (See Appendix B for more complete discussion of the performance measures.) The model residual measures the amount of model underestimation. The relative residual, i.e., the percent underestimation by the model, should be calculated as supplementary information. The relative residual provides information more readily communicated to those with nontechnical backgrounds.

The model residuals are analyzed to provide values for the following aspects of model performance; accuracy of the prediction of peak concentrations, average model bias, model precision and model gross variability.

3.1.1 Accuracy of Peak Prediction

The accuracy of the peak predictions should be evaluated and reported to conform with the somewhat conflicting requirements of evaluations responsible to regulatory standards or increments and those responsible to the needs of statistical reliability. Therefore, the performance measures to evaluate the accuracy of the peak predictions consist of the set of residuals D_n, paired in various combinations of space and time which measure the amount of underestimation of the nth highest estimation and the more complete analysis of the set of residuals for the highest 5% of the observations or for the highest 25 observations, whichever is greater.

Observed and estimated peak concentrations can be paired in space and time in the four ways listed in Table 3.1. Each measure in Table 3.1 should be calculated for each short-term averaging period specified in regulations in addition to the one-hour averaging time. (The appropriate relative residual set should also be calculated.) Thus, for example, the residuals D_2 may be required for a problem involving possible violations of the three-hour NAAQS for SO_2 where the highest, second-highest concentrations are at issue.

The accuracy of the highest or second-highest estimate is, however, difficult to evaluate statistically. Statistical evaluations have greater meaning when applied to a larger number of values than to one or two extremes. Therefore, the set of residuals D_n, where n extends over the top 5% of the observed concentrations, is evaluated for the properties of model bias, and model precision as discussed in Sections 3.2.1 and 3.1.3. If there are fewer than 500 observations, then n extends over the top 25 observations.

Table 3.1 Residuals to Measure Accuracy of Peak Prediction

Paired in	Residual Set
Space & Time	$D_n (L_n, T_n) = C_o (L_n, T_n) - C_p (L_n, T_n)$
Space not Time	$D_n (L_n, T) = C_o (L_n, T_n) - C_p (L_n, T_j)$
Time not Space	$D_n (L, T_n) = C_o (L_n, T_n) - C_p (L_j, T_n)$
Unpaired	$D_n (L, T) = C_o (L_n, T_n) - C_p (L_r, T_m)$

L_n = monitor site for nth highest observed concentration.

T_n = time of nth highest observed concentration.

T_j = time of nth highest estimated concentration at site Ln.

L_j = site of nth highest estimated concentration during time T_n.

$C_p (L_r, T_m)$ = nth highest estimated concentration.

$C_o (L_n, T_n)$ = nth highest observed concentration.

(L_r, T_m) = site and time of highest estimated concentration (Generally, $L_r \neq L_n$ and $T_m \neq T_n$).

Since statistical analysis cannot supply all desired information concerning model performance, supplementary case studies should be included which examine whether the model is able to replicate a number of the peak concentrations. The following analyses are suggested:

(1) Measured and calculated concentrations and patterns are compared for those periods corresponding to the highest 25 observed values. The case study should include consideration of the meteorological conditions associated with the events and should consider averaging times of one hour as well as for averaging times important to regulatory standards.

(2) For any critical monitoring location, compare the meteorological conditions such as stability and wind speed class, producing the highest 25 measured and calculated concentrations. The number and type of meteorological conditions will be determined by the model input parameters.

The case study approach can identify problems with a model which might not be so readily apparent from a statistical performance measure. For example, if high measured concentrations occur during a period for which the model estimated zero values everywhere, then the treatment of mixing height penetration by a plume or the value of σ_z may be wrong. Similarly, if most of the highest concentration measurements occur with slightly unstable conditions, while the highest concentration estimates occur with very unstable conditions then either the method used for assigning stability or the choice of dispersion curves associated with different stabilities may be in error. The results of these case studies may indicate the physical reasons for poor performance values of any of the measures listed in Table 3.1. The degree of interpretation and conclusions to be derived from these analyses depend on the confidence placed on the accuracy and representativeness of the model input data. If data from tracer networks are available, the case studies should include analysis of those periods with meteorological conditions of poor dispersion.

3.1.2 Average Model Bias

Model bias is measured by the value of the model residual averaged over an appropriate range of values. Large over and underestimations may cancel in computing this average. Supplementary information concerning the distribution of residuals should therefore be supplied. This supplementary information consists of confidence intervals about the mean value, calculated according to the methods presented in Appendix B and histograms or frequency distributions of model residuals.

For certain applications, especially cases in which the candidate model is designed to simulate concentrations occurring during important meteorological processes, it can be important to estimate model bias under different meteorological conditions. Data disaggregation must compromise between the desired goals of defining a large enough number of meteorological categories to cover a wide range of conditions and having a sufficient number of observations in each category to calculate statistically meaningful values. For example, it may be appropriate to stratify data by lumped stability classes, unstable (A-C), neutral (D) and stable (E-F) rather than by individual classes A, B, C, D, E, and F.

3.1.3 Model Precision

Model precision refers to the average amount by which estimated and observed concentrations differ as measured by residuals with no algebraic sign. While large positive and negative residuals can cancel when model bias is calculated, the unsigned residuals comprising the precision measures do not cancel and thus provide an estimate of the error scatter about some reference point. This reference point can be the mean error or the desired value of zero. Two types of precision measure are the noise, which delineates the error scatter about the mean error, and the gross variability, which delineates the error scatter about zero error.

The performance measure for noise is either the variance of the residuals, S_d^2, or the standard deviation of the residuals, S_d. The performance measure for gross variability is the mean square error, or the root mean square error. An alternate performance measure for the gross variability is the mean absolute residual, $\overline{|d|}$. The mean absolute residual is statistically more robust than the root-mean-square-error; that is, it is less affected by removal of a few extreme values.

Supplementary analyses for model precision should include tables or histograms of the distribution of performance measures and computation of these measures for the same meteorological categories discussed in Section 3.1.2.

3.1.4 Correlation Analyses

Correlation analyses involve calculating parameters resulting from linear least squares regression and presenting associated graphical analyses and their interpretation. The numerical results constitute quantitative measures of the association between estimated and observed concentrations. The graphical analyses constitute supplementary qualitative measures of the same information. There are three types of correlation analysis and temporal analysis.

Coupled space-time correlation analysis involves computing the Pearson's correlation coefficient, r, and parameters, a and b, of the linear least squares regression equation. A scattergram of the C_o (L, T), C_p(L, T) data pairs is supplementary information which should be presented.

Spatial correlation analysis involves calculating the spatial correlation coefficient and presenting isopleth analyses of the estimated and observed concentrations for particular periods of interest. The spatial coefficient measures the degree of spatial alignment between the estimated and observed concentrations. The method of calculation involves computing the Pearson's correlation coefficient for each time period and determining an average over all time periods. Specifics are discussed in Appendix B.

Estimates of the spatial correlation coefficient for single source models are most reliable for calculations based on data intensive tracer networks. Isopleths of the distributions of estimated and observed concentrations for periods of interest should be presented and discussed.

Temporal correlation analysis involves calculating the temporal correlation coefficient and presenting time series of observed and estimated concentrations or of the model residual for each monitoring location. The temporal correlation coefficient measures the degree of temporal alignment between observed and estimated concentrations. The method of calculation is similar to that for the spatial correlation coefficient. Time series of C_o and C_p or of model residuals should be presented and discussed for each monitoring location.

3.2 Protocol for Model Comparison

The model performance measures described in Section 3.1 are appropriate for most regulatory applications where the relative performance of two competing air quality models is to be evaluated. Each performance measure, when calculated for the proposed model and the reference model, provides certain statistics, or in some cases somewhat more qualitative measures, which can be used to discriminate between the capabilities of the two models to reproduce the measured concentration.

The objective scheme for considering the relative importance of each performance measure and significance of the difference in performance of the two models is called the model comparison protocol. This section discusses the factors to be considered in establishing such a protocol for an individual performance evaluation. Lack of experience with performance evaluations prevents writing sets of objective protocols to cover all types of problems. Rather, a specific protocol needs to be written for each performance evaluation. The objective of the protocol is to establish objective weights for each performance measure and for the degree of intermodel difference. It is very important that such a protocol be written before the data base is selected or collected and before any performance measures are calculated so as not to bias the final outcome.

The model comparison protocol basically addresses two questions: (1) What relative importance should each performance measure hold in the final decision scheme? For example, would model bias be a more important factor than gross variability or good spatial correlation? Or, for example, is accurate prediction of the magnitude of the peak concentration more important than accurate prediction of the location of that peak? Answers to these questions may vary according to the application. (2) What consideration should be given to the degree of difference in performance between the two models? It seems apparent that the more confidence one has that one model is performing better than the other, the more weight that result would carry in the final decision on the appropriateness of using that model. Clearly this is important when at least one of the models is performing moderately well. For example if only one model appears to be unbiased, the degree to which the other is more biased can be a factor in weighing the relative advantage of the apparently unbiased model.

Section 3.2.1 discusses criteria to be considered in determining the relative importance of performance measures. Section 3.2.2 covers techniques for establishing relative confidence in the ability to discriminate between the performance of the two models. Section 3.2.3 provides a rationale for combining these two schemes and a suggested format for the protocol.

3.2.1 Relative Importance of Performance Measures

This subsection discusses factors to be considered when determining what relative weights the various performance measures should carry in the overall evaluation of model performance. The assumption is that the performance results may suggest that the proposed model performs better for some aspects and the reference model for others. Those measures of performance which best characterize the ability of either model to more accurately estimate the concentrations that are critical to decision making should carry the most weight. For example, the reference model may exhibit better performance in estimating the overall concentration field but perform poorer in estimating the concentrations in the vicinity of the maximum concentration. If the estimated maximum concentration controls the emission limit on the source(s) then more weight should to given to performance measures that assess the models' capability to accurately estimate the maximum. In this example, however, some weight should still be given to the relative model performance over the entire domain since this is a measure of the models' capabilities to correctly account for atmospheric processes that influence ambient concentrations and thus adds to (subtracts from in this example) the credibility of the conclusion that the proposed model more accurately predicts the maximum.

A suggested scheme for determining the relative importance of the performance measures is to: (1) define a set of "modeling objectives" or desirable attributes of model performances appropriate to the regulatory problem (the intended application); (2) rank these objectives in order of importance and; (3) assign a maximum possible numerical score that each objective should carry in the overall performance evaluation. Each performance measure and analysis which supports the objective is listed under that objective and perhaps, each numerically weighted according to how well it supports the relative capability of the models to meet that objective.

The scheme is best illustrated by an example. Assume that for a given application accurate prediction of the maximum concentration is the most important modeling objective and that it should carry a weight of 50 out of a total of 100 possible points. (The other modeling objectives would encompass the remaining 50 points.) If a proposed model is clearly better than the reference model, i.e., is unequivocally supported by the performance measure statistics and analyses that characterize that objective, then a score of 50 would be assigned to the comparison between the two models. Conversely, if the reference model is clearly supported then the score of -50 would be assigned to the comparison. A score of zero would indicate that the performance of the two models is the same.

The performance measures that support the determination as to which model better meets the objective of accurate prediction of the maximum concentration are: (1) $D_n(L_n,T_n)$ and $D_n(L,T)$ where n might be the second-highest concentration; (2) the bias, noise and gross variability of D_n where n extends over the upper end of the frequency distribution of the observed data; and (3) the case studies described in Section 3.1.1. Of the total possible 50 points for this objective performance measure (1) might carry a weight of 20; (2) 15 points; and (3) 15 points. The rationale for assigning these weights (recall that this is done before any data are available) is that the proposed model might do poorly on the second-highest concentration but that if it performs better over the upper end of the frequency distribution and accounts for the meteorological variables correctly (the case studies), the comparison score could still be positive. This rationale also assumes that the peak concentration statistics are usually non-robust, i.e., only minimal confidence can be placed in single values of D_n unless they are supported by other statistical data.

To generalize the scheme on how to consider the relative importance of modeling objectives and their supporting performance measures, it is suggested that modeling objectives be ranked in order of importance as first, second and third order objectives.

First order objectives might be: Those concentrations essential to the decision in question are accurately estimated. The essential concentrations are defined by appropriate regulations and are usually given in terms of some peak concentration such as the second-high concentration. As noted in Section 3.1.1 the performance measures, D_n, for the peak estimations are not statistically very meaningful since their values could change significantly when using equivalent data from another time period. Therefore, additional statistical and qualitative analyses must be presented to lend confidence to the residuals for the estimation of the peak.

The performance measures and analysis which delineate the extent to which a model meets the first-order objectives are, therefore:

- The appropriate residuals from Table 3.1.
- Accuracy and precision measures for the top 5% (or top 25) of the observed concentrations.
- Results of case studies described in Section 3.1.1.

These are summarized for the major model type-task categories in Table 3.2.

Second-order objectives might be: Pollutant concentrations are modeled accurately and precisely over an extended range of concentrations. This range of concentrations should be determined on a case-by-case basis. The performance measures which quantify the degree to which a model meets the second-order objectives are summarized in Table 3.3.

Third-order objectives might be: Concentration patterns are modeled realistically over the range of meteorological or other conditions of interest. Demonstration that a model meets these goals involves correlation analyses summarized in Table 3.4.

TABLE 3.2 Summary of Performance Measures for First Order Objectives (Example).

TYPE OF MODEL/SOURCE	TASK	PERFORMANCE MEASURES	REMARKS
Single or multiple source; stable pollutant; short term	Compliance with NAAQS; site of lesser importance	1. $D_n(L_n,T_n)$, $D_n(L,T)$ 2. Bias, noise and gross variability of $D_n(L_n,T_n)$ and $D_n(L,T)$ 3. Case studies as in Section 3.1.1	1. n Specified by the regulations 2. n extends over the upper 5% of the observations or the top 25 observations, whichever is greater
	Site critical, eg. PSD Class 1	1. $D_n(L_n,T_n)$, $D_n,T)$ 2. Bias, noise and gross variability of $D_n(L_n,T_n)$ and $D_n(L,T)$ 3. Spatial correlation of tracer network data 4. Case studies described in Section 3.1.1	1. n Specified by the regulation 2. n extends over the top 5% or top 25 observations, whichever is greater 3. Supplement with isopleths of Co and Cp for high Co periods
Single source or multiple source; stable pollutant; long term	Compliance with NAAQS site of lesser importance	1. Bias of $D_n(L,T)$ 2. Bias by meteorological category as discussed in Section 3.1.2	1. n Extends over all observations above small cutoff value
	Site critical	1. Bias of $D_n(L,T)$ 2. Bias at critical receptor by meteorological category as discussed in Section 3.1.2	1. L_n=Critical site(s); n extends over all observations above small cutoff
Multiple source; short term	Compliance with NAAQS	1. $D_n(L,T)$	1. Simulations for a few days with an urban airshed model

TABLE 3.3 Summary of Performance Measures for Second Order Objectives (Example).

YPE OF MODEL/SOURCE	TASK	PERFORMANCE MEASURES	REMARKS
Single or multiple source; Stable pollutant; short term	Compliance with NAAQS; site of lesser importance	1. Bias, noise and gross variability of $D_n(L_nT_n)$over all sites	1. n extends over all observations above a small cutoff. Also supply distributions of parameters
		2. Bias, noise and gross variability by meteorological category as in Section 3.1.2	
		3. Comparison of cumulative frequency distributions of Co and Cp	3. Tests for goodness of fit
	Site critical eg PSD Class 1	1. Bias, noise and gross variability of $D_n(L_n,T_n)$ at critical sites	1. See Remark 1 above
		2. Bias, noise and gross variability at critical sites by meteorological category as in Section 3.1.2	
		3. Comparison of cummulative frequency distribution at critical sites	3. Tests for goodness of fit
Single or multiple source; Stable pollutant; long term	Compliance with NAAQS	1. Noise and gross variability of $D_n(L_n,T_n)$	1. See Remark 1 above
		2. Noise and gross variability of $D_n(L_n,T_n)$ as in Section 3.1.2	
	Site critical	1. Noise and gross variability at critical sites	1. See Remark 1 above
		2. Noise and gross variability at critical sites by meteorological category	
Multiple source; photochemical; short term	Compliance with NAAQS	1. Bias, noise and gross variability of $D_n(L_n,T_n)$	1. See Remark 1 above

TABLE 3.4 Summary of Performance Measures for Third Order Objectives (Example).

TYPE OF MODEL/SOURCE	TASK	PERFORMANCE MEASURES	REMARKS
Single source and multiple source; Stable pollutant; short term	Compliance with NAAQS; Site of lesser importance	1. Space-time correlation, all data 2. Spatial correlation 3. Temporal correlation	1. Supply scattergrams 2. Isopleths of Co and Cp for important categories 3. Time series of Co and Cp at each site
	Site Critical	1. Space-time correlation 2. Spatial correlation 3. Temporal correlation	1. Remark 1 above 2. Remark 2 above 3. Time series of Co and Cp at critical sites
Single Source and multiple source; stable pollutant; long term	All	1. Space-time correlation	1. Remark 1 above
Multiple source; photochemical; short term	Compliance with NAAQS	1. Space-time correlation 2. Spatial correlation 3. Temporal correlation	1. Remark 1 above 2. Remark 2 above 3. Time series of Co and Cp at each monitor site

3.2.2 Comparison of the Performance Measure Statistics for the Proposed and Reference Models

Once the relative importance of the modeling objectives and the performance measures that support each objective have been established, it is necessary to define the rationale to be used in determining the degree to which each pair of performance measure statistics (or analysis) supports the advantage of one model over the other. Stated differently, it is necessary to have a measure of the degree to which better performance of one model over the other can be established for each performance measure.

While confidence levels are useful for displaying and comparing model performance, they provide no direct statistical measure of the significance associated with comparative performance of the two models. By selecting a predetermined statistical level of significance, a reasonably objective scheme can be established for displaying and weighting the relative performance of each model. For example, it may be desirable to select the rejection probability at 5% for comparisons of the model bias or for comparison of model noise. This figure (5%) can be interpreted as the probability that the statistical test will suggest better performance by one model when, in fact, neither model is performing better. Procedures for establishing confidence limits on each model's performance and for testing the advantage of one model are described in Appendix B.

The concept is easily applied to the performance measures of precision, which measure the scatter of residuals. The most appropriate statistic, e.g., the ratio of model noise is selected using Appendix B and used to determine the statistical significance of the comparison. Higher significance levels, say 5%, would be associated with a high level of confidence that the model with the lower average precision is better. In the protocol an attempt would be made to incorporate the range of possible levels into an objective scheme. For example, if the maximum possible score associated with precision is 10 (positive indicates that the proposed model is better), a score of 10 would be assigned to the proposed model if comparative statistics were of the 5% significance level or less. A zero score would be associated with the 50% level and supportable intermediate scores are given to significance levels between 5% and 50% in some supportable fashion. Similarly, if the statistics suggested that the reference model had better precision, then analogous significance levels could be determined and used to assign negative scores to the comparison.

The concept can be easily extended to determine the relative performance of each model with respect to accuracy. The question is whether one model has less (more) bias than the other. In this case, it's unimportant whether one model tends to overestimate or underestimate, only whether one model tends to be more biased than the other.

The significance level of the difference in bias between the two models is the indicator used in assigning the relative performance score for accuracy. For example, if the reference model has a bias (either too high or too low) which is significantly less than the proposed model at the 10% significance level, then a score of -10 out of a possible -15 might be assigned to the comparison.

For single valued residuals, $D_n(L_n,T_n)$, objective tests for determining the significance associated with observed differences between residuals are not well developed. For this reason, a simple scheme seems to be a reasonable alternative to significance testing such as one which assigns the maximal permissible score to the smallest absolute residual.

For a specific case study it may not be possible to form a totally objective basis for comparing the two models. However, it is still important to clearly define in the protocol the methodology to be used so as not to compromise the decision on these performance measures once the results are known.

For the performance measures that involve correlation coefficients, the rationale is analogous to that for the unsigned residuals. The model with the higher correlation coefficient is better. The degree of advantage is based, in an objective and supportable manner, on the significance level associated with the appropriate statistic (see Appendix B) that compares the relative magnitude of the two correlation coefficients.

Caution should be applied in interpreting the statistical significance associated with comparisons of model performance for each of the various performance measures. This is especially crucial since each of the various statistical tests are based to a varying degree on the assumption that model residuals are independent of one another; an assumption that is clearly not true. For example, model residuals form adjacent time periods (e.g., hour to hour) are known to be positively correlated. Also the proposed and reference model residuals for a given time period are related since each residual is calculated by subtracting the same observed concentration from the estimates of the two models. For these reasons, classical tests described in Appendix B should be viewed as practical interim guides until more rigorous statistical tests for comparing model performance can be evaluated.

3.2.3 Format for the Model Comparison Protocol

The specification of an objective technique for considering the relative importance of the various attributes of good model performance (Section 3.2.1) and the rationale for deciding how well each attribute is supported by one model or the other (Section 3.2.2) constitutes the overall scheme for judging model superiority in the performance evaluation. A suggested format for the model comparison protocol, based on the scoring scheme discussed above, is provided as Table 3.5.

In the first column of the table the modeling objectives relevant to the regulatory problem are listed. The second column lists the performance measures that support that objective. The third column lists the maximum scores (±) that could be attained for each objective and for each of its supporting performance measures. A maximum positive score could be obtained if the proposed model is unequivocally supported; a maximum negative score if the reference model is unequivocally better. In the fourth column the reasons supporting the distribution of maximum weights among the various objectives and performance measures should be listed. The last column should describe in objective terms the rationale to be used for scoring each performance measure. (In Section 3.2.2 a rationale tied to the confidence levels was suggested, for most measures.)

In the middle of the table space is left for any proposed adjustments to the total score that are not adequately represented by the performance statistics. It might be agreed initially that for a particular attribute either the proposed model or reference model is not adequately characterized by the performance measure statistics and should be accounted for in the objective sense as described under "basis."

TABLE 3.5 Suggested Format for the Model Comparison Protocol

Modeling Objective	Supporting Performance Measure	Maximum Score	Basis for Maximum Score	Rationale for Scoring (Significance Criteria)
1.		1.	1.	
	a.	a.	a.	a.
	b.	b.	b.	b.
	c.	c.	c.	c.
	.	.	.	.
	.	.	.	.
2.		2.	2.	
	a.	a.	a.	a.
	b.	b.	b.	b.
	.	.	.	.
	.	.	.	.
	.	.	.	.
3.				
.				
.				
.				
.				
.				
		Total = 100		

Adjustments to Score	Basis	Rationale
1.	1.	1.
2.	2.	2.
.	.	.
.	.	.
.	.	.

Decision Rationale	Basis
Better: Score ≥	
Same: ≥ Score ≥	
Worse: Score ≤	
Absolute Criteria	
1.	1.
2.	2.

Below this, the total scores to be used in judging the overall model performance are defined. The positive score, above which the proposed model would be judged to perform significantly better than the reference model is listed on the bottom line. Marginal scores would form an interval (presumably symmetric) about zero and would be associated with the conclusion that one cannot really discriminate between the performance of the two models.

A number of factors should be considered in the rationale that supports the width of the marginal interval. Some of these factors are related to the representativeness and the amount of data. For example, if off-site data were used, it might be decided to reflect the uncertainty in the representativeness of the data by having a rather broad band of marginal model performance.

Finally at the bottom of the Table, space is left for any "absolute" requirements on model performance. These criteria would allow the setting of any *a priori* standards of performance. For example, the initial decision may be that if the proposed model is found to be grossly inaccurate or grossly biased (gross must be defined), it would not be acceptable for the application even though it performs better, overall, than the reference model.

3.3 Protocol When No Reference Model Is Available

When a reference model is not available, it is necessary to write a different type of protocol based on case-specific criteria for the model performance. However, at the present time, there is a lack of scientific understanding and consensus of experts necessary to provide a supportable basis for establishing such criteria for models. Thus the guidance provided in this subsection is quite general in nature. It is based primarily on the presumption that the applicant and the regulatory agency can agree to certain performance attributes which, if met, would indicate within an acceptable level of uncertainty that the model predictions could be used in decision-making.

A set of procedures should be established based on objective criteria that, when executed, will result in a decision on the acceptability of the model from a performance standpoint. As was the case for the model comparison protocol, it is suggested that the relative importance of the various performance measures be established. Tables 3.2, 3.3, and 3.4 serve as a guide. However, the performance score for each measure should be based on statistics of d, or the deviation of the model estimates form the true concentration, as indicated by the measured concentrations. For each performance measure criteria should be written terms of a statistical test. For example, it might be stated that the average model bias should not be greater than ± X at the Y% significance level. Some considerations in writing such criteria are:

1. Conservatism--This involves the introduction of a purposeful bias that is protective of the ambient standards or increments, i.e., overprediction may be more desirable than underprediction.

2. Risk--It might be useful to establish maximum or average deviation from the measured concentrations that could be allowed.

3. Case Studies--As mentioned in Section 2.5 there may be certain model assumptions or model features that are critical to the intended application. Minimum acceptable performance of the model in certain case studies designed to focus on these critical situations could be established.

4. Experience in the Performance of Models--Several references in the literature[8,9,10,11] describe the performance of various models. These references can serve as a guide in determining the performance that can be expected from the proposed model, given that an analogy with the proposed model and application can be drawn.

As was the case for the model comparison protocol, a decision format or table analogous to Table 3.5 should be established. Execution of the procedures in the table should lead to a conclusion that the performance is acceptable, unacceptable or marginal.

4.0 DATA BASES FOR THE PERFORMANCE EVALUATION

This section describes interim procedures for choosing, collecting and analyzing field data to be used in the performance evaluation. In general there must be sufficient accurate field test data available to adequately judge the performance of the model in estimating all the concentrations of interest for the given application.

Three types of data can be used to evaluate the performance of a proposed model. The preferred approach is to utilize meteorological and air quality data from a specially designed network of monitors and instruments in the vicinity of the sources(s) to be modeled (on-site data). In some cases especially for new sources, it is advantageous to use on-site tracer data from a specifically designed experiment to augment or be used in lieu of long-term continuous data. In infrequent cases where an appropriate analogy to the modeling problem can be identified, it may be possible to utilize off-site data to evaluate the performance of the model.

As a general reference for this section the criteria and requirements contained in the Ambient Monitoring Guidelines for Prevention of Significant Deterioration (PSD),[12] should be used. Much of the information contained in the PSD monitoring guideline deals with acquiring information or ambient conditions in the vicinity of a proposed source but such data may not entirely fulfill the input needs for model evaluation.

All data used as input to the air quality model and its evaluation should meet standard requirements or commonly accepted criteria for quality assurance. New site-specific data should be subjected to a quality assurance program. Quality assurance requirements for criteria pollutant measurements are given in Section 4 of the PSD monitoring guideline. Section 7 of the PSD monitoring guideline describes quality assurance requirements for meteorological data.

The procedures to be used in the performance evaluation described below in Section 4 involve a comparison of the performance of the proposed model with that of the reference model. Thus it is necessary to provide model estimates for both models for each receptor where measured data are available. Usually concentration estimates and measurements are for a one-hour period but may be for a shorter or longer period depending on the characteristics of the model or the sampling method used. All valid data and the corresponding concentration estimates from both models are needed in the performance evaluation. Circumstances bearing on the representativeness of any of the data or concentration estimates should be fully explained for consideration in weighting the results of the performance statistics.

It is also necessary to sum/average estimates and data such that the relative performance of the models can be compared for averaging times corresponding to increments/standards or other decision criteria germane to the problem. For example, SO_2 increments and standards are written in terms of 3-hour, 24-hour and annual averages. Concentration data and model estimates for these averaging times would be used in the performance evaluation discussed in Section 3.

Finally, it should be noted that, when the model is used to make estimates for comparison with standards/increments it is necessary to include a longer period of record of model input data than that collected for the performance evaluation. This is to ensure that the long-term temporal variations of critical meteorological conditions will be adequately accounted for. The Guideline on Air Quality Models provides some guidance on the length of record needed for regulatory modeling.

4.1 On-Site Data

The preferable approach to performance evaluation is to collect an on-site data base consisting of concurrent measurements of emissions, meteorological data and air quality data. Given an adequate sample of these data, an on-site data base designed to evaluate the proposed model relevant to its intended application, should lead to a definitive conclusion on its applicability. The most important goal of the data collection network is to ensure adequate spatial and temporal coverage of model input and air quality data.

In general the spatial and temporal coverage of emissions, meteorological and air quality data used in the performance evaluation should be adequate to show with some confidence how well each model is performing at all points and times for meteorological conditions of interest. Enough data should be collected to allow the calculation of each applicable performance measure discussed in Section 3.1. The data collection should emphasize the area around receptors where high concentrations are expected under critical meteorological conditions. Concurrent emissions data and meteorological data should be representative of the critical conditions for the site. The definition of receptors and meteorological conditions is best obtained from the screening analysis and model estimates described in Section 2.1.

The number of monitors needed to adequately conduct a performance evaluation is often the subject of considerable controversy. It can be argued that one monitor located at the point of maximum concentration for each averaging time corresponding to the standards or increments should be sufficient. However, the points of maximum concentration are not known but are estimated using the model or models that are themselves the subject of the performance evaluation, which of course unacceptably compromises the evaluation. It is possible that the use of data from one or two monitors in a performance evaluation may actually be worse than no evaluation at all since no meaningful statistics can be generated and attempts to rationalize this problem may lead to erroneous conclusions on the suitability of the models.

At the other extreme is a large number of monitors, perhaps 40 or more, that cover the entire modeling domain or area where significant concentrations, above a small cutoff can be reasonably expected, and with enough density such that the entire concentration field (isopleths) can be established. Such a concentration field will allow the calculation of the needed performance statistics and, given adequate temporal coverage as discussed below, would likely result in narrow confidence bands on the model residuals, as discussed in Section 3.1. With these narrow confidence bands it is easier to distinguish between the relative capabilities of the proposed model vs. the reference model to more accurately estimate observed concentrations. When the data field is more sparse, the confidence bands on the residuals for the two models will be broader. As a consequence, the probability of statistically distinguishing the difference between the performance of the two models will be lower.

Thus, the number of monitors needed to conduct a significantly meaningful performance evaluation should be judged in advance. Some other factors that should be considered are:

1. The more accurate the emissions data are, the less noise in the model residuals.

2. Similarly, the more accurately one can pinpoint the location of the plume(s) the less noise that will occur in the model residuals. This can be done by increasing the spatial density and degree of sophistication in meteorological input data, for models that are capable of accepting such data.

3. Models or submodels that are designed to handle special phenomena would logically only be evaluated over the spatial domain where that phenomena would result in significant concentrations. Thus, the monitoring network should be concentrated in that area, perhaps with a few outlying monitors for a safety factor.

In the temporal sense some of the above rationale is also appropriate. A short-term study will lead to low or no confidence on the ability of the models (proposed and reference) to reproduce reality. A multi-year effort will yield several samples and model estimates of the second-highest short-term concentrations thus providing some basis for statistically significant comparison of models for this frequently critical estimate. Realistically, multi-year efforts are usually prohibitive and one has to rely on somewhat circumstantial evidence, the upper end of the frequency distribution, to establish confidence in the models' capabilities to reproduce the second-highest concentration.

In general, the data collected should cover a period of record that is truly representative of the site in question, taking into account variations in meteorological conditions, variations in emissions and expected frequency of phenomena leading to high concentrations. One year of data is normally the minimum, although short-term studies are sometimes acceptable if the results are representative and the appropriate critical concentrations can be determined from the data base. Thus short-term studies are adequate if it can be shown that "worst case conditions" are limited to a specific period of the year and that the study covers that period. Examples might be ozone problems (summer months), shoreline fumigation (summer months) and certain episode phenomena.

Other considerations on the length of record for the performance evaluation are analagous to the considerations for spatial coverage:

1. Accurate emissions data over the period of record diminishes the noise in the temporal statistics. Although data contained in a standard emissions inventory can sometimes be used, it is generally necessary to obtain and explicitly model real time (concurrent with the air quality data used in performance evaluation) emissions data from significant sources. "In-stack" monitoring is highly recommended to insure the use of emission rates comparable in time to the measured and estimated ground-level concentrations.

2. Continuous (minimum of missing data) collection of representative meteorological input data is important.

3. Models designed to handle special phenomena need only have enough temporal coverage to provide an adequate (produce significant statistical results) sample of those phenomena. For example, a downwash algorithm might be evaluated on the basis of 50 or so observations in the critical wind speed range.

It is important that the data used in model development be independent of those data used in the performance evaluation. In most cases, this is not a problem because the model is either based on general scientific principles or is based on air quality data from an analogous situation. However, in some semi-empirical approaches where site-specific levels of pollutants are an integral part of the model, an independent set of data must be used for performance evaluation. The most common examples of these models are statistical approaches where concentrations for various averaging times utilize probability curves derived from site-specific data and for approaches requiring calibration.

When actual air quality data are used in the performance evaluation, it is necessary to distinguish between the contribution to the measured concentration from sources that are included in the model and the contribution attributable to background (or baseline levels). Section 5.4 of the Guideline on Air Quality Models discusses some methods for estimating background. Considerable care should be taken in estimating background so as not to bias the performance evaluation. Incorporation of background data consistently in the proposed model and the reference model is necessary to ensure that no artificial differences in the performance statistics are generated. For example, a "calibrated" model may implicitly include background and if it were compared to a model where background is accounted for differently some biases may be introduced.

4.2 Tracer Studies

The use of on-site tracer material to simulate transport and dispersion in the vicinity of a point or line source has received increasing attention in recent years as a methodology for evaluating the performance of air quality simulation models. This technique is attractive from a number of standpoints.

1. It allows the impacts from an individual source to be isolated from those of other nearby sources which may be emitting the same pollutants.

2. It allows a precise definition of the emission rate.

3. It is generally possible to have a reasonably dense network of receptors in areas not easily accessible for placement of a permanent monitor.

4. It allows for the emissions from a proposed source to be simulated.

There are some serious difficulties in using tracers to demonstrate the validity of a proposed model application. The execution of the field study is quite resource intensive, especially in terms of manpower. Samplers need to be manually placed and retrieved after each test and the samples need to be analyzed in a laboratory. In many cases an aircraft is required to dispense the tracer material. Careful attention must be placed on quality control of data and documentation of meteorological conditions. As a result most tracer studies are conducted as a short term (a few days to a few weeks) intensive campaign where large amounts of data are collected. If conducted carefully, such studies provide a considerable amount of useful data for evaluating the performance of the model. However, the performance evaluation is limited to those meteorological conditions that occur during the campaign. Thus, while a tracer study allows for excellent spatial coverage of pollutant concentrations, it provides a limited sample, biased in the temporal sense, and leaves an unanswered question as to the validity of the model for all points on the annual frequency distribution of pollutants at each receptor.

Another problem with tracer studies is that the plume rise phenomena may not be properly simulated unless the tracer material can be injected into the gas stream from an existing stack. Thus, for new sources where the material is released from some kind of platform, the effects of any plume rise submodel cannot be evaluated.

Given these problems, the following criteria should be considered in determining the acceptability of tracer tests:

1. The tracer samples should be easily related to the averaging time of the standards in question;

2. The tracer data should be representative of "worst case meteorological conditions";

3. The number and location of the samplers should be sufficient to ensure measurement of maximum concentrations;

4. Tracer releases should represent plume rise under varying meteorological conditions:

5. Quality assurance procedures should be in accordance with those specified or referenced in the PSD monitoring guideline as well as other commonly accepted procedures for tracer data;

6. The on-site meteorological data base should be adequate;

7. All sampling and meteorological instruments should be adequately maintained;

8. Provisions should be made for analyzing tracer samples at remote locations and for maintaining continuous operations during adverse weather conditions where necessary.

Of these criteria, items 1 and 2 are the most difficult to satisfy because the cost of the study precludes collection of data over an annual period. Because of this problem it is generally necessary to augment the tracer study by collecting data from strategically placed monitors that are operated over a full year. The data are used to establish the validity of the model in estimating the second-highest short term and the annual mean concentration. Although it is preferable to collect these data "on site," this is usually not possible where a new plant is proposed. It may be possible to use data collected at a similar site, in a model evaluation as discussed in the next subsection.

As is the case for a performance evaluation that uses routine air quality data, sufficient and relevant meteorological data must be collected in conjunction with the tracer study to characterize transport and dispersion and to characterize the model input requirements. Since tracer study data are difficult to interpret, it is suggested that the data and methodologies used to collect the data be reviewed by individuals who have experience with such studies.

4.3 Off-Site Data

Data collected in another location may be sufficiently representative of a new site so that additional meteorological and air quality data need not be collected. The acceptability of such data rests on a demonstration of the similarity of the two sites. The existing monitoring network should meet minimum requirements for a network required at the new site. The source parameters at the two sites should be similar. The source variables that should be considered are stack height, stack gas characteristics and the correlation between load and climatological conditions.

A comparison should be made of the terrain surrounding each source. The following factors should be considered:

1. The two sites fall into the same generic category of terrain:

   a. flat terrain

   b. shoreline conditions

   c. complex terrain:

      (1) three-dimensional terrain elements, e.g., isolated hill
      (2) simple valley
      (3) two dimensional terrain elements, e.g., ridge
      (4) complex valley

2. In complex terrain the following factors assist in determining the similarity of the two sites

   a. aspect ratio of terrain, i.e., ratio of:
      (1) height of valley walls to width of valley
      (2) height of ridge to length of ridge
      (3) height of isolated hill to width of hill base

   b. slope of terrain

   c. ratio of terrain height to stack/plume height

   d. distance of source from terrain, i.e., how close to valley wall, ridge, isolated hill

   e. correlation of terrain feature with prevailing winds

It is very difficult to secure data sets with the above emission configuration/terrain similarities. Nevertheless, such similarities are of considerable importance in establishing confidence in the representativeness of the performance statistics. The degree to which the sites and emission configuration are dissimilar is a measure of the degree to which the performance evaluation is compromised.

More confidence can be placed in a performance evaluation which uses data collected off-site if such data are augmented by an on-site tracer study (See Section 4.2). In this case the considerations for terrain similarities still hold, but more weight is given to the comparability of the two sets of the observed concentrations. On-site tracer data can be used to test the ability of the model to spatially define the concentration pattern if a variety of meteorological conditions were observed during the tracer tests. Off-site data must be adequate to test the validity of the model in estimating maximum concentrations.

5.0 MODEL ACCEPTANCE

This section describes interim criteria which can be used to judge the acceptability of the proposed model for the specific regulatory application. This involves execution of the performance protocol which will lead to a determination that the model performs better, about the same as, or worse than the reference model or performs acceptably, marginally, or unacceptably in relation to established site-specific criteria. Depending on the results of the performance evaluation, the overall decision on the acceptability of the model might also consider the results of the technical evaluation of Section 2. Finally, because the procedures proposed in this document are relatively new and untested, it is advisable to reexamine the conclusion reached to see if it makes good common sense.

5.1 Execution of the Model Performance Protocol

Execution of the model performance protocol involves: (1) collecting the performance data to be used (Section 4.0); (2) calculation and/or analysis of the model performance measures (Section 3.1); and (3) combining the results in the objective manner described in the protocol (Section 3.2 or Section 3.3) to arrive at a decision on the relative performance of the two models.

Table 5.1 shows a format which can be used to accommodate the results of the model comparison protocol described in Section 3.2.3. If a different protocol format is prepared, it should have the same goal, i.e., to arrive at a decision on whether the proposed model is performing better, about the same, or worse than the reference model.

TABLE 5.1 Suggested Format for Scoring the Model Comparison

Modeling Objective	Supporting Performance Measures	Score	Statistics, Analyses and Calculations that Support the Score
1. ________	a. ________ b. ________ c. ________ . ________ . ________ . ________	1. ________ a. ________ b. ________ c. ________ . ________ . ________ . ________	1. ________ a. ________ b. ________ c. ________ . ________ . ________ . ________
2. ________	a. ________ b. ________ c. ________ . ________ . ________ . ________	2. ________ a. ________ b. ________ c. ________ . ________ . ________ . ________	2. ________ a. ________ b. ________ c. ________ . ________ . ________ . ________
3. ________ . ________ . ________ . ________		3. ________ . ________ . ________ . ________	3. ________ . ________ . ________ . ________
Preliminary Score ________ Adjustments to Score ________		1. ________ 2. ________ 3. ________	1. ________ 2. ________ 3. ________

Final Score ________

Decision on Performance Evaluation (Better, Same, Worse) ________

Absolute Requirements Satisfied? ________

The first two columns in the upper half of Table 5.1 are analogous to those in Table 3.5. The third column contains the actual score for each modeling objective as well as the sub-scores for each supporting performance measure. The scores in this column cannot exceed the maximum scores allowed in the protocol. The last column is for the statistics, graphs, analyses and calculations that determine the score for each performance measure, although most of this information would probably be in the form of attachments.

The bottom part of the table is for the preliminary score (obtained from totaling the scores from each objective), adjustments to the score with supporting data, analysis, etc. and the final score. The final score would determine whether the proposed model is performing better, marginally or worse in comparison to the reference model. This result is used in Section 5.2 to determine overall acceptability of the model.

At the bottom of the table, space is available to include the results (yes or no) of any absolute requirements that may be specified in the protocol. Failure to meet these requirements presumably means the model is unacceptable.

If the decision scheme is based on performance criteria alone, a scoring table based on the procedures contained in Section 3.3 should be employed. The resulting conclusions of acceptable, marginal or unacceptable is used in Section 5 to determine the overall acceptability of the model.

5.2 Overall Acceptability of a Proposed Model

Until more objective techniques are recommended, it is suggested that the final decision on the acceptability of the proposed model be based primarily on the results of the performance evaluation. The rationale is that the overall state of the modeling science has many uncertainties in the basic theory regardless of what model is used, and that the most weight should be given to actual proven performance. Thus when a proposed model is found to perform better than the reference model, it should be accepted for use in the regulatory application. If the model performance is clearly worse than that of the reference model, it should not be used. Similarly, if the performance evaluation is not based on comparison with a reference model, acceptable performance should imply that the model be accepted, while unacceptable performance would indicate that it is inappropriate.

When the results of the performance evaluation are marginal or inconclusive, then the results of the technical evaluation discussed in Section 2 should be used as an aid to deciding on the overall acceptability. In this case, a favorable (better than the reference model) technical review would suggest that the model be used, while a marginal or worse determination would indicate that the model offers no improvement over existing techniques. If Section 2.5 were used to determine technical acceptability, a marginal or inconclusive determination on scientific supportability combined with a marginal performance evaluation would suggest that the model not be applied to the regulatory problem.

5.3 Common Sense Perspective

One objective of this document is to provide a framework for organizing the procedures and criteria for model evaluation such that the evaluation can be conducted in as consistent and objective a manner as possible. However, this framework must of necessity be flexible to allow for incorporating additional knowledge about model evaluation, performance measures and criteria.

For example, truly objective criteria for evaluating the technical aspects of models and scientifically acceptable model performance standards are not yet available.

The user should realize that there are many unresolved items at this time. Especially lacking is a totally scientific methodology for combining: (1) various performance statistics, (2) confidence in the scientific basis for the model, (3) data accuracies, and (4) other positive and negative attributes of the model into a single overall determination of the applicability and validity of the model. Given these concerns, two caveats are:

1. The procedures proposed are relatively untested. Although they are based on inputs and comments from a number of scientists in the field, it remains to be seen what problems may turn up in a real situation. Thus, when an evaluation is completed, it seems only prudent to look back over the analyses and the results to see if they really make sense.

2. The assumption has been made that a given regulatory problem requires a model estimate and that the best way to determine the appropriate technique is to evaluate the relative applicability of available models. No determination is made on whether the models are "accurate enough" to be accepted. This is the realm of performance standards, which are not addressed. However, the analysis will produce performance statistics which could be compared to standards, if they existed. If the statistics suggest gross inaccuracies or biases, even in the better of the models, it might be prudent to advise the decision maker that other modeling or monitoring information should be used to resolve the regulatory problem.

6.0 REFERENCES

1. Environmental Protection Agency. "Guideline on Air Quality Models," EPA-450/2-78-027, Office of Air Quality Planning and Standards, Research Triangle Park, N. C., April 1980.

2. American Meteorological Society. "Judging Air Quality Model Performance," Draft Report from Workshop on Dispersion Model Performance held at Woods Hole, Mass., September 1980.

3. Environmental Protection Agency. "Guideline for Use of Fluid Modeling to Determine Good Engineering Practice Stack Height," Draft for public comment, EPA 450/4-81-003, Office of Air Quality Planning and Standards, Research Triangle Park, N. C., June 1981.

4. Environmental Protection Agency. "Guideline for Fluid Modeling of Atmospheric Diffusion," EPA 600/8-81-008, Environmental Sciences Research Laboratory, Research Triangle Park, N. C., April 1981.

5. Environmental Protection Agency. "Guideline for Determination of Good Engineering Practice Stack Height (Technical Support Document for Stack Height Regulations)," EPA 450/4-80-023, Office of Air Quality Planning and Standards, Research Triangle Park, N. C., July 1981.

6. U. S. Congress. "Clean Air Act Amendments of 1977," Public Law 95-95, Government Printing Office, Washington, D. C., August 1977.

7. Environmental Protection Agency. "Workbook for Comparison of Air Quality Models," EPA 450/2-78-028a, EPA 450/2-78-028b, Office of Air Quality Planning and Standards, Research Triangle Park, N. C., May 1978.

8. Bowne, N. E. Preliminary Results from the EPRI Plume Model Validation Project--Plains Site. EPRI EA-1788-SY, Project 1616 Summary Report, TRC Environmental Consultants Inc., Wethersfield, Connecticut, April 1981.

9. Lee, R. F., et. al. Validation of a Single Source Dispersion Model, Proceeding of the Sixth International Technical Meeting on Air Pollution Modeling and Its Application NATO/CCMS, September 1975.

10. Mills, M. T., et. al. Evaluation of Point Source Dispersion Models, Draft Report Submitted by Teknekton Research, Inc. to U. S. EPA, January 1981.

11. Londergan, R. J., et. al. Study Performed for the American Petroleum Institute--An Evaluation of Short-Term Air Quality Models Using Tracer Study Data, Submitted by TRC Environmental Consultants, Inc. to API., October 1980.

12. Environmental Protection Agency. "Ambient Monitoring Guideline for Prevention of Significant Deterioration (PSD)," EPA 450/4-80-012, Office of Air Quality Planning and Standards, Research Triangle Park, N. C., November 1980.

APPENDIX A

REVIEWER'S CHECKLIST

Each proposal to apply a nonguideline model to a specific situation needs to be reviewed by the appropriate control agency which has jurisdiction in the matter. The reviewing agency must make a judgment on whether the proposed model is appropriate to use and should justify this judgment with a critique of the applicant's analysis or with an independent analysis. This critique or analysis would normally become part of the record in the case. It should be made available to the public hearing process, used to justify SIP revisions or used in support of other proceedings.

The following checklist serves as a guide for writing this critique or analysis. It essentially follows the rationale in this document and is designed to ensure that all of the required elements in the analysis are addressed. Although it is not necessary that the review follow the format of the checklist, it is important that each item be addressed and that the basis or rationale for the determination on each item is indicated.

CHECKLIST FOR REVIEW OF MODEL EVALUATIONS

I. Technical Evaluation

A. Is all of the information necessary to understand the intended application available?

1. Complete listing of sources to be modeled including source parameters and locations?

2. Maps showing the physiography of the surrounding area?

3. Preliminary meteorological and climatological data?

4. Preliminary estimates of air quality sufficient to (a) determine the areas of probable maximum concentrations, (b) identify the probable issues regarding the proposed model's estimates of ambient concentrations and, (c) form a partial basis for design of the performance evaluation data base?

B. Is the reference model appropriate?

C. Is enough information available on the proposed model to understand its structure and assumptions?

D. Are the results of the technical comparison of the proposed and reference models supportable?

1. Were procedures contained in the Workbook for Comparison of Air Quality Models followed? Are deviations from these procedures supportable or desirable?

2. Are the comparisons for each application element complete and supportable?

3. Do the results of the comparison for each application element support the overall determination of better, same or worse?

E. For cases where a reference model is not used, is the proposed model shown to be applicable and scientifically supportable?

II. Model Performance Protocol

A. Are all the performance measures recommended in the document to be used? For those performance measures that are not to be used, are valid reasons provided?

B. Is the relative importance of performance measures stated?

1. Have modeling objectives that best characterize the regulatory problem been properly chosen and objectively ranked?

2. Are the performance measures that characterize each objective appropriate? Is the relative weighting among the performance measures supportable?

C. How are the Performance Measure Statistics for the Proposed and the Reference Model to be Compared?

1. Are significance criteria used to discriminate between the performance of the two models established for each performance measure?

2. Is the rationale to be used in scoring the significance criteria supportable?

3. Is the proposed "scoreband" associated with marginal model performance supported?

4. Are there appropriate performance limits or absolute criteria which must be met before the model could be accepted?

D. How is Performance to be Judged When No Reference Model is Used?

1. Has an objective performance protocol been written?

2. Does this protocol establish appropriate site-specific performance criteria and objective techniques for determining model performance relative to these criteria?

3. Are the performance criteria in keeping with experience, with the expectations of the model and with the acceptable levels of uncertainty for application of the model?

III. Data Bases

A. Are monitors located in areas of expected maximum concentration and other critical receptor sites?

B. Is there a long enough period of record in the field data to judge the performance of the model under transport/dispersion conditions associated with the maximum or critical concentrations?

C. Are the field data completely independent of the model development data?

D. Where off-site data are used, is the situation sufficiently analogous to the application to justify the use of the data in the model performance evaluation?

E. Will enough data be available to allow calculation of the various performance measures defined in the protocol? Will sufficient data be available to reasonably expect that the performance of the model relative to the reference model or to site-specific criteria can be established?

IV. Is the Model Acceptable

A. Was execution of the performance protocol carried out as planned?

B. Is the model acceptable considering the results of the performance evaluation and the technical evaluation?

C. Does the result of the model evaluation make good common sense?

APPENDIX B

CALCULATION OF PERFORMANCE MEASURES AND RELATED PARAMETERS

This appendix presents methods of calculation of performance measures and related parameters and procedures for applying and interpreting statistical tests of model performance. The parameters and tests recommended follow the results of the AMS Workshop on Dispersion Model Performance (Fox 1980). A summary of this Workshop appears as Appendix C.

Two concerns of Workshop participants were that air quality data are often not normally distributed and that sequential values of meteorological and air quality parameters are not independent of one another. This latter concern results from persistence of meteorological events.

These two concerns are not directly addressed in this appendix since both have been identified as areas needing research. In the majority of calculations and procedures discussed in this appendix, methods are given both for situations in which data follow a normal distribution and for situations for which the normal distribution does not apply.

In evaluation studies using large data sets, some randomized data selection to form a subset of independent data can answer the second concern.

B.1 Definition of Residuals

This subsection discusses the calculation of residuals described in the report. The first type of residual discussed in the report was the difference between observed and estimated concentrations paired in time and space and covering a range of observed concentrations from some small cutoff value to the highest observed. The second type discussed included differences between observed and estimated concentrations paired in various ways in time and space. The data set for this second type of residual includes at most the upper five percent or upper 25 observed concentrations, whichever is greater.

B.1.1 Residuals Covering a Wide Range of Observed Values

Air quality model performance evaluation is primarily based on analysis of the differences between observed and estimated concentrations. The primary parameter for this analysis is the model residual, d, defined as

$$d\ (l,t) = Co\ (l,t) - Cp\ (l,t) \qquad \text{(B.1)}$$

where: d (l,t) is the model residual at location, l, and time, t.

Co = observed concentration

Cp = predicted concentration.

To avoid possible misinterpretation one should note that the residual, d, measures the amount of under-estimation by the model.

Although subsequent statistical analysis of residuals is most valid when performed on d (l,t) as defined in Equation (B.1), there are situations when source strengths may vary significantly over the period of record for which the model will be evaluated. In these cases it may be more meaningful to define a model residual prorated to the source strength. In this case the prorated residual, dq (l,t), is defined as

$$dq\ (l,t) = d\ (l,t) * Qo/Q(t) \qquad (B.2)$$

where Qo = the nominal constant source strength used as the base for prorating the residuals

Q(t) = the actual source strength during the period t.

It is difficult to specify how much variation in source strength may be significant, but a variation of ± 25% about the mean is suggested. The information derived from model performance evaluation must often be communicated to persons with nontechnical backgrounds. Such persons may not know if a 10 ppm average underestimation of carbon monoxide is better or worse than a 0.10 ppm average error in SO_2. Therefore, for ease of communication, we suggest that analysis results include the analysis of the relative residual, dr (l,t), defined as

$$dr\ (l,t) = \frac{d\ (l,t)}{Co\ (l,t)} *100. \qquad (B.3)$$

The behavior of the relative residual causes some statistical problems and, therefore, should not be used as a basis for making decisions, but for communication of results.

B.1.2 Residuals for Peak Concentration

Important peak concentrations are specified in regulations. The residuals which measure the accuracy of prediction of these peak concentration are determined by comparing observed and predicted concentrations paired in various ways in space and time. Table B.1 shows the residuals which measure the accuracy of estimation of the peak. The symbols are interpreted

Table B.1 Residuals to Measure Accuracy of Peak Prediction

Paired in	Residual Set
Space & Time	$D_n (L_n, T_n) = C_o (L_n, T_n) - C_p (L_n, T_n)$
Space not Time	$D_n (L_n, T) = C_o (L_n, T_n) - C_p (L_n, T_j)$
Time not Space	$D_n (L, T_n) = C_o (L_n, T_n) - C_p (L_j, T_n)$
Unpaired	$D_n (L, T) = C_o (L_n, T_n) - C_p (L_r, T_m)$

L_n = monitor site for nth highest observed concentration.

T_n = time of nth highest observed concentration.

T_j = time of nth highest estimated concentration at site Ln.

L_j = site of nth highest estimated concentration during time T_n.

$C_p (L_r, T_m)$ = nth highest estimated concentration.

$C_o (L_n, T_n)$ = nth highest observed concentration.

(L_r, T_m) = site and time of highest estimated concentration (Generally, $L_r \neq L_n$ and $T_m \neq T_n$).

as follows: The subscript, n, on D_n indicates the rank order of the observation, i.e., D_2 is the residual for the second highest prediction and D_{19} that for the 19th highest. The parameters in parentheses indicate the degree of pairing in time and space. L indicates a station location; the subscript n indicates the station of the nth highest observation. T indicates an observation period; the subscript n indicates the time period of the nth highest observation.

Table B.2 presents example values of observed and estimated concentrations for three stations and four measurement periods. We want to determine the accuracy of prediction of the second-highest concentration (n = 2). The second-highest observed concentration (Co = 1.25) occurs at Station 1 (L_2 = Station 1) time period three (T_2 = Period 3).

Table B.2. Example Observed and Estimated Concentrations of SO_2 (ppm)

	Station 1		Station 2		Station 3	
	Co	Cp	Co	Cp	Co	Cp
1	1.02	1.07	1.01	0.82	0.96	0.87
2	1.14	0.85	1.03	0.96	1.02	0.94
3	1.25	1.05	1.07	1.03	1.13	1.04
4	1.36	1.11	1.23	1.10	1.22	1.09

Then D_2 (L_2, T_2) = 1.25-1.05 = 0.20 ppm is the residual for prediction of the second-high concentration paired in space and time.

The unpaired concentrations produce the residual

$$D_2\ (L,\ \bar{T}) = 1.25\text{-}1.10 = 0.15 \text{ ppm.}$$

B.2 Analysis of Bias and Gross Error

Determining the bias and gross error of model predictions involves analyzing the distribution of the residual, d, and/or simple functions of d such as the absolute value or the square.

B.2.1 Bias of Model Predictions

The bias of model predictions is measured by the mean value,

$$\bar{d} = \frac{\Sigma d}{N} \qquad \text{(B.4)}$$

and is the statistical first moment of the distribution of the residual. The number of observations, N, extends over the range of concentrations or over the meteorological conditions of interest. For first order modeling objectives, N can extend over the upper 5% or upper 25 observations.

The 95% confidence limits about the mean are calculated using Student's "t" distribution. The value of $t_{(0.025;\ N-1)}$ is found from any standard table of Student's "t" (e.g., Selby, 1972, p. 617). The value 0.025 is the probability that a value will be greater than the upper limit of the 95% confidence bound. The value N-1 is the number of degrees of freedom for the variable "t." The true value of the mean, μ_d, is then given by

$$\bar{d} - \frac{t_{(0.025,\ N-1)}\ S_d}{\sqrt{N}} \leq \mu_d \leq \bar{d} + \frac{t_{(0.025,\ N-1)}\ S_d}{\sqrt{N}},$$

where S_d is the sample standard deviation discussed in the next section. If N is sufficiently large (> 100 or so) the value for t is 1.96.

This method of calculating the confidence limits assumes that the distribution of the values $\overline{d}$ is normal. This assumption is more nearly satisfied with a large number of observations, N.

B.2.1.1 Confidence limits for nonnormally distributed variables.

Javitz and Ruff (1979) discuss a procedure for calculating the confidence limits about the mean value of a nonnormally distributed variable. The method uses the results of the central limit theorem which states that the sample means of any variable are normally distributed if the size of the sample is large enough. The method is:

Step 1: Subdivide the data set into five data subsets so that each subset contains data from every fifth time period. The first subset would contain data from time periods 1, 6, 11, etc. The second subset would contain data from time periods 2, 7, 12, etc. If the full data set has a periodicity of five time periods (e.g., average daily values for week days only), then the data should be divided into a different number of subsets.

Step 2: Compute the average value of the desired parameter, e.g., d, for each subset. These are labeled d_1, d_2 or d_k, where k is the number of subsets. The value d_1 is then the average value of d for subset 1.

Step 3: Compute the sample standard deviation of the subset means:

$$S = \frac{1}{k - 1} \sum_{1}^{k} (d_i - \overline{d}) ,$$

where $\overline{d}$ is the mean value for the whole data set (equation B-4).

Step 4: The 95% confidence material for the true value of the mean value, $\bar{d}$, is given by:

$$\bar{d} \pm t_{(0.975,k-1)} \frac{S}{\sqrt{k}}$$

where $T_{0.975,k-1}$ is the upper 97.5 percent critical value of the student's "t" distribution with k-1 (e.g., four) degrees of freedom.

B.2.2 Model Precision

Model precision, also known as gross error, refers to the average amount by which estimated and observed concentrations differ as measured by residuals with no algebraic sign. While large positive and negative residuals can cancel when measuring model bias, the unsigned residuals comprising the precision measures do not cancel and thus provide estimation of the error scatter about some reference point. This reference point can be the mean error or the desired value of zero. Two types of precision measure are the noise, which delineates the error scatter about the mean error, and the gross variability, which delineates the error scatter about zero error.

The performance measure for noise is either the variance of the residuals or the standard deviation of the residuals. The standard deviation is the square root of the variance, where

$$S_d^{\,2} = \Sigma \frac{(d-\bar{d})^2}{N-1} \qquad (B.5)$$

is the variance of the sample of the residuals and N the number of observations.

The performance measure for gross variability is the mean square error, or the root-mean-square-error. The mean square error is defined by

$$MSE_d = \Sigma \frac{d^2}{N} \qquad (B.6)$$

The bias, noise and gross variability are related by

$$MSE_d = (\frac{N-1}{N}) \; S_d{}^2 + (\overline{d})^2 \qquad (B.7)$$

An alternate performance measure for the gross variability is the mean absolute residual defined by

$$\overline{|d|} = \Sigma \frac{|d|}{N} \qquad (B.8)$$

The mean absolute residual is statistically more robust than the root-mean-square-error; that is, it is less affected by removal of a few extreme values. The confidence limits on the variance are calculated using the chi-squared distribution in the following manner.

1. Choose values of $\chi^2{}_{(0.025,N-1)}$ and $\chi^2{}_{(0.975,N-1)}$ from standard χ^2 tables (e.g., ERC Standard Mathematics Tables, 20th Edition, pg. 619), where N-1 is the number of degrees of freedom.

2. The confidence limits are given by

$$\frac{(N - 1)S_d^2}{\chi^2{}_{(0.975,N-1)}} \leq \sigma^2 \leq \frac{(N - 1)S_d^2}{\chi^2{}_{(0.025,N-1)}}$$

where σ^2 is the true variance.

The confidence limits on the average absolute residual are calculated as outlined in Section B.2.1.

B.3 Correlation Analysis

Correlation analysis consists of coupled space-time analysis, spatial analysis and temporal analysis.

B.3.1 Space-time Analysis

Coupled space-time correlation analysis involves computing the Pearson's correlation coefficient and parameters of the linear least squares regression equation. For space-time analysis, observed and estimated concentrations from all stations and time periods are used in the calculations. The overall Pearson's correlation coefficient, r, is defined by:

$$r = \frac{\Sigma\ (C_o - \bar{C}_o)\ (C_p - \bar{C}_p)}{\{\Sigma\ (C_o - \bar{C}_o)^2\ .\ \ (C_p - \bar{C}_p)^2\}^{1/2}} \qquad \text{(B.9)}$$

The linear least squares regression line is

Co = a + bCp,

where Co is the estimated "true" concentration, b is the slope of the regression line

$$b = \frac{N\ \Sigma\ C_oC_p - (\Sigma\ C_o)(\Sigma\ C_p)}{N\ \Sigma\ C_p^{\ 2} - (\Sigma\ C_p)^2} \qquad \text{(B.10)}$$

and a is the intercept defined by

$$a = \bar{C}_o - b\bar{C}_p \qquad \text{(B.11)}$$

A scattergram of the Co and Cp data pairs is supplementary information which should be presented.

B.3.2 Spatial Correlation Analysis

Spatial correlation analysis involves calculating the spatial correlation coefficient and presenting isopleth analyses of the estimated and observed concentrations for particular periods of interest. The spatial coefficient measures the degree of spatial alignment between the estimated and observed concentrations. The method of calculation essentially involves calculating the Pearson's correlation coefficient for each time period and determining an average over all time periods. The specifics of the method are:

Calculate the Pearson's correlation coefficient, r_t, for each averaging period, t, from Equation B.9. Change the variable to ϕ_t for each time period from

$$\phi_t = \frac{1}{2} \ln \left[\frac{1 + r_t}{1 - r_t} \right]. \tag{B.12}$$

Calculate the mean value, $\overline{\phi}_t$, by averaging over the number of time periods. Estimate the average spatial correlation coefficient

$$\overline{r}_t = \frac{\exp(2\overline{\phi}_t)-1}{\exp(2\overline{\phi}_t)+1} \cdot \tag{B.13}$$

The 95% confidence limits about the estimate of the spatial correlation coefficient are calculated from

$$\text{Limits} = \tanh\left(\overline{\phi}_t \pm \frac{1.96}{\sqrt{N_t(K-3)}}\right) \tag{B.14}$$

where N_t = the number of time periods and

K = the number of Co, Cp data pairs in each time period (the number of monitoring locations).

Estimates of the spatial correlation coefficient for single source models are most reliable for calculations based on data intensive tracer networks. Isopleths of the distributions of estimated and observed concentrations for periods of interest should be presented and discussed.

B.3.3 Temporal Analysis

Temporal correlation analysis involves calculating the temporal correlation coefficient and presenting time series of observed and estimated concentrations or of the model residual for each monitoring location. The temporal correlation coefficient measures the degree of temporal alignment between observed and estimated concentrations. The method of calculation is similar to that for the spatial correlation coefficient. Calculate the Pearson's correlation coefficient, r_l, for each monitoring location, l, from Equation B.9 Change the variable to ϕ_l for each monitor location using

$$\phi_l = \frac{1}{2} \ln \left[\frac{1 + r_l}{1 - r_l} \right] \quad (B.15)$$

Average over the number of monitor locations to produce the value $\bar{\phi}_l$. Estimate the average temporal correlation coefficient $\bar{r}_l$ from

$$\bar{r}_l = \frac{\exp (2\bar{\phi}_l)-1}{\exp (2\bar{\phi}_l)+1} \quad (B.16)$$

The 95% confidence limits about the mean temporal correlation coefficient are calculated from

$$\text{limits} = \tanh\left(\bar{\phi}_1 \pm \frac{1.96}{\sqrt{N_1(M-3)}}\right) \qquad \text{(B.17)}$$

where N_1 = the number of monitoring locations and

M = the number of Co, Cp data pairs for each monitoring location (the number of time periods).

Time series of Co and Cp or of model residuals should be presented and discussed for each monitoring location.

B.4 Statistical Tests

This section discusses the use of the statistical test of hypotheses mentioned in the body of the report. A general discussion of the concept of statistical hypothesis testing can be found in any statistical text (e.g., Panofsky and Brier, 1965, Chapter III).

B.4.1 Comparison of Cumulative Distribution Functions

Comparison of the cumulative distribution functions involves constructing quantile-quantile (Q-Q) plots and testing for statistically significant differences between the distributions. Karl (1978) presents examples of the technique applied to ozone measurements in St. Louis. The techniques discussed here can be used to analyze differences between distributions of Co and Cp at a given station or distributions of the residual, d, for reference and candidate models. Ogives of the cumulative distributions of the two parameters to be compared are first plotted as in the example in Figure B.1. The Q-Q plot for data such as those shown in the example simply consists of plotting values of one parameter at a given cumulative frequency percentages against the values of the second parameter for the same cumulative frequency percentages as shown in the example in Figure B.2. If the two distributions are identical, then points will fall along the straight line with slope equal to one.

Q-Q plots are useful in detecting differences in distributions in data sets. The plots do not require any assumptions regarding the form of the distributions of the two data sets and the statistical significance of any differences can be determined by non-parametric methods. The Wilcoxen-matched pair, signed-rank test is used to test the null hypothesis that there is no significant difference between the two distributions. (See Panofsky and Brier, 1965, pp. 64-66 or Siegel, 1956 for more complete discussion of the test.)

Step 1. Form the differences

$$\Delta(q) = X(q) = Y(q)$$

where $\Delta(q)$ = the difference between values at cumulative frequency quantile, q

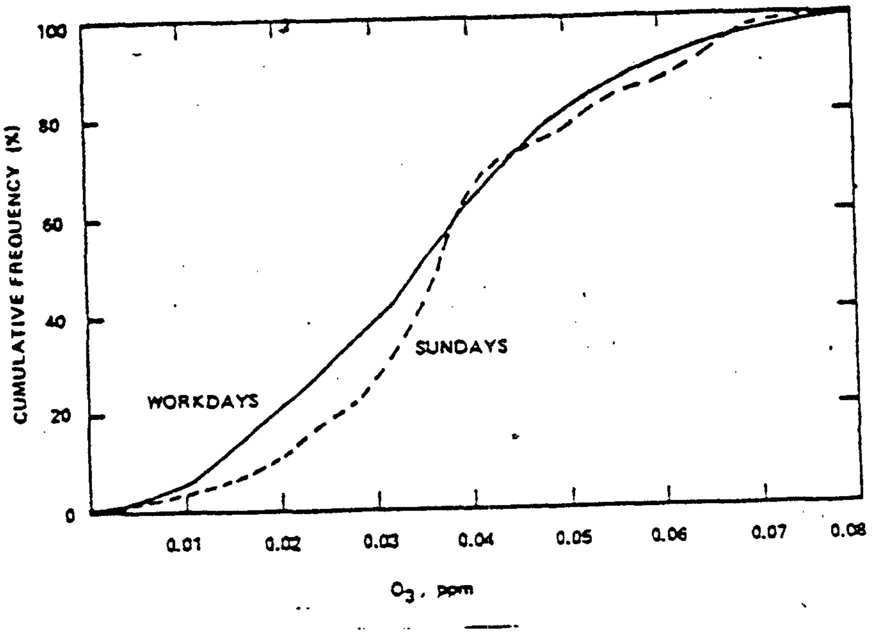

Figure B.1. Ogives for 16-h average O_3 concentrations on Sundays and workdays for the inner sites. (Karl, 1978)

Day of the week variations of pollutants

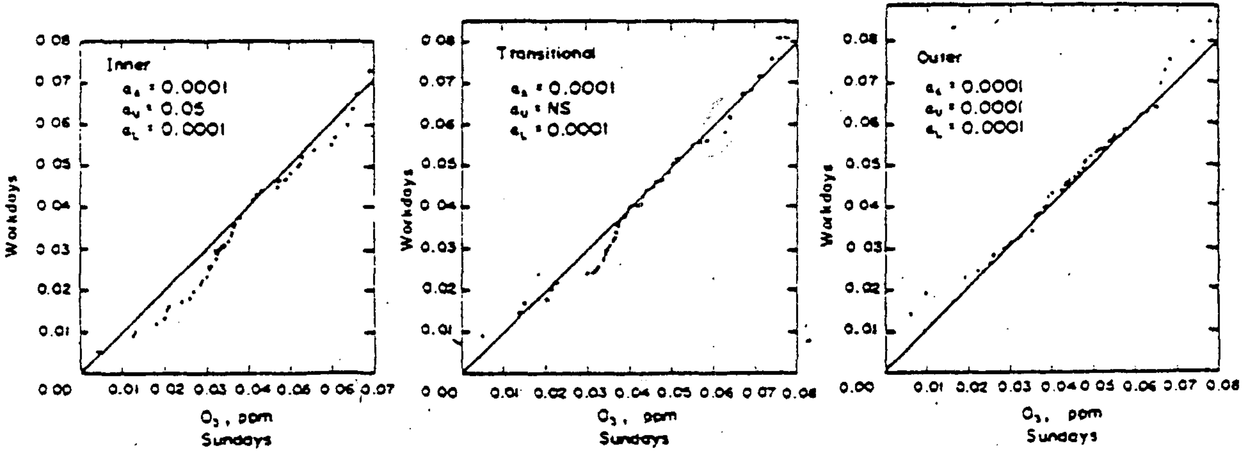

Figure B.2. Quantile-quantile plots of ozone for Sundays vs workdays. α gives the significance level for rejection of the null hypothesis that there is no difference between paired values. Subscript A denotes all quantiles plotted; U denotes quantiles at or above the 50th percentile; L denotes quantiles at or below the 50th percentile. NS implies no significance, the null hypothesis must be accepted. (Karl, 1978)

$X(q)$ = value of parameter X at q

$Y(q)$ = value of parameter Y at q

Step 2. The absolute values of $|\Delta(q)|$ are ranked from highest to lowest, the highest value is given Rank = 1, the second highest Rank = 2, etc. The same average value is given to a number of identical differences. Zero differences are excluded before the ranking

Step 3. The algebraic sigh of each $\Delta(q)$ is assigned to its corresponding rank value.

Step 4. The test statistic, R, is calculated by adding the rank values for the fewest cases of the same sign.

Step 5. Determine the critical value of the test statistic by entering Table B.3 with the number of non-zero differences.

Step 6. If the absolute value of R is less than the critical value of the test statistic, reject the null hypothesis and conclude that the two distributions are significantly different.

If the absolute value of R is greater than the critical value of the test statistic, do not reject the null hypothesis and conclude that there is no significant difference between the distributions.

B.4.2 Comparison of Means

Tests for comparison of two mean values test the null hypothesis that the means are equal. The alternate hypothesis is that

Table B.3. Critical values for 5% significance Wilcoxen matched-pair, signed-rank test.[1]

Number of Non-Zero Differences	Critical Value of Test Statistic 5% Level
6	0
7	2
8	4
9	6
10	8
11	11
12	14
13	17
14	21
15	25
16	30
17	35
18	40
19	46
20	52
25	89
> 25	$\frac{M(M+1)}{4}-1.96\left[\frac{M(M+1)(2M+1)}{4}\right]^{\frac{1}{2}}$

[1] Kreysig (1970, P. 461)

one mean is greater than the other. The two tests discussed in this subsection are the student's "t" test, a parametric test used when data approximate a normal distribution, and the Wilcoxen-Mann-Whitney test, a nonparametric test which does not assume any form for the distribution.

B.4.2.1 The Student's "t" Test

The Student's "t" test should be used to test the equality of two sample means when the distributions approximate a normal distribution. When the distributions are nonnormal, the test might still be used, but will be less powerful (Till, 1974, p. 61). If the distributions are known to be much different from normal, a nonparametric test should be used (see Section B.4.2.2).

The procedure tests the null hypothesis that the two means are equal. The alternate hypothesis is that one mean is greater than the other. The alternate hypothesis results from inspection of the sample values of the two means. For example, if we are testing for differences between the mean residual of the candidate model, $\overline{d}_{can}$, and the mean residual of the reference model, $\overline{d}_{ref}$, and inspection of the values shows $\overline{d}_{ref} > \overline{d}_{can}$, then the alternate hypothesis would be:

$$\overline{d}_{ref} > \overline{d}_{can} \ .$$

There are two possible cases for this test. Case A where the variances are equal but unknown and Case B where the variances are unequal and unknown.

Case A: Variances unknown but equal. The F-test for equality of variances is discussed in Section B.5. As stated above the null hypothesis is:

$$H_o: \ \mu_x = \mu_y,$$

and the alternate hypothesis is:

$$H_o: \ \mu_x > \mu_y,$$

where μ_x and μ_y are determined by inspection of the two sample means.

Step 1: The critical value of "t" is determined from any standard Student's "t" tables (e.g., Selby, 1972, P. 617) at the 95% confidence level and $n_1 + n_2 - 2$ degrees of freedom. The values n_x and n_y are the number of observations for parameters X and Y respectively.

Step 2: Calculate the test statistic

$$T = (\overline{X} - \overline{Y}) \left[\frac{n_x n_y (n_x + n_y - 2)}{(n_x + n_y)(n_x s_x^2 + n_y s_y^2)} \right]^{1/2} \qquad (B.18)$$

where $\overline{X}$, $\overline{Y}$ = means of parameters X and Y

S_x^2, S_y^2 = variances of parameters X and Y.

Step 3: If the value of the test statistic T (step 2) is less than the critical value of "t" (step 1), do not reject the null hypothesis. If the value of T is greater than the critical value of "t", reject the null hypothesis.

For further discussion see any standard statistics text (e.g., Panofsky and Brier, 1965, pp. 58-64; Till, 1974, Section 4.3).

Case B: Variances unknown and unequal. If the variances can not be assumed equal, an approximate Student's "t" test is given by Hoel (1971, p. 265). The null hypothesis and alternate hypothesis are the same as in Case A.

Step 1: Calculate the sample variances $S_x{}^2$ and $S_y{}^2$ (Equation B.5).

Step 2: Calculate the number of degrees of freedom for "t".

$$\text{d.f.} = \frac{(s_x^2/n_x)^2 + (s_y^2/n_y)^2}{\dfrac{(s_x^2/n_x)^2}{n_x + 1} + \dfrac{(s_y^2/n_y)^2}{n_y + 1}} - 2 \qquad \text{(B.19)}$$

If the number of degrees of calculated in (B.19) is not an integer, round to the nearest integer.

Step 3: Determine the critical value of "t" as in Case A, Step 1.

Step 4: Calculate the value of the test statistic

$$T = \frac{\overline{X} - \overline{Y}}{(s_x^2/n_x + s_y^2/n_y)^{1/2}} \qquad \text{(B.20)}$$

Step 5: Reject or do not reject the null hypothesis as in Case A, Step 3.

B.4.2.2 The Wilcoxen-Mann-Whitney Test.

If the distribution of variables is known to be far from normal, the Wilcoxen-Mann-Whitney test should be used. This test is discussed in Section B.4.1.

B.5 Tests for the Equality of Variances

Tests for the comparison of two variances test the null hypothesis that $\sigma^2_x = \sigma^2_y$. The alternate hypothesis is that $\sigma^2_x > \sigma^2_y$. The two tests discussed in this subsection are the F-test, a parametric test used when data closely follow a normal distribution, and a variation of Student's "t" test used when data deviate from normality.

B.5.1 The F-test for Normally Distributed Variables

The F-test should be used to test the equality of two sample variances when the distributions closely approximate normal distributions. The procedure tests the null hypothesis that the two sample variances are equal. The alternate hypothesis is that one variance is greater than the other. The alternate hypothesis results from inspection of the sample values of the two variances. For example, if we are testing for differences between the variance of the residual of the candidate model, $S^2_{d,can}$, and of the reference model, $S^2_{d, ref}$, then the alternate hypothesis would be

$$S^2_{d,\ can} > S^2_{d,\ ref}$$

For a general discussion of the test see Kreyszig (1970, Sec. 13.6) or Hoel (1970, pp. 271-273).

Step 1: From the sample results determine the value of the larger sample variance, S_x^2 and the smaller sample variance, S_y^2.

Step 2: Determine the critical value of the parameter, F, from any standard table of the F distribution (e.g., Selby, 1972, p. 620). The parameters for the table are: 95% confidence level, $n_x - 1$ = degrees of freedom for the numerator of F (greater mean square) and $n_y - 1$ = degrees of freedom for the denominator (lesser mean square).

Step 3: Calculate the value of the test statistic

$$F = S_x^2 / S_y^2$$

Step 4. If the calculated value of the test statistic, F, (Step 3) is greater than the critical value (Step 2), reject the null hypothesis. If the calculated value of F is less than the critical value, do not reject the null hypothesis.

B.5.2 Tests for Nonnormally Distributed Errors

The F-test can be shown to be sensitive to deviations from normality. Kreyszig (1970, pp. 217-218) suggests the following:

Step 1: Compute the means of the following new random variables

$$\overline{|X|} = \overline{|Xi - \overline{X}|} \quad \text{and}$$

$$\overline{|Y|} = \overline{|Yi - \overline{Y}|}$$

It can be shown that $\overline{|Y|}$ and $\overline{|Y|}$ are proportional to σ_x and σ_y respectively.

Step 2: Test for the differences between the means $\overline{|X|}$ and $\overline{|Y|}$ using the Student's "t" test as in Section B.4.2.1.

REFERENCES

Fox, D. G., 1980. Judging Air Quality Model Performance. American Meteorological Society, 60 pp., draft.

Javitz, H. S. and R. E. Ruff, 1979. Evaluation of the Real-Time Air-Quality Model Using the RAPS Data Base; Vol. II: Statistical Procedures, EPA-600/4-81-013b, U. S. Environmental Protection Agency, Research Triangle Park, N. C. 27711.

Karl, T. R., 1978. Day of the Week Variations of Photochemical Pollutants in the St. Louis Area, Atmos. Environ. 12, 1657-1667.

Kreysig, E., 1970. Introductory Mathematical Statistics, Principles and Methods, 468 pp, John Wiley & Sons, Inc., New York.

Panofsky, H. A. and G. W. Brier, 1965. Some Applications of Statistics to Meteorology, 223 pp, Pennsylvania State University, University Park, Pennsylvania.

Siegel, S., 1956. Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill, New York.

Selby, S. M. ed., 1972. CRS Standard Mathematical Tables, 20th edition, Chemical Rubber Co., Cleveland, Ohio.

Till, R., 1974. Statistical Methods for the Earth Scientist, John Wiley & Sons, New York.

APPENDIX C

JUDGING AIR QUALITY MODEL PERFORMANCE

-REVIEW OF THE WOODS HOLE WORKSHOP-

Douglas G. Fox*

1. INTRODUCTION

Atmospheric dispersion models are used to support laws and regulations aimed at protecting the nation's air resources. For this reason, models have become something more than approximations of nature designed to provide a scientifically reasonable connection between a source and receptor of air pollutants. Courts have interpreted them to be legally binding mechanisms for negotiating levels of emission control from sources. In view of this expanded role, it has become particularly critical that models be correct and be correctly applied. As a result of this need, the U.S. Environmental Protection Agency (EPA) has entered into a cooperative agreement with the American Meteorological Society to aid in the scientific and professional development and application of atmospheric dispersion models. The AMS is not involved with the regulatory process, rather our actions are motivated by a desire to advance the use of scientifically valid models.

The problem of evaluating the performance of models is among the most important facing the modeling community. To this end, the American Meteorological Society (AMS) convened a small expert working group in September 1980 to discuss current practices in model evaluation, recommend model performance evaluation measures and methods and if possible set standards for model performance. An additional task was to discuss the need for further work in this area.

2. PERFORMANCE MEASURES

Common ideas which evolved from the workshop are listed below.

2.1 Using Models for Regulatory Decisions

Models are the most rational and equitable means available to support the nation's air quality goals. It is however, the modeler's responsibility to provide "decision makers" with an estimate of the significance of model output. Where possible some statement of confidence in model results is recommended. For this to be done in a scientific and professionally acceptable manner it is necessary for the air quality modeling community to implement statistical performance evaluation.

2.2 Statistical Aspects of Air Quality Goals

Current law and regulations require that models simulate the second highest 1-, 3-, or 24-hour concentration likely to occur in a year to evaluate short term (1-hour, 3-hour, or 24-hour) standards. Since there are 8760 hours in a year for a 1-hour standard this is equivalent to predicting the 0.02 percentile (.3 percentile for 24-hour average). It is difficult to predict such a rare event with any degree of confidence. Participants recommended that models should be compared against a more robust statistic, such as the upper 2 to 5 percentile of values (something like 10-50 values). The workshop recognized that political and technical problems will need to be resolved before this recommendation can be realized.

2.3 Scientific Evaluation

Statistical performance evaluation cannot be used exclusively for determining the acceptability or unacceptability of a model. There are many scientific considerations which can provide critical input to model evaluation.

*This DRAFT SUMMARY is prepared from a workshop report currently under review. It is presented in order to provide wide distribution of model evaluation ideas in the hope of focusing discussion and encouraging work. Participants in the workshop included: D. G. Fox†, USDA Forest Service; Robert Bornstein, San Jose State U.; Norman Bowne, TRC Env. Consultants; R. L. Dennis, NCAR; Bruce Egan†, ERT; Steven Hanna†, NOAA; Glenn Hilst, EPRI; Stuart Hunter, Princeton U.; Michael Mills, Teknekron Research Inc.; Larry Neimeyer, EPA; Hans Panofsky, Pennsylvania State U.; Darryl Randerson†, NOAA; Philip Roth, Systems Applications, Inc.; Ronald Ruff, SRI International; Lloyd Schulman, ERT; Jack Shreffler, EPA; Herschel H. Slater, Consultant; Joseph Tikvart, EPA; A. Venkatram, Ontario Min. of the Environ., Canada; Jeffrey C. Weil, Martin Marietta Corp.; and Fred D. White†, Chairman AMS Steering Committee. Dr. Fox, representing the AMS/EPA Steering Group was Chairman of the Woods Hole Dispersion Model Workshop and is Chief Meteorologist, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, 240 West Prospect Street, Fort Collins, Colorado 80526.

†NOTE: AMS Steering Committee members

Not the least of these is the recognition that the atmosphere is a stochastic system and as such there are limits to its predictability. Air quality models operating within this system are limited to the degree of predictability they can attain. This in effect, provides a scientific limit to model accuracy. More effort should be expended in determination and communication of such scientific limitations for particular problems.

2.4 Data for Model Evaluation

Available data bases are not equal to the task of model evaluation. Efforts such as the EPRI-PMV, and the EPA Complex Terrain study receive a strong commendation from the participants. However, since good data are not usually available for evaluating a model, procedures to utilize existing data are needed. The value of statistical methodologies depends upon such characteristics of the data as independence and normality. Recognizing that meteorological data in general are not independent, time series analysis is an appropriate tool which should be utilized. Transformation of data to approach normality should also be considered.

2.5 Specific Performance Measures

Performance can be measured in two general ways, namely by comparing the magnitude of differences between observations and predictions and comparing the correlation or association between observations and predictions.

Magnitude differences can be expressed in terms of differences or discrepancies, defined as

$$d \equiv C_o(\underline{x},t) - C_p(\underline{x},t)$$

where C_o is an observed concentration and C_p is a predicted concentration. Three measures of magnitude difference are of importance:
(1) Estimated Bias (average) of the differences,

$$\overline{d} \equiv \frac{1}{N}\sum d$$

where N is the number of observations;
(2) Estimated noise (variance) of the differences

$$S_d^2 \equiv \frac{1}{N-1}\sum(d-\overline{d})^2$$

(3) gross variability of the differences, either as the average absolute gross error,

$$\overline{|d|} \equiv \frac{1}{N}\sum |d|$$

or as the RMS error,

$$RMS_d \equiv \frac{1}{N}\sum d^2$$

The RMS_d error is related to the variance and bias since

$$RMS_d^2 = \left(\frac{N-1}{N}\right)S_d^2 + (\overline{d})^2$$

Workshop participants recommended that these measures be applied to
(A) total Fields of Differences,

$$C_o(\underline{x},t) - C_p(\underline{x},t)$$

for all $\underline{x}, t$

(B) Selectively paired maximum values of the differences, for example, where L(n), T(n) are the coordinates for highest (n=1), second highest (n=2), etc. concentrations, then

$$d_M(L_{(n)},T_{(n)}) \equiv C_o(L_{(n)},T_{(n)}) - C_p(L_{(N)},T_{(n)})$$

paired in both space and time;

$$d_M(L,T_{(n)}) \equiv C_o(L_{(n)},T_{(n)}) - C_p(L_j,T_{(n)})$$

paired in time but Lj is the location of maximum predicted concentration at T(n);

$$d_M(L_{(n)},T) \equiv C_o(L_{(n)},T_{(n)}) - C_p(L_{(n)},T_j)$$

paired in space but Tj is the time of maximum predicted concentrate at location L(n), and

$$d_M(L,T) \equiv C_o(L_{(n)},T_{(n)}) - C_p(L_k,T_j)$$

unpaired since Co and Cp are simply the maximum observed and predicted values without regard to time or space.

(C) Totally unpaired comparisons of frequency distributions of observations with frequency distributions of predictions can be conducted by using statistical methods of comparison for the bias (t, z, Wilcoxon/Mann-Whitney statistics), for the variance (F or χ^2 Statistics) and for the gross variability (χ^2 or Kolmogorov-Smirnov statistics)

Correlation can be measured by considering the data paired for the entire field as discussed in (A) above and selectively paired for maximum values as discussed in (B) above. Correlation is measured by the correlation coefficient r_{op}, defined as

$$r_{op} \equiv \frac{\sum(C_o-\overline{C}_o)(C_p-\overline{C}_p)}{\left[\sum(C_o-\overline{C}_o)^2\cdot\sum(C_p-\overline{C}_p)^2\right]^{1/2}}$$

where the overbar is as defined above. Three specific ways of considering correlation were suggested:

(1) Temporal $r(\Delta T)$ the cross correlation coefficient, where the C's are paired at a particular location in space as in (A) above and separately as in (B) above. All time lags ΔT between observation and predictions including $\Delta T = 0$ can be considered.

(2) Spacial r_s the spacial correlation coefficient, using C's paired at particular times for the entire field as (A) above.

(3) Coupled, r the correlation coefficient using C's for the entire field.

Various other statistics are suggested to compare, for peak values, the displacement in time and space of observations from predictions.

2.6 Qualifications on Performance Measures

The measures suggested will likely prove most meaningful when applied to large data sets more typical of urban problems and tracer studies conducted for limited time periods. Difficulties in measuring the performance of point source models exist because the concentration pattern resulting from such a source has very sharp gradients. Generally, the peak concentration which is routinely calculated (center line value) is not measured. Special considerations recommended for the point source problem, therefore, include data preprocessing for wind direction and possibly stability. Further it should be realized that as a community we have only very limited experience with many of the measures of performance. It will, therefore, require some time to realize the significance of them.

3. PERFORMANCE STANDARDS

Participants agreed that it was unrealistic to attempt to establish standards at this time. There is an overriding concern that criteria for setting standards are not available. Just how accurate must a model be for the various regulatory applications? Secondly, there is a conspicuous lack of experience with existing models tested against performance measures such as we recommend. Finally, data bases of high enough quality to be capable of discriminating between performance of various models are not abundant. More good data must be collected.

In spite of these concerns the participants did develop two recommendations related to judging air quality models.

3.1 Statistical Inference Testing

Statistical tests such as the Student's t for means and the χ^2 for variances can be utilized to establish confidence intervals about the calculated values of performance measures. This allows a quantitative indication of a model's validity. It was, however, recognized that often such statistical testing is of very limited value because it is based upon close adherence of the data to theoretical distributions. The point is that such tests mav suggest that model results are less believable than in fact they should be.

3.2 Develop Performance Profiles Referenced Against EPA Guideline Models

The performance of models recommended in the Guideline could provide a reference value for comparison of other models. The reference concept, however, to some participants, implied that the Guideline Models are "good enough" while in fact they may not be. For this reason it was suggested that the reference be considered like 0°F—namely an arbitrary number representing nothing physically significant, but one against which other temperatures can be quantified. At any rate it seems appropriate to develop profiles of performance for models by comparing performance against what are currently accepted regulatory procedures.

4. RESEARCH NEEDS

The workshop participants recommended five specific areas in which research is needed. They are (1) development and refinement of performance measures; (2) application of performance measures to (especially) point source models; (3) analysis of the characteristics of meteorological data; (4) analysis of the characteristics of air quality data; and (5) the evaluation of diffusion models. In addition to these specific tasks it was reitterated that much better data are needed. Data collection with special field programs, for example, can be quite expensive. The cost, however, is small compared to the amount of money expended on pollution emission controls and, therefore, on implications resulting from the applications of models. It is possible that this data will show how poorly we are able to predict concentrations. They may result in a major new round of research into the fundamental physics and chemistry of the atmosphere.

5. CONCLUSIONS

How shall we judge the performance of air quality simulation models? The AMS/EPA Woods Hole Workshop was convened in part to focus the attention of the professional air quality modeling community on this important task. Although the workshop may raise more questions than it answers, a few ideas have emerged which are described in this short summary.

A set of statistics which can provide a rational framework for quantitatively evaluating the nature of differences between observations and predictions by models are proposed. Statistics are suggested as a tool to provide confidence in model predictions as well as to compare new models against those models recommended by EPA in their "Guidelines". The task is not complete. We have only extremely limited experience with these measures of performance. A recommendation is, therefore, to test models using the framework suggested in this paper and detailed in our forthcoming report. Only through such experiences will it be possible to learn the most appropriate procedures for evaluating models.

It was a strong feeling of the participants that statistical measures alone may not be sufficient to judge between models. Scientific evaluations based upon accepted laws of physics will always provide a good basis for critiquing models.

Finally, it was unanimously agreed that data on which to evaluate models is lacking. The collection of good data must remain a high priority activity for air pollution modelers.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711

DATE: 7/30/81

SUBJECT: Role of Models in Regulatory Decision-Making

ROM: Joseph A. Tikvart, Chief J. Tikvart
Source Receptor Analysis Branch (MD-14)

TO: Chief, Air Programs Branch, Regions I - X

As you are aware, OAQPS sponsored a workshop on the role of atmospheric models in regulatory decision-making. The workshop was held in May 1981 at Airlie House. A summary report has been prepared and distributed which will serve as the focal point for the modeling conference. We have previously communicated with you concerning both the report and the conference.

Section 4 of the summary report (see attachment) provides recommendations on actions that EPA can take to better reflect the uncertainties of air quality model estimates in its regulatory decisions. Many of these recommendations require further study, technical development, coordination, and review of current policies. Some of the recommendations, if implemented, could have a direct effect on Regional Office and State procedures for SIP revisions and the review of new sources, as well as resources required for these programs.

We anticipate that the summary report will be well received and widely endorsed at the modeling conference. It suggests a flexibility that many in the industrial and regulatory communities believe is necessary to relieve the current regulatory climate which is perceived to be overly stringent. We want to seriously explore these recommendations and their ramifications.

The purpose of this memo is to solicit your views on the summary report recommendations, in particular those that could directly affect Regional Office and State programs. To this end, several subsections of the attachment are marked for your attention. These subsections deal with (1) planning meetings and criteria for model selection; (2) development of protocol documents; (3) use of "arbitration panels," and (4) more explicit consideration of model uncertainty in decisions. To what extent are these issues factored into current Regional Office and State programs? How would you implement the recommendations? What modifications to current programs would be required? What problems and benefits would be created? What would be the effect on resources and the timeliness of reviews?

Other portions of Section 4 provide observations and recommendations concerning (1) screening, long-range transport and complex terrain models; (2) performance evaluation of models; (3) design concentrations; (4) modifications to modeling guidelines; (5) a modeling "center" and (6) a quality assurance program. We also solicit any comments you might have on these issues.

It would be appreciated if I could have your views on the workshop recommendations, either verbally or in writing, by the end of August. Please contact me if you have any questions.

Attachment

cc: R. Campbell
T. Helms
C. Hopper
R. Rhoads
R. Smith
B. Steigerwald
Modeling Contacts, Regions I - X

Workshop Summary Report

ROLE OF ATMOSPHERIC MODELS IN REGULATORY DECISION-MAKING

EPA Contract No. 68-01-5845

July 1981

Prepared for

Charlotte Hopper
Source Receptor Analysis Branch
Office of Air Quality Planning and Standards
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

Prepared by

C. Shepherd Burton
Systems Applications, Inc.
101 Lucas Valley Road
San Rafael, California 94903

4 SUMMARY OF WORKSHOP FINDINGS AND RECOMMENDATIONS

4.1 OVERVIEW

In the course of independently addressing the four overall questions, each workgroup was required to develop an approach to its specific problem and the criteria by which issues and needs could be identified and recommendations made. Their efforts were separately documented through on-site reports written and edited by the workgroup participants. These three documents, which represent three separate reports, are provided as appendixes to the final report. Workgroup I addressed the four questions using the PSD permitting problem as a vehicle for examining alternative answers; Workgroup II's problem focused on SIP revisions; and Workgroup III explored the workshop questions in light of concerns over the transport of pollutants across political boundaries.

The principal findings and recommendations of the individual work group reports are integrated and presented in this closing section of the workshop summary report. Attention is given to the needs identified by the workgroups and to the procedural or process changes recommended by groups for the consideration or implementation of the EPA's Office of Air Quality

Planning and Standards. Particular emphasis is given to needs and recommendations for

(1) Assuring the wide acceptance by interested parties and by the public of modeling in air quality management.

(2) Identifying the factors affecting the needed balance between standardization, consistency, and flexibility in model selection and application and in the interpretation and presentation of model results in AQM decisions.

In keeping with the basic workshop structure and the structure of the reports, the summary findings and recommendations are organized according to the four general workshop questions. At the end of this section, some additional proposals are advanced and some concepts that were recommended by the workgroups are extended. Although some of these proposals were not explicitly mentioned by any workgroup, they appear to be consistent with the needs and recommendations provided in the workgroup reports.

In light of the broad cross section of skills and interests of the participants, it is worth noting the harmony within, and among, the groups concerning the needs identified and the recommendations presented. This harmony is reflected in both the specifics and the spirit of workgroup findings and recommendations. Although at the outset of the workshop, participants were informed that a consensus view was not sought, it appears that by at least one measure--the level of harmony--consensus was achieved.

4.2 CRITERIA PERTINENT TO THE APPROPRIATE SELECTION OF AN AIR QUALITY MODEL

Whether the model is intended for use in a PSD permitting effort, a revision of a SIP, or in policy setting, all workgroups endorsed the concept of early, open, and cooperative participation in model selection by all affected and interested parties. A model selected in this manner, using the additional criteria presented next, is likely to be supported and accepted by not only the regulatory agencies and industry being regulated, but also by interested labor, civic, and environmental groups.

Other selection criteria recommended by the workgroups included

(1) A suitable match between (a) the technical attributes and capabilities of the selected model, (b) the operational requirements of the selected model, and (c) the air quality issue(s) of concern.

(2) A suitable means for estimating, evaluating, or examining the uncertainty associated with model predictions.

(3) A means for addressing and satisfying consistency requirements and concerns of equity with respect to prior use of the same or similar models and similar air quality issues.

Throughout the model selection process, resource constraints were also acknowledged by the workgroups as warranting consideration. However, the groups recommended that such considerations be given after one or more modeling alternatives has met the other (technical) criteria. That is, apparently the resource constraint criteria should principally serve to distinguish between technically acceptable alternatives.

Within each of these criteria, the workgroups identified many detailed criteria for consideration in model selection, including (1) the spatial and temporal scales of the problem, (2) defensible treatments of recognized important physical and chemical atmospheric processes, (3) the spatial and temporal representativeness of the meteorological data and record, (4) emissions from individual and interacting sources, including variability, (5) the match between the data needs of the model and the availability of input data, (6) the compatibility of the model's output(s) with the requirements of the ambient standard, increment, or air quality goal, (7) documentation of the model's algorithms, computing requirements, computer software, test/example cases, and prior evaluation activities and regulatory applications, and (8) simplicity, adaptability, and flexibility in its transferability to different geographical settings, emission source configurations, and (possibly) political boundaries. Appendixes to the final report provide additional selection criteria and corresponding discussions about each.

A special mention is in order regarding the selection of screening models and long-range transport models in PSD new source reviews. It was noted by Workgroup I that though screening models may originally appear to ease and simplify the permitting process, they subsequently could cause complications involving equity issues, including the following:

(1) Premature determination of increment consumption, which in turn can elicit from potential industrial developers a

variety of responses that may complicate subsequent permitting actions.

(2) Predatory actions by industrial developers, including attempts to bank the increment and tactics to discourage acquisition of adjacent development sites, and so on.

(3) Distortions in the time phasing of industrial develop ment to ensure being one of the first developers of a region.

(4) Inequities in Best Available Control Technology (BACT) determinations.

These concerns will be of particular importance in areas of concentrated development, such as the oil shale area. In one respect, the use of screening models in areas of potential concentrated development can confuse air quality management decisions and planning, since such models do not and cannot (because they are not intended to) provide a reliable measure of the consumption of the air resource. Answers to questions about the ultimate potential development cannot be addressed by either industrial developers or government policy-makers using these models. Furthermore, the use of a multiplicity of models--screening, guideline, and nonguideline--can cause even greater complications. The requirement for consistency and standardization in such situations appears paramount.

The requirement for consistency was noted to be important in another instance. This circumstance involves the use of long-range transport models--also in PSD new source review but also, possibly, in SIP revision actions, assuming the transport of pollutants across state boundaries is of concern. In selecting such models, the principal consideration must be given to the soundness of the scientific principles upon which they are based and the implementation of those principles in the model, since empirically based model evaluations are probably several years away.

The balance between flexibility, standardization, and consistency is particularly vexing in regions of complex terrain. The likely wide variations in meteorological, topographical, and source configurations, when combined with the absence of generally accepted modeling approaches, suggests a strong need for flexibility in selecting a modeling approach--especially in PSD and SIP revision actions. However, in regions of concentrated development, standardization and consistency in model selection are also required to reduce the potential for inequities between sources and to reduce administrative burdens. Workgroup I suggested that for a particular geographical region, a requirement for performance evaluation of a proposed nonguideline model, using an available data base having suitable similarities to the impact assessment of interest, would likely impose the nécessary (regional) consistency and also allow the desired flexibility. Questions involving the characterization and specification of such similarity criteria were not addressed by Workshop I (e.g., what these criteria could/should be and who should identify and specify them).

4.3 CRITERIA PERTINENT TO THE APPROPRIATE USE OF AN AIR QUALITY MODEL

The previous subsection addressed the criteria that decision-makers and modelers should adopt in selecting from a set of available models the single model or subset of models to be used in a particular situation. This subsection focuses on the workgroup-recommended principles that should structure the process by which the model is agreed upon and set up, the input data prepared, the runs made, the output formulated, and the entire process documented.

All workgroups recommended that all models used in air quality regulation undergo standarized performance evaluation according to the Woods Hole recommendations. In those instances where a bonafide dispute exists concerning the suitability of a particular application of an EPA-recommended model, an application-specific performance evaluation was recommended by Workgroup II as the preferred means of resolution. Workgroup II recommended that efforts be made to develop minimum acceptable levels of model performance (i.e., standards). Furthermore, Work group II recommended that models be required to meet some minimum level of performance prior to acceptance for regulatory use; however, the workgroup also recognized that an explicit level of performance cannot currently be specified.

All the workgroups recommended and endorsed the concept of instituting a protocol concept in the use of models. Such a protocol would be developed through open, cooperative meetings between the regulator and other interested parties prior to the use and application of a particular model or set of models. The workgroups also noted that efforts should be made to identify in advance, to the greatest extent possible, the specifics regarding modeling procedure, including as needed (a) model performance evaluation methods (e.g., measures, standards, and so on), (b) data sources, uses, and adjustments, (c) model computation and parameter selection options, (d) receptor selection, (e) model output formats, (f) interpretation and, if necessary, adjustment of model results, (g) identification of model limitations and biases, and (h) the examination of the likely effects of model limitations and uncertainties on the estimated impacts (e.g., through sensitivity analysis or Monte

Carlo simulation). Much of the recommended information would be available from user manuals and guidelines for the model so that a protocol document need not be an extensive volume; rather, it should be a substantive one.

The workgroups also recommended that during these meetings every reasonable effort be made by all parties to identify potential uncertainties and conflicts and to identify the means to resolve them. Among the approaches suggested for resolving conflicts were the establishment of technical review committees composed of interested parties, or the designation of project arbitrators. Their judgment of both would be final. These and other possible approaches to reconciling disputes or conflicts are recommended for inclusion in the protocol document. The information that can be used during the resolution of a dispute, as well as limits governing its use, should also be identified and specified in the protocol.

It appeared to the workgroups that a natural balance can be found between flexibility and consistency with the institution and practice of a protocol concept. It was noted by Workgroup II that when a high degree of flexibility is being sought, all procedures should be agreed upon a priori by all interested parties and that a suitable forum (ut sup.) should be identified and established to resolve anticipated or unanticipated issues.* Furthermore, this practice would facilitate the recommendations of Workgroup III that any party or decision-maker who uses a model and seeks to base a regulatory decision on the output of that model is obligated to (a) publicly document the input data and actual model used, and (b) reproduce the methods/techniques employed in preparing input data and model exercise.

4.4 INCORPORATING AIR QUALITY MODELING UNCERTAINTIES INTO THE DECISION-MAKING PROCESS

All workgroups endorsed the need, and recommended that approaches be sought, to identify, quantify, reduce (if possible), and incorporate the uncertainties associated with air quality modeling into the regulatory and decision-making process. Furthermore, all the workgroups recommended that priority be given to (a) the identification and quantification

* A priori means here that procedures should be specified prior to obtaining, or being able to infer, the final result of the impact assessment.

of uncertainties and (b) the incorporation of such uncertainties into the regulatory process. The workgroups recognized that the rate of reduction of modeling uncertainties, through model and data input improvements, generally occurs more slowly than the rate at which such information is needed in the regulatory setting. In addition, such a prioritization presumes that technically sound modeling is being practiced (i.e., a best available modeling approach is selected and properly used). The workgroups also recommended that model research and development be continued.

In recommending that modeling uncertainties be reflected in regulatory decisions and in the exploration of alternative control strategies, Workgroup III noted that it should be the *modeler's* responsibility to the decision-maker to identify, describe (when possible), and quantify the sources of uncertainty in each air quality analysis. In addition, it should be the *modeler's* responsibility to express modeling results in a manner that clearly communicates the uncertainty in an understandable and utilizable form to the decision-maker and the decision-making process. Workgroup III also noted that it should be the *decision-maker's* responsibility to become knowledgeable concerning, and conversant with, results that express and contain uncertainty. Determining how best to utilize the results presented with their attendant uncertainty was noted by Workgroup III to be the responsibility of the decision-maker.

Furthermore, for AQM and the regulatory process to ignore modeling uncertainty and to continue to base decisions on best estimate single-value measures, such as the high, second-high concentrations, places an unduly heavy burden on modelers, who essentially are being required to make, or are implicitly making, policy decisions when they select models and choose model inputs.

The workgroups recognized that modeling uncertainty can be incorporated into the decision-making process by

(1) Developing procedures for quantifying uncertainty.

(2) Giving attention to the strengths and weaknesses of modeling in fashioning the measures of achievement for air pollution control programs.

(3) Explicitly describing the uncertainties (and their likely implications, if known) that cannot be eliminated.

With respect to item (2), Workgroup I noted that uncertainty could be reduced substantially and rather quickly if measures other than the high, second-high were employed, or if the high, second-high measure were augmented with additional information available from a model. Another measure Workgroup I identified was the 95th percentile value of the distribution of ground-level concentrations. This workgroup noted, however, that the selection of such a concentration value, if chosen to be consistent with current practice, could raise equity issues vis-a-vis individual or groups of sources, with the former possibly leading to more favorable outcomes than the latter. Workgroup II noted that an alternative approach to using the high, second-high concentration value would be to calculate the 95th percentile concentration and then extrapolate the resulting value to the percentile corresponding to the high, second-high value.

Workgroup I also identified additional information that currently available models could readily provide to decision-makers, including the

(1) Number of times concentration values exceed 80 or 90 percent of the standard/increment.

(2) Average of the 10 highest concentration values at the worst receptor.

(3) Episodic character of the highest concentration values (i.e., the extent to which such values are uniformly distributed throughout the year or are grouped together).

(4) Location and extent of the geographic area where standards/increments are most likely threatened.

(5) Exposure or dosage estimates. Workgroup I, in suggesting the use of such information by decision-makers, recognized that significant changes in both the current values embodied in the clean air legislation and the regulatory process are necessary.

All workgroups recommended that the strengths and limitations of models be examined in light of the need to incorporate the resulting understanding into the design of a decision-making process that

(1) Reduces the sensitivity of decisions to model uncertainties.

(2) Seeks to manage the risk of incorrect decisions.

One or more of the workgroups provided the following recommendations regarding the explicit incorporation of uncertainties into decision making:

(1) Make uncertainties explicit, through the best available means, in all modeling-related decisions. As appropriate, use data from site-specific performance evaluation studies, use the understanding of departures from underlying model assumptions, and use the results of sensitivity analyses. For the immediate future, sensitivity analyses and Monte Carlo simulations probably represent the only available approaches for providing uncertainty estimates for medium- and long-range transport models. In effect, the workgroups recommend providing decision-makers with estimates of error bars on model estimates. It was noted that sensitivity analyses are likely to provide lower estimates of uncertainty.

(2) Use confidence bounds (i.e., error limits), or preferably, probability distributions to express uncertainties.

(3) Continue the process already started with the ExEx and Multi-Point Rollback (MPR) methods of incorporating probabilistic concepts into the modeling framework and into convenient and understandable formats for use by decision-makers. Examples noted included using expected exceedance, violation probability, and Type I and Type II error approaches.

(4) Develop structures and models for the decision process itself to provide a basis for accommodating and analyzing model uncertainty in the overall process, with its attendant uncertainties. In effect, develop a mathematical framework for decision analysis in air quality management.

Workgroup I also noted that additional flexibility is probably needed in the decision process, especially with respect to PSD permitting issues, to reflect the various purposes/goals of the PSD provisions regarding air-quality-related values.

4.5 INCORPORATING IMPROVEMENTS IN AIR QUALITY METHODS INTO THE REGULATORY PROCESS

The previous subsections have focused on the selection and use of models and the interpretation of model results and their attendant uncertainties in a somewhat static regulatory environment--one in which no explicit recognition is given to either the evolutionary (e.g., through the introduction of new dispersion coefficients) or revolutionary (e.g., through the introduction of visibility impairment models) nature of developments in modeling methods.

The fourth question explored by the workshop recognized that air quality modeling is a rapidly expanding, evolving, and advancing field whose raison d'etre is to identify, address, meet, and serve the needs of air quality managers, policymakers, and decision-makers. The growth of this field and its potential contributions can be estimated by considering the increased number of technical conferences, publications, organizations that sponsor and perform research, and organizations (including state and local agencies) that provide services in air quality modeling and related activities. The approaches to encouraging, controlling, and facilitating the embodiment of the most suitable methods and data bases in the regulatory process are still to be developed. The publication of air quality modeling guidelines by the EPA represents an initial effort toward attaining this goal. In this subsection, the recommendations of the workshop vis-a-vis the introduction of modeling improvements into the regulatory process are provided.

All workgroups noted that the issue related to consistency, standardization, and flexibility lay at the core of this overall problem. All workgroups either explicitly or implicitly recommended that consistency should be achieved by selecting and using a new or modified approach rather than by insisting that the same guideline or nonguideline model be used for all circumstances.

Although each workgroup emphasized somewhat different elements of the process for achieving this goal, all the groups recommended that

(1) Improvements be made in the methods used to convey changes in models, methodology, and processes to interested participants.

(2) Consideration be given to the establishment of a concept to provide for the centralization of certain modeling activities and to provide some insulation of the technical modeling tasks from the political decision-making process.

In addition, Workgroups I and III noted that such a center would require extensive peer review and technical oversight, a recommendation also implicitly recommended by Workgroup II. The remainder of this subsection elaborates on the nature of these recommendations.

Workgroup II recommended that the regular updating of modeling guidelines constitutes the most reasonable means of conveying changes. Recommendations were not advanced regarding the frequency of updates, though the criteria recommended for their selection were the significance and acceptability of the change to the technical community. Workgroup II also recommended that procedures be instituted to establish criteria for change and to communicate methodologies, practices, and so on, to the community of practitioners and other interested parties.

In addition, Workgroup II recommended the possibility of adopting a regular schedule for revisions, even if the announcement at the scheduled time were only that no significant revisions were expected during the subsequent interval.

In making its recommendations for guideline revisions, Workgroup II acknowledged the importance of proposed model changes on past regulatory actions, along with the implications of proposed model changes for future regulatory actions. Thus, this workgroup recommended the establishment of a function within the EPA for dealing with the implications of new methods or practices. This function would address, in advance, the policy, legal, and regulatory issues raised by any proposed changes and would recommend methods for resolving such issues. Workgroups II and III recommended that strong consideration be given to grandfathering affected facilities, provided that past modeling efforts had been carried out in good faith.

Recommendations varied regarding the scope and function of the modeling center concept. The responsibilities identified by one or more workgroups of such a function included

(1) Maintenance and updating of model costs.

(2) Maintenance of test data bases.

(3) Undertaking model performance evaluation studies and archiving their results.

(4) Maintenance of a repository for all actions involving nonguideline models.

(5) Maintenance of information concerning model application results.

(6) Provision of certain defined services for selected modeling studies, including third-party model exercise in some cases and the exercise of models whose costs or technical requirements, either in the form of expertise or hardware demands, are extensive.

As noted, the modeling center would of necessity require extensive peer review and technical oversight and, thus, the workgroup recommended an advisory or review committee as the preferred means of reaching consensus and according legitimacy to proposed changes in modeling practices. Such a body would be composed of both government and nongovernment representatives having backgrounds in policy and technical areas. This committee would periodically review proposed revisions to the guidelines originating from, say, the modeling center. The committee would also review and comment on the suitability of new modeling techniques and advances in modeling practice.

4.6 SOME LOGICAL EXTENSIONS OF WORKGROUP REPORTS, AND FURTHER RECOMMENDATIONS

The previous sections have attempted to report with acceptable fidelity the recommendations of the workgroups. The similarities and the lack of dissimilarities between the recommendations and their possible implications for additional recommendations can be clearly noted. This subsection provides comments and recommendations resulting from the efforts to integrate the conclusions of all the workgroups.

First, dissimilarities among workgroup's recommendations, either in specifics or in spirit, despite the disparity among workshop participants, were not noticeable. This does not mean that areas of disagreement do not exist or that disagreements did not occur. It does appear to mean however, that in areas involving the practice of air quality modeling there is much room for agreement. Further, it may also mean that the majority of participants see the practice of air quality modeling and the AQM approach as the preferred way to accomplish clean air goals.

Second, the nexus for efficiently achieving clean air objectives through the AQM approach and impact assessment lies in establishing and preserving a balance among flexibility, standardization, and consistency. More effort needs to be devoted to defining the dimensions of this issue and the parameters that will secure and assure the continuance of that balance in the regulatory setting. It appears that many of the essential elements for dealing with this issue were identified by the workshop participants:

(1) Utilization of cooperative processes, whenever possible, that provide for early and substantive involvement of interested parties and that encourage the anticipation, definition, and resolution of potential areas of conflict. Such processes can be expected to accord broad acceptance and

legitimacy to the results.

(2) Utilization of air quality (or increment consumption) assessment plans or protocols to identify and define methods, tasks, analysis steps, data bases, potential disputes (and the means for resolving them), and the schedule (including the times for periodic meetings) for accomplishing the impact assessment.

(3) Utilization of advisory groups to provide oversight, guidance, and peer review.

It is recommended that, whenever appropriate, the EPA provide guidance to the regions and states regarding the purpose and role of the foregoing elements and that in the case of item (2), the agency provide guidance, by way of examples, regarding the use of such plans or protocols.

An element not identified by the workgroups, but which appears to be implicit in, and consistent with, their recommendations involves the establishment of quality assurance (QA) in the AQM process. Offered here as an additional recommendation, this function would attempt to reduce doubts and risks concerning the modeling methods employed in, and conclusions derived from, air quality impact assessments. At least two basic activities are recommended for a QA activity: (1) certification, and (2) evaluation.

Certification would be primarily directed at verifying, among other possibilities, the correctness of the impact analysis against established accepted practice, model design, and user manual specifications. Certification of an organization's capability to offer impact assessment services could also be provided. The certification would follow a rigorous test plan established in advance. The result of this activity would be either acceptance (certification) or rejection (no certification); the basis for the rejection and the deficiencies would be noted. Further attention is needed to designate the entity responsible for setting standards, defining the test plan, and other related activities. A quality assurance board composed of professional organization, government, and nongovernment membership that encompasses a broad range of skills and interests could be a part of this function.

The evaluation activity would be mainly directed at examining items of concern that have been identified at some point during the impact assessment. This activity follows an investigative approach; problems or issues that are raised are examined for their importance to, and effect on, a particular outcome. The correctness of the method or methods in question would be evaluated, and the effect of using the method(s) on a particular outcome would be assessed. The result of the evaluation would, at least, be brought to the attention of the decision-maker or other interested participants in the impact assessment.

It is recommended that further consideration be given to the scope and use of a QA activity, especially in relation to the

(1) Function(s).

(2) Elements of impact assessments to be included.

(3) Establishment of standards, and their relationship to acceptance tests and independent verification and validation.

(4) Need for oversight.

(5) Roles and types of audits.

(6) Documentation requirements.

(7) Requirement for reducing administrative and resource burdens on all parties and for preserving cost-effectiveness.