United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park, NC 27711

EPA-454/R-92-025
September 1992

AIR

EPA

PROTOCOL FOR DETERMINING THE BEST PERFORMING MODEL

EPA-454/R-92-025

Protocol for Determining the Best Performing Model

U. S. ENVIRONMENTAL PROTECTION AGENCY
Office of Air Quality Planning and Standards
Technical Support Division
Research Triangle Park, NC 27711

December 1992

DISCLAIMER

This report has been reviewed by the Office of Air Quality Planning and Standards, EPA, and approved for publication. Mention of trade names or commercial products is not intended to constitute endorsement or recommendation for use.

CONTENTS

FIGURES

TABLES

1.0 BACKGROUND AND PURPOSE

EPA has an extensive ongoing program to evaluate the performance of air quality simulation models. The program has resulted in a standard set of statistical and graphical procedures for analyzing and displaying the performance of models, both from an operational and scientific viewpoint.[1,2,3,4] Application of these procedures has produced considerable information about the performance of models. While the information has provided numerous insights into model behavior, it is not entirely suitable for comparing the overall performance of the models. Model comparisons are difficult because the statistics that were generated cannot be easily composited and also because methods for calculating confidence limits for complex statistics were unavailable at that time.

Since these studies were published, advances have been made in the statistical methodology needed to compare the performance of models.[5] With this newer methodology, it is feasible to combine results from different averaging periods and different data bases into a probabilistic framework. For example, it is possible to make statements such as "The overall difference in performance between model A and model B is X units and this difference is significant at the 95 percent confidence level". The purpose of this document is to present a statistical method for comparing the performance of models using the statistical techniques that are now available.

EPA has also published a document which describes a process for selecting a best model for case-by-case regulatory applications.[6] The document, "Interim Procedures for Evaluation of Air Quality Models", provides guidance in areas such as selection of

the data bases, the statistics, the performance measures, and the weights to be given to each evaluation component. The statistical procedures discussed in the interim procedures document are still believed to be sound. However, if computer resources are available, the newer statistical procedures described in this document (see also reference 7) may be more appropriate.

The statistical approach for determining which model(s) perform better than other competing models involves two steps. The first step is a screening test to eliminate models that fail to perform at a minimum operational level. Although a completely objective basis for choosing a minimum level of performance is lacking, accumulated results from a number of model evaluation studies suggests that a factor-of-two is a reasonable performance target a model should achieve before it is used for refined regulatory analyses. The second step applies only to those models that pass the screening test. The analysis is based on a computer intensive resampling technique known as bootstrapping which generates a probability distribution of feasible data outcomes. The outcomes are used to calculate a composite measure of performance that combines information among averaging periods, data bases and integrates both the scientific and operational components of model performance. Comparison of the distributions of the composite measures of performance for each pair of models provides evidence of the degree to which one model performs better than other competing models.

Appendix A provides an example application of the protocol using data from six large mid-western power plants to compare the performance of two rural air quality models.

2.0 SCREENING TEST

Each competing model is subjected to a screening test to determine if it meets minimum standards for operational performance. The fractional bias is used as the performance measure. The general expression for the fractional bias (FB) is given by:

$$FB = 2\left[\frac{OB - PR}{OB + PR}\right]$$

The fractional bias of the average is computed using this equation where OB and PR refer to the averages of the observed and predicted highest 25 values. The same expression is used to calculate the fractional bias of the standard deviation where OB then refers to the standard deviation of the 25 highest observed values and PR refers to the standard deviation of the 25 highest predicted values.

The fractional bias has been selected as the basic measure of performance in this evaluation because it has two desirable features. First, the fractional bias is symmetrical and bounded. Values for the fractional bias range between -2.0 (extreme overprediction) and +2.0 (extreme underprediction). Second, the fractional bias is a dimensionless number which is convenient for comparing the results from studies involving different concentration levels or even different pollutants.

Figure 1 is a graphical illustration of model performance in which the fractional bias of the standard deviation (y-axis) is plotted against the fractional bias of the average (x-axis). Models that plot close to the center of the graph (0,0) are relatively

free from bias, while models that plot further away from the center tend to over or underpredict. Values

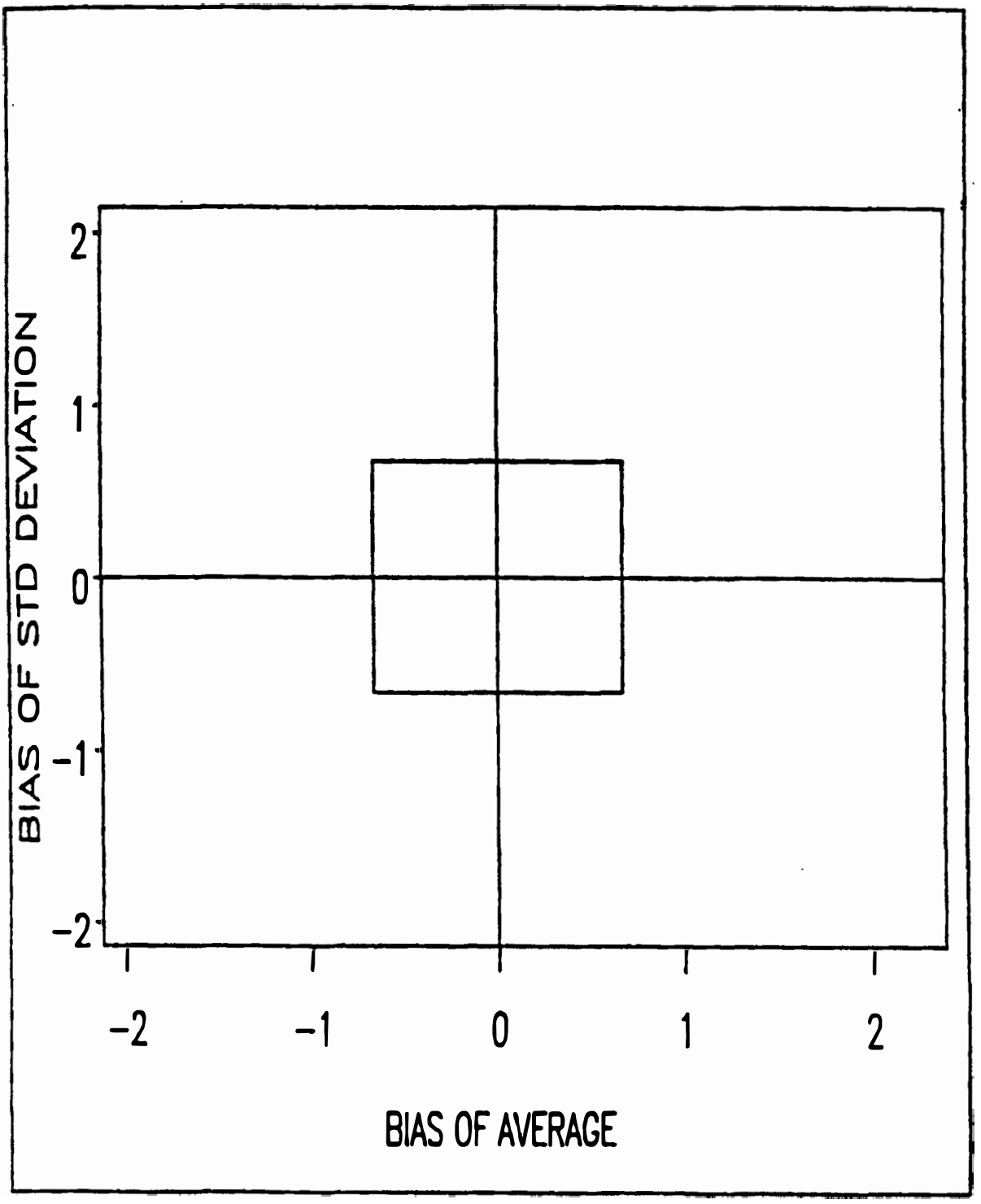

Figure 1. Example fractional bias plot for a given averaging period, e.g., 3-hour.

of the fractional bias that are equal to -0.67 are equivalent to overpredictions by a factor-of-two, while values that are equal to +0.67 are equivalent to underpredictions by a factor-of-two.

Since neither underprediction nor overprediction is desirable, the calculations are simplified by determining the absolute fractional bias (AFB) which is the absolute value of the fractional biases computed for the average and the standard deviation. The absolute fractional bias is calculated for each averaging period for which ambient standards or goals have been established. Separate calculations are made using information from each available data base. If the computed AFB statistic tends to exceed the value of 0.67 for any averaging period or data base, consideration may be given to excluding that model from further evaluation due to its limited credibility for refined regulatory analysis.

3.0 STATISTICAL TEST

Models that pass the screening test are then subjected to a more comprehensive statistical comparison that involves both an operational and scientific component. The rationale for the operational component is to measure the model's ability to estimate concentration statistics most directly used for regulatory purposes. For a pollutant such as SO_2 for which short-term ambient standards exist, the statistic of interest involves the network-wide highest concentrations. In this example, the precise time, location and meteorological condition is of minor concern compared to the magnitude of the highest concentrations actually occurring. The scientific component is necessary to evaluate the model's ability to perform accurately throughout (1) the range of

meteorological conditions that might be expected to occur and (2) the geographic area immediately surrounding the source(s) for which model estimates are needed.

Because of the emphasis on highest concentrations, a robust test statistic is calculated that represents a "smoothed" estimate of the highest concentration.* A performance measure, based on the fractional bias, is calculated which compares the air quality and model test statistics. The performance measures obtained from the operational and scientific components and from among the various data bases are combined to create a composite performance measure. The bootstrap procedure is used to estimate the standard error for the composite performance measure for each model. Using the estimate of standard error obtained from the bootstrap, the statistical significance of the difference between models is assessed.

Test Statistic

The test statistic used to compare the performance of the models is a robust estimate of the highest concentration (RHC) using the largest concentrations within a given data category. The same robust estimator is used in both the operational and scientific phases of the statistical comparison. The robust estimate is based on a tail exponential fit to the upper end of the distribution and is computed as follows:[7,8]

---

*Typically, the network-wide highest value from among the annual second highest concentrations at each monitor is used to determine attainment/non-attainment of ambient standards. Because the highest concentration value is subject to extreme variations, the robust highest concentration is preferable in this analysis because of its stability. Also, the bootstrap distribution of robust highest concentrations is not artificially bounded at the maximum concentration, which allows for a continuous range of concentrations that in fact may exceed the highest concentration.

$$RHC = X(N) + [\bar{X} - X(N)] \ln [\frac{3N - 1}{2}] ,$$

where:

$\bar{X}$ = average of the N-1 largest values
X(N) = Nth largest value
N = number of values exceeding the threshold value ($N \leq 26$)

The value of N is nominally set equal to 26 so that the number of values averaged ($\bar{X}$) is arbitrarily 25. The value of N may be lower than 26 whenever the number of values exceeding the threshold is lower than 26. Whenever N is less than 3, the RHC statistic is set equal to the threshold value where the threshold is defined as a concentration near background which has no impact on the determination of the robust highest concentration.

The robust estimator of the highest value is strongly related to the two statistics used in the screening test. Increasing values of the average and standard deviation have the effect of increasing the central location and spread of the 25 highest values. Increases in the central location and spread tends to increase the magnitude of the highest value within the 25 highest concentrations. The robust highest value in effect is a direct measurable result of the composite impact of the central location of the highest values and their spread about that central location.

Performance Measures

The operational component of the evaluation compares the performance of the models in terms of the largest network-wide RHC test statistic. The robust highest value is calculated separately for each monitoring station within the network. The

largest measurement based RHC value in the monitoring network and the largest model based RHC value from the model estimates are used to calculate the absolute fractional bias for each model. An absolute fractional bias is calculated for each averaging period for which short-term ambient standards exist, e.g., 3- and 24-hour.

The scientific component of the evaluation is also based on the absolute fractional bias as the basic measure of performance. The absolute fractional bias for each model is calculated using the robust highest statistic determined for each meteorological condition and monitoring station. Only data for 1-hour averaging periods are used in this component of the evaluation. The meteorological conditions used are a function of atmospheric stability and wind speed. Six unique meteorological conditions are defined from two wind speed categories (below and above 4.0 m/s) and three stability categories: unstable (A,B,C), neutral (D), and stable (E and F). Other categories or combinations of meteorological conditions might be chosen at the discretion of the evaluator.

Composite Performance Measures

A composite performance measure is computed for each model as a weighted linear combination of the individual fractional bias components. Within the operational evaluation component, results for the various averaging periods are given equal weight. For the scientific component, each of the combinations of the meteorological conditions is given equal weight. Because the operational evaluation component is deemed to be the more important of the two, it is given a weight that is twice the weight of the scientific component. Finally, results from the various data bases are

given equal weight unless it is determined that differences in such factors as data quality and geographical coverage of the monitoring networks suggest otherwise. The algebraic expression for the composite performance measure (CPM) is:

$$CPM = \frac{1}{3}\overline{(AFB)_{i,j}} + \frac{2}{3}\left[\frac{(AFB)_3 + (AFB)_{24}}{2}\right],$$

where:

$(AFB)_{i,j}$ = Absolute Fractional Bias for meteorological category i at station j
$(AFB)_3$ = Absolute Fractional Bias for 3-hour averages
$(AFB)_{24}$ = Absolute Fractional Bias for 24-hour averages

Because the purpose of the analysis is to contrast the performance among the models, the composite performance measure is used to calculate pairs of differences between the models. For discussion purposes, the difference between the composite performance of one model and another is referred to as the model comparison measure. The expression for the model comparison measure is given by:

$$(MCM)_{A,B} = (CPM)_A - (CPM)_B ,$$

where:

$(CPM)_A$ = Composite Performance Measure for Model A
$(CPM)_B$ = Composite Performance Measure for Model B

When more that two models are being compared simultaneously, the number of model comparison measure statistics is equal to the total of the number of unique combinations of two models. For example, for three models, three comparison measures are computed, for four models, six comparison measures are computed, etc. The model comparison measure is used in judging the statistical significance of the apparent superiority of any one particular model over another.

Standard Error Determination

The yardstick used to determine if the composite difference between models is statistically significant is known as the standard error. At the simplest level, the ratio of the composite difference to the standard error provides a convenient measure of the significance for the resulting difference. Nominally, ratios that are larger than ± 1.7 are significant, while values < 1.7 indicate no significance at approximately the 90 percent level.

Because the model comparison measure is a rather involved statistic, the usual statistical methods for estimating the standard error do not apply. Fortunately, computing speeds have increased so that resampling techniques such as the "jackknife" and "bootstrap" are feasible methods for estimating the standard error and for determining confidence limits. Because of its simplicity, the blocked bootstrap method[9] is used to generate an estimate of the standard error. The bootstrap is basically a resampling technique whereby the desired performance measure is recalculated for a number "trial" years. To do this, the original year of data (nominally 365 days), is partitioned into 3-day blocks. Within each season, 3-day blocks (approximately 30 blocks per season) are sampled with replacement until a total season is created. This process is repeated using each of the four seasons to construct a complete bootstrap year. Three day blocks are chosen to preserve day to day meteorological persistence, while sampling within seasons guarantees that each season will be represented by only days chosen from that season. Since sampling is done with replacement, some days are represented more than once, while other days are not represented at all. Next, the data generated for the bootstrap year are used to calculate

the composite performance measures for each model. This process is repeated until sufficient samples are available to calculate a meaningful standard error for each of the model performance statistics. The standard error is calculated as simply the standard deviation of the bootstrap generated outcomes for the model comparison measure.

Selecting the Best Models

The magnitude and sign of the model comparison measure are indicative of the relative performance of each pair of models. The smaller the composite performance measure the better the overall performance of the model. This means that for two arbitrary models, Model A and Model B, a negative difference between the composite performance measure for Model A and Model B implies that model A is performing better (Model A has the smaller composite performance measure), while a positive value indicates that model B is performing better. When only two models are compared, the test statistic is simply the ratio of the composite difference to the standard error calculated from the bootstrap outcomes.

When more than two models are being compared, it is convenient to calculate simultaneous confidence intervals for each pair of model comparisons.[10] For each pair of model comparisons, the significance of the model comparison measure depends upon whether or not the confidence interval overlaps zero (0). If the confidence interval overlaps zero, the two models are not performing at a level which is statistically different. If the confidence interval does not overlap zero, (upper and lower limits are both negative or both positive), then there exists a statistically significant difference between the two models at the stated level of confidence.

The level of confidence chosen has an important impact on the decision. The larger the probability or confidence level, the larger the length of the confidence limits required to satisfy the confidence statement. Choosing a confidence level that is overly demanding (e.g., 99.9999%) would almost surely result in such wide limits that no decision could be reached regarding which model(s) are performing better. At the other extreme, choosing a confidence level that is very lenient (e.g., 70%) may lead to a decision that one or more models are superior when in fact no real difference exits. This choice must be such that the two competing needs are balanced which requires judgement from the evaluators. A confidence level in the vicinity of 90 to 95 percent represents a reasonable compromise between these two needs.

Limitations

This protocol document contains very specific requirements for conducting the statistical comparisons believed necessary to compare the performance of models. These requirements are based on experiences gained from EPA's model evaluation activities over the past several years. The reader is reminded that there may be more logical choices of meteorological conditions and specific weights for compositing performance among various data categories. Likewise, the specific test statistic, performance measure and range of data may be different depending on the nature of the data bases being used and the judgement of those conducting the evaluation.

4.0 DISPLAY OF RESULTS

To fully understand the final outcome from applying the methodology, each of the component results should be examined. For example the absolute fractional bias does

not provide any information about the direction of the bias, i.e., it does not indicate if a particular model tends to under or overpredict. Greater understanding about the relative performance of each model can be obtained through graphic display of the fractional bias for the various data categories used in the evaluation.

For the screening test, results are displayed graphically using the fractional bias of the average vs the fractional bias of the standard deviation as illustrated in Figure 1 (see reference 2). Information is presented separately for each averaging period and each of the data bases used in the analysis. For the statistical test, this is accomplished with the use of box diagrams (see Appendix A) which display the magnitude of selected percentiles for the fractional bias using the outcomes of the bootstrap process. Although these diagrams are not intended for use in making the final decision, they are useful in summarizing and presenting the outcome. Also, the scientific results should prove useful for the improvement of existing models and in the development of new models.

5.0 REFERENCES

1. Environmental Protection Agency, 1982. Evaluation or Rural Air Quality Simulation Models (EPA-450/4-83-003). U.S. Environmental Protection Agency, Research Triangle Park, NC.

2. Environmental Protection Agency, 1985. Evaluation of Rural Air Quality Simulation Models, Addendum B: Graphical Display of Model Performance Using the Clifty Creek Data Base (EPA-450/4-83-003b). U.S. Environmental Protection Agency, Research Triangle Park, NC.

3. Cox, W. M. and J. A. Tikvart. Assessing the Performance of Air Quality Models, Paper presented at the 15th International Technical Meeting on Air Pollution and its Applications (NATO/CCMS), April 16-19, 1985, St Louis, MO.

4. Baldridge, K. W., 1985. Standardized SAS Graphics Subsystem User's Manual. Prepared by Computer Sciences Corporation for EPA, Computer Sciences Corporations, Research Triangle Park, NC.

5. Efron B., 1982. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, PA.

6. Environmental Protection Agency, 1984. Interim Procedures for Evaluating Air Quality Simulation Models (Revised) (EPA-450/4-83-023). U.S. Environmental Protection Agency, Research Triangle Park, NC.

7. Cox, W. M. and J. A. Tikvart, 1990. A statistical procedure for determining the best performing air quality simulation model. *Atmos. Environ.*, 24A(9): 2387-2395.

8. Breiman, L., J. Gins and C. Stone, 1978. Statistical Analysis and Interpretation of Peak Air Pollution Measurements (TSC-PD-A190-10). Technology Service Corporation, Santa Monica, CA.

9. Tukey, J. W., 1987. Kinds of bootstraps and kinds of jackknifes, discussed in terms of a year of weather-related data (Technical Report No. 292). Department of Statistics, Princeton University, Princeton, NJ.

10. Cleveland, W. S., and R. McGill, 1984. Graphical Perception Theory, Experimentation, and Application to the Development of Graphical Methods. *J. Am. Stat. Assoc.*, 79(387): 531.

APPENDIX A

EXAMPLE COMPARISON OF MODEL PERFORMANCE

A.1 INTRODUCTION

This appendix demonstrates an example application of the protocol for comparing models using six data bases available around four large power plants located in the midwest. For purposes of this example, MPTER (Version 6) and an alternative point source model are evaluated and compared. The MPTER model was chosen since it is the EPA preferred model for regulatory applications. The alternative model was chosen as a state-of-the-art rural point source model incorporating advanced features for simulating plume behavior in flat terrain.

The six model evaluation data bases used in the analysis consisted of Clifty Creek (1975 and 1976), Muskingum River (1975 and 1976), Paradise (1976) and Kincaid (1980/1981). The Clifty Creek plant is a coal fired, base-load facility located along the Ohio River in southern Indiana. Terrain surrounding the plant consists of low ridges and rolling hills at elevations that are below the top of the stacks. Hourly SO_2 data is available for six monitoring stations located at distances ranging from approximately 3 - 15km from the power plant. The Muskingum River plant is also a coal-fired plant, located in Ohio surrounded by low ridges and rolling hills. Four SO_2 monitoring stations are located at distances ranging from 4 - 20km from the plant. The Paradise plant, located in Kentucky, has 12 monitors located at distances from 3 - 17km from the plant. The Kincaid plant, located in central Illinois, employed a dense network of SO_2 monitors ranging from approximately 2 - 20 km from the plant. Each of these data bases is documented in greater detail elsewhere.[1,2,3,4] For each of these data bases, 1, 3 and 24-hour average measured and predicted concentrations have been assembled for each of the operating monitoring stations. In addition, wind speed and atmospheric stability are

available for each of the hourly records. An hourly background concentration was estimated and subtracted from the measured hourly concentrations using the EPA method.[5] In addition, a threshold check is used to eliminate low observed or predicted values that have no effect on the performance statistics. For 1-hour averages, a threshold value of 25 $\mu g/m^3$ is used while, for 3-hour and 24-hour averages, a value of 5 $\mu g/m^3$ is used. The threshold checks are applied independently to the measured and predicted concentrations.

A.2 SCREENING TEST RESULTS

For each data base, the observed concentrations from all monitoring stations were pooled and sorted by averaging period. From the sorted data, the 25 highest observed concentrations, unpaired in space or time, were used to calculate a mean and standard deviation. The same procedure was applied to the predicted concentrations obtained from MPTER (Version 6) and the alternative model. Using these statistics, a fractional bias for the mean and a fractional bias for the standard deviation was determined for each model for 3-hour averages and for 24-hour averages.

Figure 1 shows the results in which the fractional bias of the average and fractional bias of the standard deviation are plotted for 3-hour averages. For both MPTER (left panel) and the alternative model (right panel), the results for all six of the data bases are shown. For both models, the predicted and observed averages and standard deviations are within a factor-of-two except at Kincaid where underpredictions are apparent. Figure 2 shows similar results for the 24-hour averages. Again, most of the data points indicate performance within a factor-of-two except for the alternative model at Clifty Creek where a tendency for overpredictions is evident. Since both models satisfactorily meet the

High 25 Concentrations: 3-hour Averages

1 = Clifty Creek (1975) 2 = Clifty Creek (1976) 3 = Muskingum River (1975)
4 = Muskingum River (1976) 5 = Paradise (1976) 6 = Kincaid (1980/81)

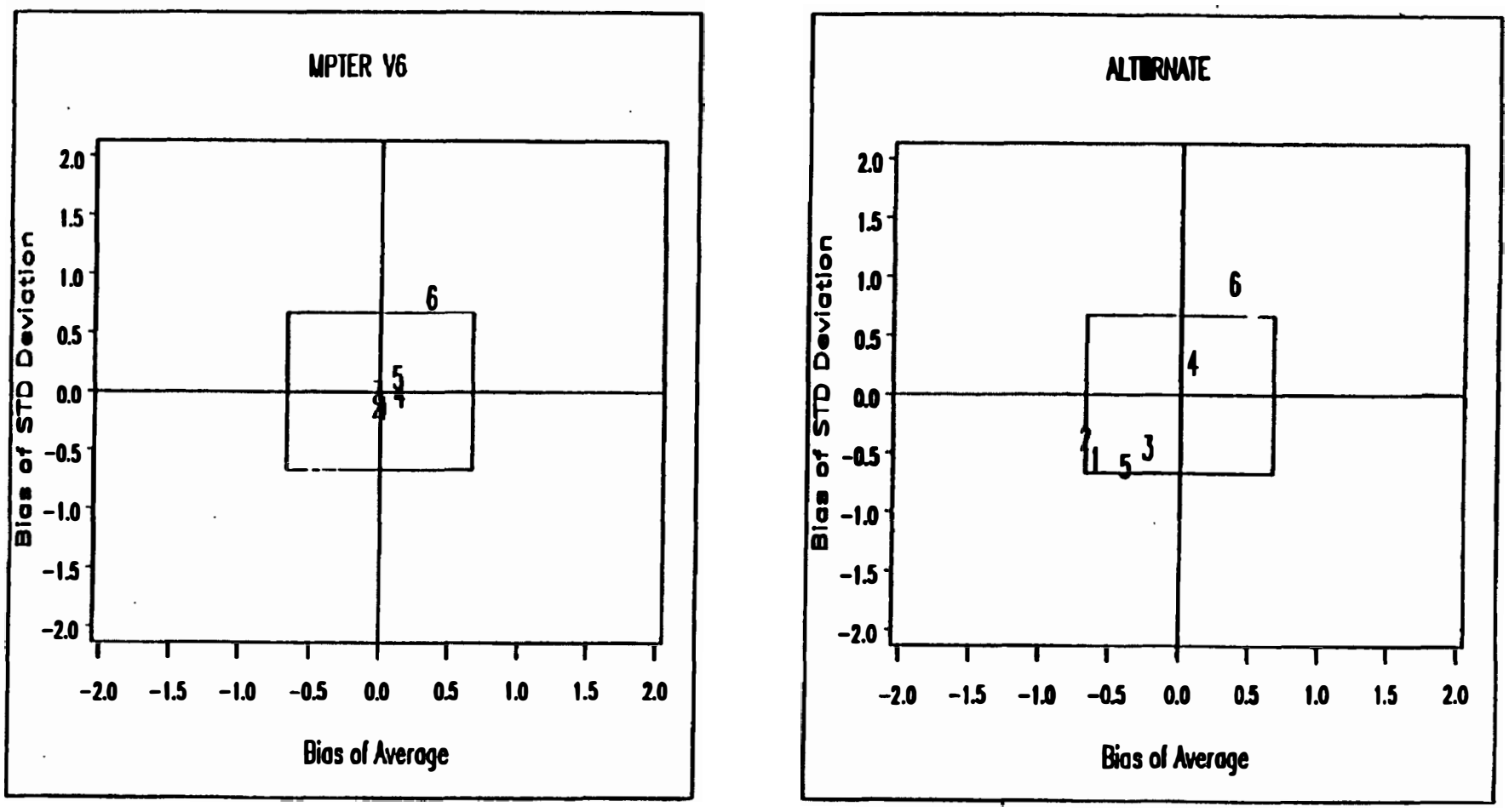

Figure A-1. Fractional bias of the average and standard deviation: 3-hour averages.

High 25 Concentrations: 24-hour Averages

1 = Clifty Creek (1975) 2 = Clifty Creek (1976) 3 = Muskingum River (1975)
4 = Muskingum River (1976) 5 = Paradise (1976) 6 = Kincaid (1980/81)

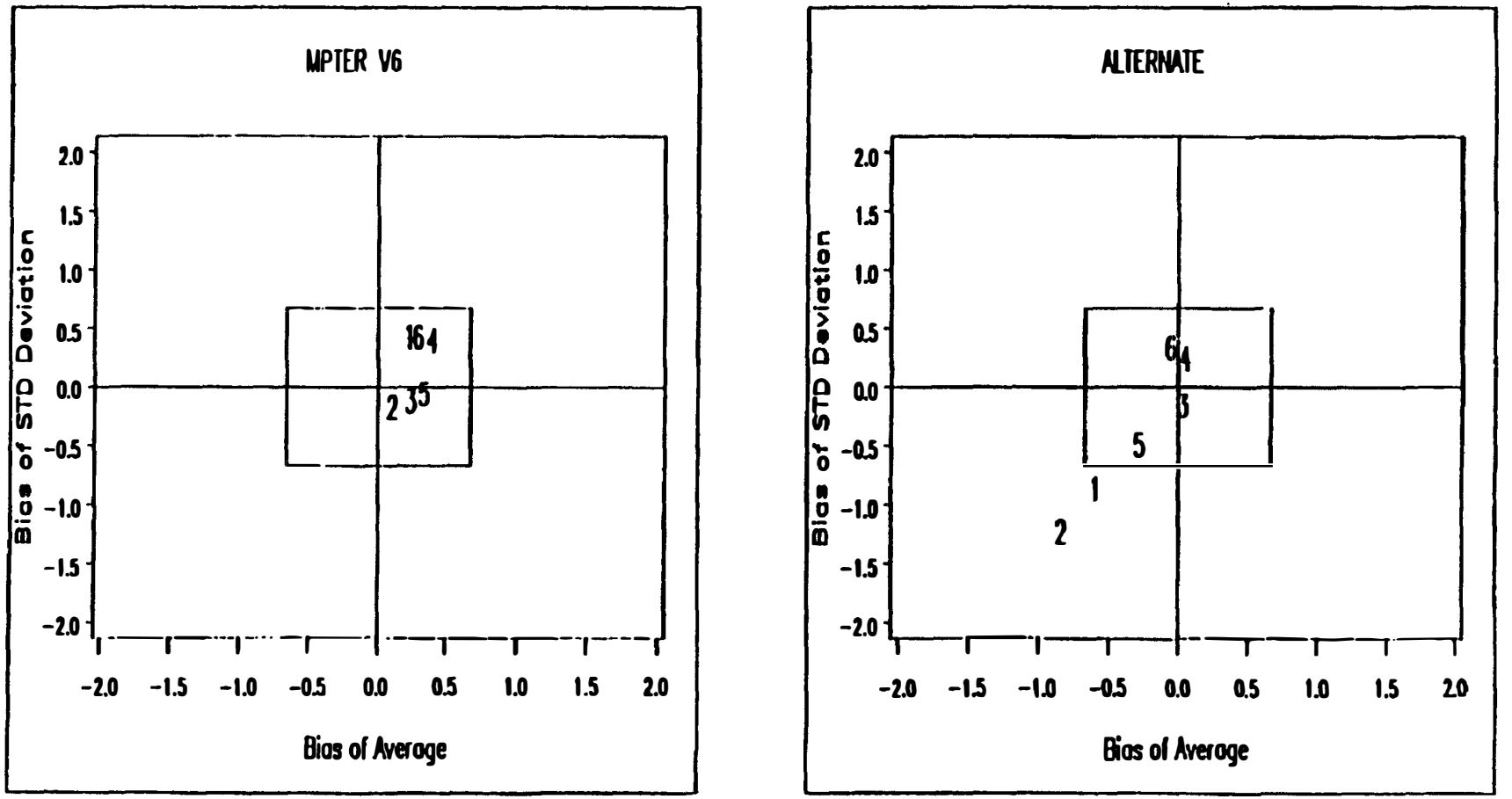

Figure A-2. Fractional bias of the average and standard deviation: 24-hour averages.

minimum level of performance, the two models are subjected to a more comprehensive statistical comparison.

A.3 STATISTICAL COMPARISONS

The performance of MPTER is compared with the performance of the alternative model using a composite statistical measure that combines the performance within the operational component (3-hour and 24-hour averages) and the scientific component (1-hour averages). For purposes of the operational component, the observed and predicted concentrations were sorted separately by station and averaging period. Using the 25 largest values, the statistical procedures described in the protocol were applied to calculate the robust highest concentration (RHC) at each station. A network based absolute fractional bias was computed for each averaging period and model using the largest observed RHC and the largest predicted RHC value from among the monitoring stations in each data base.

For the scientific component, six meteorological categories were defined from two wind speed categories and three stability categories. The two wind speed categories are: low ($\leq$4.0 m/s) and high (>4.0 m/s). The three stability categories are: unstable (class A, B, C), neutral (class D), and stable (class E, F). To minimize distortions associated with small counts, data categories having fewer than 100 observations were eliminated from the analysis.

The hourly observed and hourly predicted concentrations within each data category were sorted. The 25 highest values were used to calculate a separate robust highest

concentration for each of the station/meteorological data categories. A composite absolute fractional bias was computed by averaging the individual absolute fractional biases. A composite performance measure for each model was then calculated by averaging three quantities: (1) the absolute fractional bias based on 3-hour averages, (2) the absolute fractional bias based on 24-hour averages, and (3) the composite absolute fractional bias based on 1-hour averages. The difference between the composite performance measure for MPTER and the alternative model (Model comparison measure) is actually the statistic used in judging the overall difference in performance between the two models.

Following the procedure outlined in the protocol, 100 bootstrap trial years were generated.* For each trial year, the statistics and model performance measures described above were recalculated resulting in 100 sets of statistical outputs. The statistics in each set included the fractional biases, absolute fractional biases, composite absolute fractional biases, composite performance measure and the differences between the composite performance measures for MPTER and the alternative model. For this example demonstration, a confidence level of 90 percent was selected for determining statistical significance for the difference in performance between the two models.

A.4 STATISTICAL RESULTS: CLIFTY CREEK

Figure 3 presents an example comparison of the bias for the two models using the 1975 Clifty Creek data. The figure presents the results for 1-, 3- and 24-hour averages

*The number of bootstrap trials was limited to 100 by available computing resources. Nominally, 500 to 1000 bootstrap trials would be used if computational resources were not a prime consideration.

Clifty Creek (1975)

MPTER V6 (□) and Alternate (X)

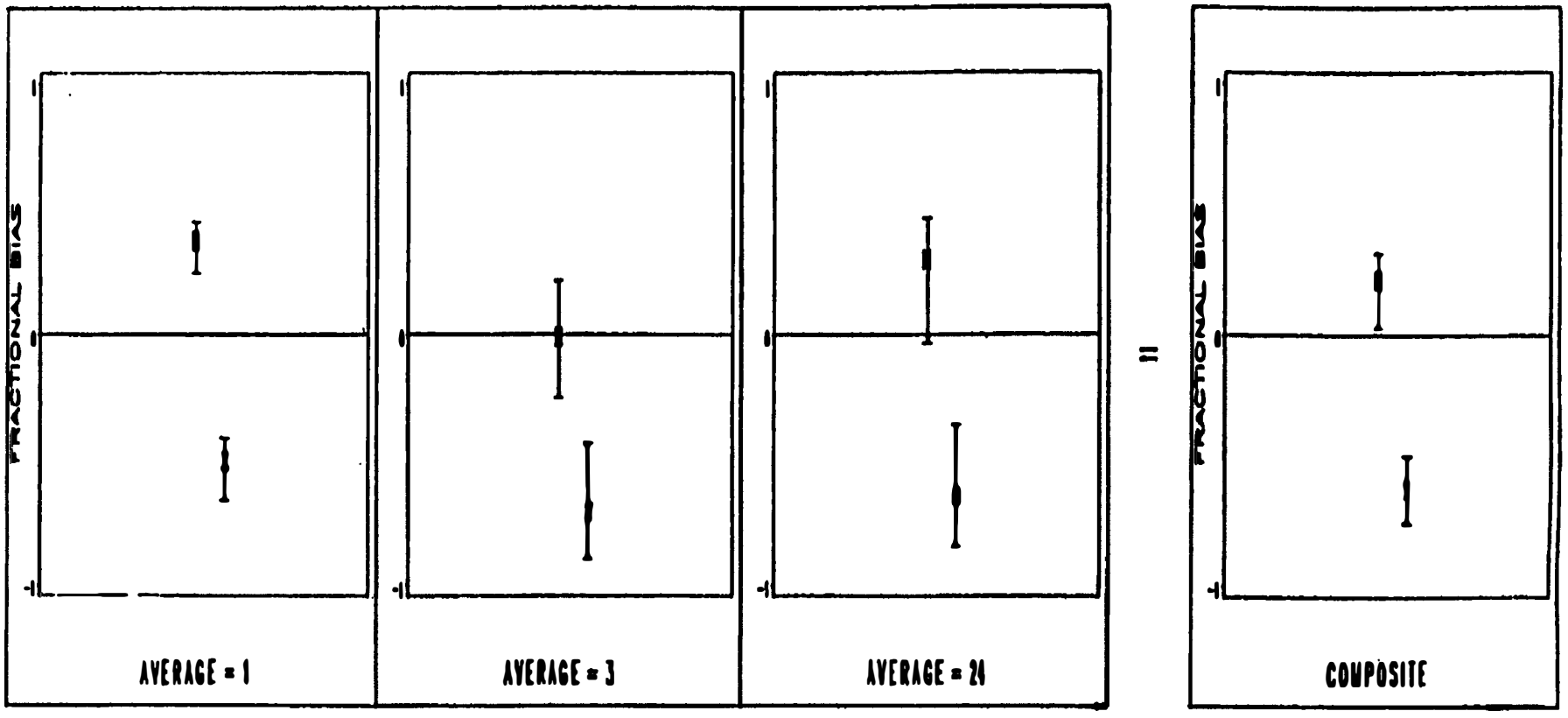

Figure A-3. Fractional bias and bootstrap percentiles (5 and 95) for Clifty Creek (1975): MPTER and Alternate Model

in terms of the RHC test statistic for Clifty Creek(1975). The data displayed consists of the fractional bias for each averaging period along with the 5th and 95th percentiles resulting from the bootstrap. The results for 1-hour averages is the composite average over the individual stations and six meteorological conditions, while the 3-hour and 24-hour results are based on the largest RHC test statistic across the six monitoring stations. The composite 1-hour fractional bias indicates an overall tendency for MPTER to underpredict peak 1-hour concentrations at Clifty Creek.* Since the upper and lower percentiles are far above the zero reference line, the underpredictions are "significant" in a statistical sense. Composite 1-hour results for the alternative model indicate a clear tendency for overprediction at Clifty Creek. For 3-hour averages, MPTER appears to be essentially unbiased while the alternative model shows a tendency for overpredictions. For 24-hour averages, MPTER shows a tendency for modest underpredictions while the alternative model shows a clear tendency for overprediction. The overall composite fractional bias shown in the last panel suggests a tendency for underpredictions by MPTER and overpredictions by the alternative model. The composite underprediction by MPTER is dominated by the 1-hour results while the composite overpredictions by the alternative model are more evenly spread across each of the three averaging periods. Table 1 summarizes the results of the comparison in performance between the two models in terms of the absolute fractional bias for each averaging period. The ratio of the difference between the two models to the standard error provides a rough measure of the statistical significance of the difference in composite performance between the two

*For 1-hour concentrations unpaired in space or time, both models actually overpredict the observed 1-hour concentrations. If ambient standards existed for 1-hour average data, the operational component of this evaluation would include 1-hour average comparisons equivalent to those described for 3-hour and 24-hour averages.

Table A-1. Composite Performance of MPTER and the Alternative Model for Six Rural Data Bases

Data Base	Averaging Period	Absolute Fractional Bias MPTER	Absolute Fractional Bias Alternate	Diff. (d)	Std. Dev. (s)	Ratio (d/s)
Clifty Creek (1975)	1-hr	0.81	0.83	-0.02	0.05	-0.4
	3-hr	0.01	0.71	-0.70	0.14	-5.0
	24-hr	0.31	0.64	-0.33	0.24	-1.4
	Composite	0.37	0.73	-0.35	0.10	-3.5
Clifty Creek (1976)	1-hr	0.58	0.47	0.11	0.04	2.8
	3-hr	0.25	0.73	-0.48	0.12	-4.0
	24-hr	0.12	0.91	-0.79	0.14	-5.6
	Composite	0.41	0.78	-0.37	0.07	-5.3
Muskingum River (1975)	1-hr	1.04	0.60	0.43	0.07	6.1
	3-hr	0.10	0.22	-0.12	0.15	-0.8
	24-hr	0.08	0.02	0.06	0.17	0.4
	Composite	0.40	0.28	0.12	0.08	1.5
Muskingum River (1976)	1-hr	0.54	0.38	0.41	0.08	5.1
	3-hr	0.04	0.11	-0.06	0.10	-0.6
	24-hr	0.34	0.03	0.31	0.12	2.6
	Composite	0.44	0.23	0.21	0.05	4.2
Paradise (1976)	1-hr	1.25	0.75	0.50	0.03	16.7
	3-hr	0.09	0.55	-0.46	0.10	-4.6
	24-hr	0.25	0.48	-0.23	0.14	-1.6
	Composite	0.53	0.59	-0.06	0.13	-0.5
Kincaid (1980/81)	1-hr	0.68	0.59	0.09	0.04	2.2
	3-hr	0.29	0.53	-0.24	0.39	-0.6
	24-hr	0.29	0.69	-0.40	0.35	-1.1
	Composite	0.42	0.60	-0.18	0.22	-0.8
Grand Composite (All 6 Data Bases)		0.43	0.54	-0.11	0.05	-2.2

models. Absolute values for the ratio that exceed a nominal value of 1.7 indicate significance at approximately the 90 percent confidence level. For the Clifty Creek 1975 data base, the composite results indicates that the overall performance of MPTER is significantly better than the performance of the alternative model. The difference between the composite absolute fractional bias statistics for the two models is -0.35 which is 3.5 times as large as the standard error for the difference. Before discussing the results for all six data bases, it is instructive to examine the results at Clifty Creek more closely. Although the composite difference between the two models is large, there are noticeable differences among the three averaging periods. For 1-hour averages, the two models performed about the same but for different reasons. By referring to Figure 3, it is clear that underpredictions by MPTER and overpredictions by the alternative model are of approximately the same magnitude. The net effect is that both models are penalized by approximately the same degree. The performance of the two models for 3-hour averages appears to be the dominant factor contributing to the overall difference. The fractional bias for MPTER is essentially zero while for the alternative model the fractional bias is 0.71 leading to a large difference compared with its standard error (-0.70 vs. 0.14). For 24-hour averages, MPTER is also the better performing model. The absolute fractional bias for MPTER is low (0.31) but not significantly lower than for the alternative model (0.64).

A.5 STATISTICAL RESULTS: ALL DATA BASES

The relative performance between the two models varies somewhat among the six data bases. The ratio of the composite difference to its standard error ranged from -5.3 at Clifty Creek (1976) to 4.2 at Muskingum (1976). At Clifty Creek, MPTER clearly

performs better than the alternative model while at Muskingum River, the alternative model performs at least as well or better than MPTER. For Paradise and Kincaid the composite results indicate that MPTER is performing slightly better than the alternative model; however, the results are not statistically significant. The grand composite result over the six data bases indicates that MPTER performs statistically better than the alternative model. The composite difference between the two models is -0.11 which is more than twice the estimated standard error.

Figures 4 and 5 present the composite results graphically for each averaging period and for the overall grand composite. Clearly, the overall tendency is for MPTER to perform better for 3-hour and 24-hour averages (operational component), while the alternative model performs better for 1-hour averages (scientific component). The composite statistics shown in the last panel of Figures 4 and 5 suggest that the better performance of MPTER in the operational component more than compensates for the better performance by the alternative model in the scientific component. Note that in this example comparison, the overall statistical significance was small and also there were rather large differences in performance between data bases. In practice, these facts might be taken into consideration when choosing the model for applications and/or in considering whether additional data were necessary before arriving at a final decision.

Composite for All Inventories

MPTER V6 (□) and Alternate (X)

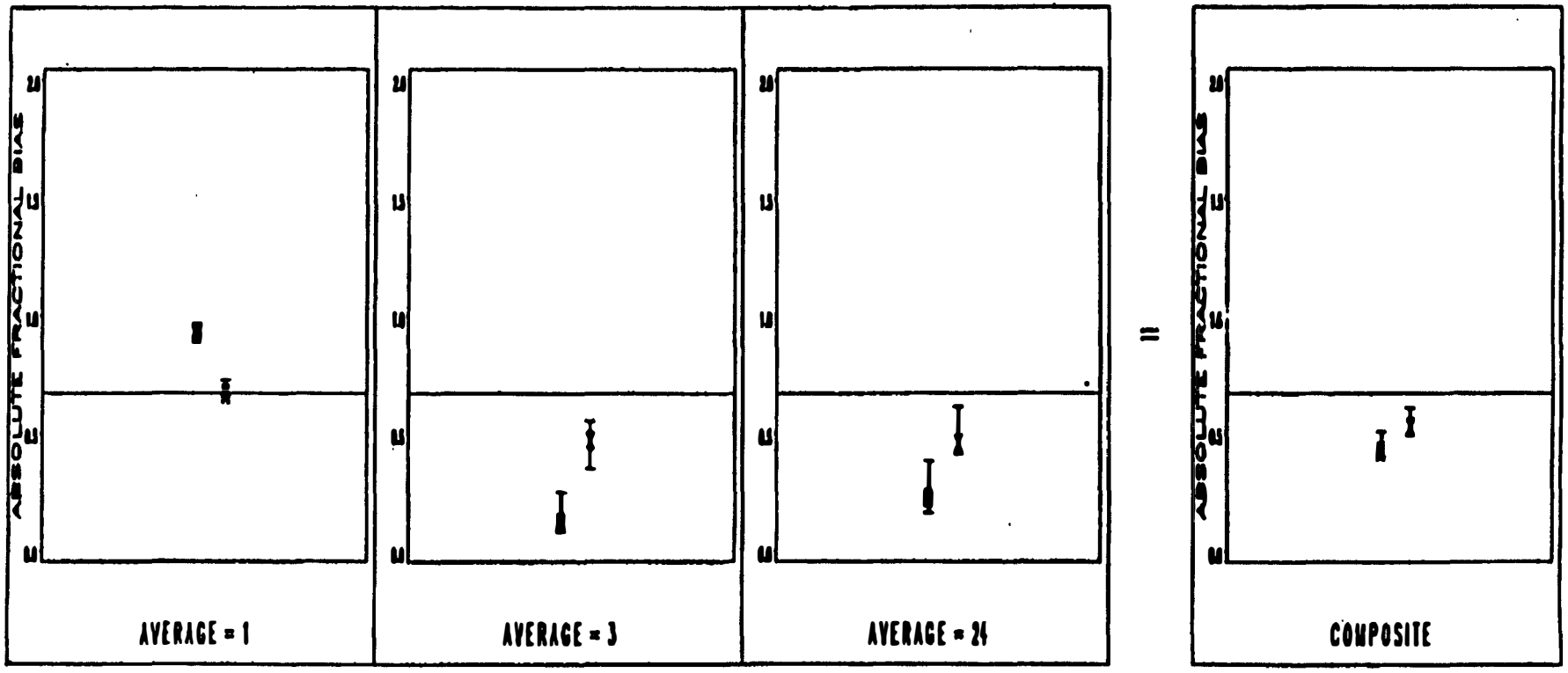

Figure A-4. Absolute fractional bias and bootstrap percentiles (5 and 95), composite for six inventories: MPTER and Alternate Model

Composite for All Inventories

Difference between MPTER V6 and Alternate

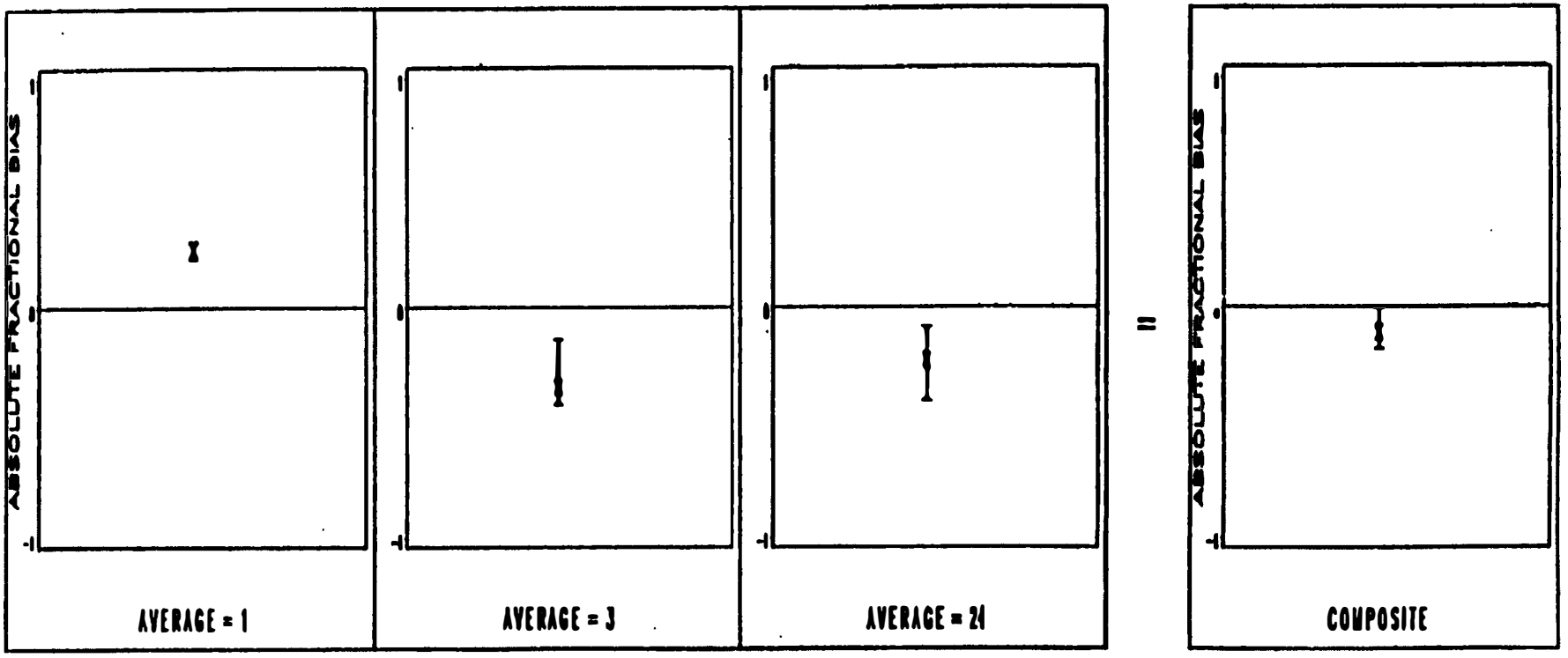

Figure A-5. Absolute fractional bias and bootstrap percentiles (5 and 95), composite difference for six inventories: MPTER and Alternate Model

A.5 REFERENCES

1. Environmental Protection Agency, 1982. Evaluation or Rural Air Quality Simulation Models (EPA-450/4-83-003). U.S. Environmental Protection Agency, Research Triangle Park, NC.

2. Environmental Protection Agency, 1985. Evaluation of Rural Air Quality Simulation Models, Addendum A: Muskingum River Data Base (EPA-450/4-83-003a). U.S. Environmental Protection Agency, Research Triangle Park, NC.

3. Environmental Protection Agency, 1986. Evaluation of Rural Air Quality Simulation Models, Addendum C: Kincaid SO_2 Data Base (EPA-450/4-83-003c). U.S. Environmental Protection Agency, Research Triangle Park, NC.

4. Environmental Protection Agency, 1987. Evaluation of Rural Air Quality Simulation Models, Addendum D: Paradise SO_2 Data Base (EPA-450/4-83-003d). U.S. Environmental Protection Agency, Research Triangle Park, NC.

5. Environmental Protection Agency, 1987. Guideline on Air Quality Models (Revised) and Supplement A (EPA-450/2-78-027R). U.S. Environmental Protection Agency, Research Triangle Park, NC.

TECHNICAL REPORT DATA

(Please read Instructions on reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA-454/R-92-025		

4. TITLE AND SUBTITLE	5. REPORT DATE
Protocol for Determining the Best Performing Model	December 1992
	6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
William M. Cox	

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
U.S. Environmental Protection Agency Office of Air Quality Planning and Standards Technical Support Division Research Triangle Park, NC 27711	11. CONTRACT/GRANT NO.

12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This document describes a recommended procedure for evaluating the performance of air quality dispersion models, and for selecting the best performing model for particular regulatory applications. The procedure is based on direct comparisons between measured and model-predicted concentrations to establish the accuracy of each model. The document includes an example evaluation using SO_2 data collected at four midwestern power plants in which the performance of MPTER and one other rural air quality model were compared.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollution Atmospheric Dispersion Modeling Model Performance Statistical Evaluation		

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(Report)*	21. NO. OF PAGES
Release Unlimited	Unclassified	30 (incl. appendix)
	20. SECURITY CLASS *(Page)*	22. PRICE
	Unclassified	