UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711
AUG 9 1988

MEMORANDUM

SUBJECT: Model Accuracy and Uncertainty

FROM: William G. Laxton. Director
Technical Support Division (MD-14)

TO: Gerald A. Emison, Director
Office of Air Quality Planning and Standards (MD-10)

In the application of air quality models, decision makers are usually constrained to using calculated concentrations as "best estimates" for determining the adequacy of emission limits. The details of model accuracy and uncertainty are frequently too complex to consider within the confines of pollution control decisions that are routinely made by Regional Offices and State/local agencies. Thus to maintain the credibility of these mathematical tools, the Source Receptor Analysis Branch has a continuing concern with the accuracy of air quality models and with the appropriate use Of Model estimates in decision making. This concern is most acute for point source applications involved in new source review or in setting emission limits as part of a SIP revision. In the past our studies have shown that highest estimated concentrations for point source models typically have an accuracy of ± 10 to 40 percent, or well within the often quoted factor-of-two that has long been recognized for these models. However, through an extensive program we have been conducting over the last few years, it is now possible to make a definitive statement about the accuracy associated with MPTER ' and related UNAMAP models that are commonly applied to large point sources. In addition, we are able to suggest how model accuracy might be used along with "best estimates" to calculate the probability that model-based emission limits will attain ambient standards. This information is summarized below and explained in more detail in the attached report.

Model Accuracy

Based on data for four major midwestern power plants, we can now say that the MPTER model appears to be essentially unbiased in estimating highest concentrations for averaging periods consistent with existing S02 ambient standards. Actually, exact calculation of model bias, defined as the ratio of model predicted design concentrations to measurement based design concentrations, shows small departures from a true absence of bias, i.e., a ratio of 1.0. The composite data indicate that MPTER (1) slightly overestimates highest 3-hour average concentrations (by about 9 percent or a bias ratio of 1.09) and (2) slightly underestimates highest 24-hour average concentrations (by about 15 percent or a bias ratio of 0.85). More explicitly we can say that for 3-hour average concentrations the ratio of highest estimated to measured concentrations is 1.09, and the true value of this ratio lies between 0.97 and 1.23 with 95% confidence. For 24-hour average concentrations the ratio is 0.85 and the true value (with 95% confidence) lies between 0.74 and 0.98. However, it should be noted that these statements are properly limited to major isolated point sources where accuracy of estimates in "space and time" is not an issue; they are not applicable to multi-source situations. Nevertheless, we believe the statements to be typical of accuracy associated with a wide variety of point source applications.

Using Model Accuracy in Decision Making

A potentially important by-product of the model accuracy studies is an ability to assess the likelihood that emission limits based on model estimates will result in attainment of ambient standards. With this goal in mind, we sponsored development of a technique known as CUE (Calculation of Uncertainty Estimates). CUE combines information on accuracy with "best estimates" from the model in order to calculate probability distributions of attainment for alternative emission limits. Although it was originally intended for source specific applications and computationally intensive, simplifying assumptions make it possible to generalize the CUE calculations for the four midwestern power plants. Tables 1 and 2, taken from the attached report, provide a composite of the probability of attainment for assumed levels of model bias.

In Table 1 it can be seen that as the "best estimate," or the predicted design concentration, increases from a value below the standard to a value above the standard, the probability of attainment decreases for a specified emission limit. That is, a decision maker should be more certain of attainment for an emission limit that results in a best estimate that is less than the standard than for emissions that produce a best estimate that is greater than the standard. For example, with an unbiased model (i.e., a bias ratio of 1.0), a best estimate that is only 70 percent of the standard results in near certainty (96%) of attainment. Whereas a best estimate that is 30 percent greater than the standard results in a clear lack of certainty (9%) of attainment, or rather certainty of nonattaiment. Similarly, as the bias ratio increases from systematic underestimates to systematic overestimates, the degree of certainty in attainment increases. A model that is biased to overestimate provides greater assurance of attainment (67% at a bias ratio of 1.09) for a best estimate equal to the standard than a model that underestimates (21% at a bias ratio of 0.85). While these concepts may be intuitively obvious, this is the first time they have been clearly quantified in tabular form.

This information may also be used to choose an emission limit that will result in a predefined probability of attainment as shown in Table 2. For example, while not commonly considered in the decision-making process, it is a fact that an unbiased model provides only a 50-50 chance of attainment whin the best estimate is exactly equal to the standard. Thus, if a decision maker wishes to be more certain of attainment than 50%, the best estimate for an emission limit must actually be less than the targeted standard. For a decision maker to be at least 70% sure of attainment, the emission limit for the source must be such that the "best estimate" of an unbiased model is 90% of the standard or less. If the decision maker were to require a 95% probability of attainment, the emission limit would have to be tightened so that the best estimate is 72% or less of the standard.

The procedure for dealing with uncertainty is simplified and highly dependent on the variance (scatter in model error), as well as the bias of model estimates. Thus the calculated probabilities should be considered as only an approximation to results expected from a site specific application of the CUE technique. Issues such as data extrapolation, emission variability and load conditions must be considered to determine how accurate such calculations are for application to an isolated point source.

The information presented herein is intended to enhance and supplement the basis for using the “best estimates” provided by recommended EPA models. It is not intended to change or modify how those estimates are used by Regional Offices and State/local agencies. However, we believe that this documents, with the greatest precision that can be mustered, a statement on the accuracy and uncertainty associated with standard model applications. It remains to be seen whether there is a way in which a decision maker can fully utilize such information, within the constraints of CAA requirements and our policies on various pollutants, i.e., $S0_2$. If you wish to further discuss the interpretation and use of this information, we will be happy to set up a briefing for you.

Attachments

cc: Modeling Contact, Regions I-X
J. Calcagni
W. Cox
F. Schiermeier
B. Steigerwald
J. Tikvart

Table 1

Probability of Attainment of Ambient Standards for Large SO_2 emitting Power Plants

Best Model Estimate (DV)	Probability of Attainment (Percent)				
	Bias Ratio (Model/Monitor)				
	0.5	0.85	1.00	1.09	1.5
0.7	5	83	96	99	100
0.8	1	62	87	94	100
0.9	0	39	70	83	99
1.0	0	21	50	67	98
1.1	0	10	32	48	94
1.2	0	4	18	32	87
1.3	0	2	9	19	76

Note: Assumes that the ratio of model based to measurement based design concentrations is lognormally distributed with log mean equal to the bias ratio and log standard deviation = 0.20. The "best estimate" (DV) is expressed as the ratio of the model based design value to the level of the ambient standard.

Table 2

Model Based Design Values Required to Achieve A Given Probability of Attainment For Large SO_2 Emitting Power Plants

Probability of Attainment (percent)	Design Value (DV)				
	Bias Ratio (Model/Monitor)				
	0.50	0.85	1.00	1.09	1.50
50	0.50	0.85	1.00	1.09	1.50
55	0.49	0.83	0.98	1.06	1.46
60	0.48	0.81	0.95	1.04	1.43
65	0.46	0.79	0.93	1.01	1.39
70	0.45	0.77	0.90	0.98	1.35
75	0.44	0.74	0.87	0.95	1.31
80	0.42	0.72	0.85	0.92	1.27
85	0.41	0.69	0.81	0.89	1.22
90	0.39	0.66	0.77	0.84	1.16
95	0.36	0.61	0.72	0.78	1.08

Note: Assumes that the ratio of model based to measurement based design concentrations is lognormally distributed with log mean equal to the bias ratio and log standard deviation = 0.20. The "best estimate" (DV) is expressed as the ratio of the model based design value to the level of the ambient standard.

ASSESSMENT AND USE OF ACCURACY INFORMATION FOR MPTER

July 1988

United States Environmental Protection Agency
Office of Air Quality Planning and Standards
Technical Support Division
Source Receptor Analysis Branch

Assessment and Use Of Accuracy Information for MPTER

Introduction

EPA has an ongoing program to evaluate the performance of air quality models used in regulatory applications. As part of this effort, the MPTER model and several other rural models, have been extensively evaluated using S02 data collected around four major power plants located in the midwest. Reported results indicate that MPTER is essentially an unbiased model in the sense that highest 3-hour and 24-hour average model predictions are similar in magnitude to measured values. The purpose of this paper is summarize available information describing the performance of MPTER and to demonstrate how this information might be included along with "best estimates" from routine modeling applications. More specifically, this exercise is intended to: (1) summarize the accuracy of the MPTER model in terms of bias and precision using ratios of model-based to measurement-based design concentrations and, (2) demonstrate how model bias and precision information may be used to assess the probability of attainment/non-attainment for large S02 emitting sources.

Because of its importance in determining NAAQS compliance, this analysis will focus on the most significant statistic from a regulatory perspective--the so called design concentration for averaging periods consistent with existing S02 ambient standards. Conceptually, the design concentration is determined as the maximum network-wide second highest concentration. For purpose of this analysis, the design concentration was calculated by. statistical extrapolation in which the 25 highest concentrations at each station are fit to an exponential frequency distribution. Since the fitting process smooths out the affect of potential outliers the tail estimated design value is more robust than the highest or second highest concentration. The largest of these extrapolated second highest concentrations from among the network of monitoring stations is used as the design concentration in all subsequent calculations.

The statistical measure selected to characterize model performance is simply the ratio of the model-based to monitor based design concentration. This ratio is easy to interpret since values of the ratio greater than one measure the degree by which the model overpredicts while values that are less than one measure the degree by which the model underpredicts. This ratio was also selected because it simplifies calculation of attainment/non-attainment probabilities using "best estimates" resulting from MPTER.

Procedure

As indicated above, the bias of the model is directly estimated as simply the ratio of model-based to monitor-based design concentration. The concept and calculation of model precision is slightly more involved and requires further discussion.

Conceptually, model precision is measured by the scatter of the ratio of model-based to monitor-based design concentration. Scatter is typically calculated as the standard deviation although in instances where the ratio appears to be log-normally distributed, the log standard deviation is more appropriate.

Calculation of model precision requires knowledge of the probability distribution of the design value ratios. Since this distribution is not easily obtained using classical statistical methods, the bootstrap procedure was used to generate an approximation to this distribution. The bootstrap, which has been widely applied in air pollution analysis, is a resampling method in which the statistic (ratio) is regenerated a large number of times. Using the bootstrap outcomes, the standard deviation and log standard deviation is easily calculated.

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). For each of the data bases, approximately one year of hourly SO2 measurements and the associated hourly predictions for MPTER are available. For each data set the bootstrap procedure was applied to generate 100 trial years of observed and predicted concentrations. For each trial year, the 3-hour and 24-hour design concentrations were calculated. Using these outcomes, the ratio of the predicted to observed design concentration was determined.

The results of these calculations indicate that the ratios appear to be log-normally distributed. Thus it is appropriate to define model precision as the standard deviation of the logarithm of the bootstrap ratios. Table I presents summary statistics for each of the data bases including the bias ratio and log standard deviation of the 3-hour and 24-hour ratios.

Model Bias

The question of model bias can be addressed at two levels. (1) Is there evidence to suggest that MPTER overpredicts or underpredicts either the 3-hour or 24-hour highest concentrations for any of the data bases? (2) What is the best composite estimate of the bias of MPTER among the data bases and does it indicate an overall tendency for either over or underprediction?

A quick answer to the first question can be obtained by comparing the difference between the log of the actual ratio and the log of the ideal ratio (1.00) with the standard deviation of the log ratios. For example, for 3-hour averages at Clifty Creek (1975), the ratio is 1.21 while the standard deviation for the log ratios is 0.17. The test statistic resulting from this calculation (log(l.21)/0.17) is 1.1 which is somewhat less than 1.96, the classical point of significance at the 5 percent probability level. Another way of looking at this result is that MPTER slightly overpredicts the measured value (by 21 percent); however, this degree of overprediction is well within the uncertainty inherent in the calculated ratio.

Since the 3-hour and 24-hour ratios are not independent of each other (positive correlation exists), it is more appropriate to test the two averaging periods simultaneously. The chi-square statistic is used to test If the two observed ratios for 3-hour averages (1.21) and 24-hour averages (0.86) are from an unbiased ratio distribution with means of 1.0 and correlation coefficient of 0.4. The results indicate a chi-square of only 3.4 which is not significant at the 5 percent level. The results for the other five data bases indicate that a detectable model bias exists only for Clifty Creek (1976) and for Muskingum River (1976). For Clifty Creek 1976, there is some indication that the 3-hour highest values are significantly overpredicted while at Muskingum River 1976, there is some indication that the 24-hour highest values are significantly underpredicted.

To estimate the composite bias among the data bases, the individual measures of bias (ratios) are combined by weighting each by the reciprocal of its variance. The bottom of Table 1, indicates the results of this calculation. For 3-hour highest concentrations, the composite ratio is 1.09 with composite log standard deviation of 0.06. Since the test statistic is small (((log(1.09))/0.06 = 1.3), the bias statistic is within the estimated error limits. This may be interpreted to mean that over the six data bases, MPTER overpredicts by approximately 9 percent;

however, this degree of overprediction is not statistically significant. For 24-hour highest concentrations, the composite ratio is 0.85. A similar calculation using the observed ratio of 0.85 and standard deviation of 0.07 indicates a modest but significant overall tendency for MPTER to underestimate 24-hour highest concentrations for these power plants.

Model Precision

The precision of the model is defined as the logarithm of the standard deviation of the bootstrap ratios. Large standard deviations translate into large uncertainty in the bias ratio while small standard deviations translate into small uncertainties in this ratio. This uncertainty can be quantified in the form of confidence limits for the actual bias ratio. Table 2 presents the actual ratio for each of the two averaging periods and data bases along with the 95 percent confidence limits for the ratio. To develop table 2, it was assumed that the logarithm of the ratios is distributed normally with mean equal to the logarithm of the actual ratio. This assumption was validated from visual examination of log-normal plots for the 100 bootstrap ratios. The confidence limits may be interpreted along with the results presented in table 1, as evidence of model bias (or lack of bias), or as particular percentiles of the ratio distribution. For example, the lower and upper 95 percent confidence limits for 3-hour averages at Clifty Creek for 1975 are 0.86 and 1.69, respectively. Since both the lower and upper limit encompass 1.00, there is no evidence to suggest significant model bias. The confidence limits for 3-hour averages at Clifty Creek (1975) indicate overprediction while the confidence limits for 24-hour averages at Muskingum River (1976) indicate underprediction since the value of 1.00 is not included within either pair of limits. Not surprisingly, these confidence limits are consistent with results from applying the Chi-square test which simultaneously tested the assumption that 3-hour and 24 hour ratios were equal to 1.0.

Determining NAAQS Compliance Probabilities

While this use of precision is important in assessing the uncertainty of model performance measures, it can also be used for estimating the uncertainty of model-based concentrations. The Calculation of Uncertainty (CUE) technique provides a method for integrating "best estimates" from the model and modeling uncertainty to project the probability of NAAQS attainment/non-attainment for specific large point sources. Simply stated, CUE combines the probability distribution of ratios with the model based design concentration to estimate the probability distribution for monitor-based design concentrations. The methodology has only been tested in a site specific application using the 1975 Clifty Creek model evaluation data base; however, the principles are directly applicable to other plants for which comprehensive model evaluation data are available. While a full application of CUE is outside the scope of this analysis, the CUE computations can be simplified for illustration purposes(**). Calculation of the probability of NAAQS attainment requires estimates of model bias and precision and the "best estimate" of the design concentration for the source. With this information, the probability of attainment is calculated as:

$$PA = \text{Probnorm}((\log(BR/DV))/LSD)$$

where PA = Probability of Attaining the Standard

BR = Bias Ratio (Predicted/Measured)

DV = Ratio of Model "Best Estimate" to Ambient Standard

LSD = Log standard deviation of Ratios (Precision)

Probnorm = Normal Probability Function

(**) CUE uses the entire probability distribution of bootstrap ratios to compute probability of attainment and to derive emission limits consistent with prescribed probability limits. Model bias and model precision, even when site specific, may not be sufficient to completely characterize this probability distribution.

For illustration purposes, five values of the bias ratio are assumed including the specific composite bias for 3-hour averages (1.09), for 24-hour averages (0.85) and for an unbiased model (1.00). The model-based design concentrations are expressed in terms of the ratio of the “best estimate” to the level of the ambient standard. A range of values are assumed from 0.7 up through 1.3. For model precision, a value of 0.20 is assumed which is slightly larger than the values at each of the data bases except Kincaid for which precision was somewhat larger (0.29). Results of these calculations are summarized in Table 3.

For a given 'best estimate", the probability of attainment <u>increases</u> as the bias ratio increases toward greater model overprediction. For example, consider the case when the "best estimate" is exactly equal to the level of the standard (DV=1.0). This example simulates the situation where the decision-maker chooses an emission limit such that the standard is just barely achieved. For a bias ratio of 0.85 (model underpzedicts by 15 percent), the actual probability of attainment is only 21 percent. For a bias ratio of exactly 1.00 (model is unbiased), the probability of attainment is exactly 50 percent. For a bias ratio of 1.09 (model overpredicts by 9 percent), the probability of attainment increases to 67 percent.

For a given bias ratio, the probability of attainment <u>decreases</u> as the "best estimate" increases. For example, consider the case for which the model is unbiased (bias ratio=1.0). When the emission limit is such that the model "best estimate" is 70 percent of the level of the standard (DV=0.7), the probability of attainment is 96 percent. When the model "best estimate" is exactly equal to the standard the (DV=l.0), the probability is again exactly 50 percent. When the emission limit is such that the model "best estimate" is 30 percent greater than the standard (DV=1.3), the probability of attainment is lowered to 9 percent.

The equation above can be reexpressed so that a fixed probability of attainment may be specified and used to calculate the design value ratio consistent with that probability.

DV = BR*EXP(- LSD*Probit(PA))

whare Probit = Inverse Hormal Probability function

Table 4 presents a summary of Design Value ratios for fixed probability of attainment values ranging from SO percent through 95 percent. For example, if the model is assumed to be unbiased (bias ratio=1.0) and the decision-maker wants to be at least 70 percent sure of attainment, then the emission limit for the source must be such that the model "best estimate" is 90 percent of the ambient standard or less. If the decision-maker requires the probability of attainment to be at least 95 percent, then the emission limit must be lowered such that the "best estimate" is 72 percent of the applicable standard.

limitations

The validity of these calculations is obviously limited by the extent to which the composite bias and precision data are applicable to other sources and environmental conditions. For sources with similar operating conditions, stack heights and meteorological regimes, use of the accuracy information derived from the data at these four power plants should be reasonably valid. Perhaps more importantly, it must be assumed that the source being modeled will be operated in a manner consistent with the way the "best estimate" was derived. Since emission limits traditionally established using modeling are conservative (constant maximum emissions, full load, etc), the probability of attainment calculated would also tend to be conservative.

Summary

The analysis of comprehensive data collected around four major SO2 emitting power plants, indicates that EPA’s MPTER model is unbiased with respect to 3-hour average design concentrations and indicates modest but statistically significant underpredictions of 24-hour design concentrations. The best estimate of the composite bias of the model as measured by the ratio of predicted to measured values is 1.09 for 3-hour averages and 0.85 for 24-hour averages.

Accumulated information about the bias and precision of the MPTER model may be used to estimate the probability of attainment of ambient standards around large isolated sources. The simple procedure for calculating this probability is based on the lognormal distribution of model-based to measurement-based design value ratios. The only required inputs are model bias, model precision and the "best estimate" for the NAAQS application. Example calculations indicate that current practice of setting emission limits such that model “best estimates" are just equal to the level of the standards results in attainment probabilities less than 50 percent for a model that underpredicts and over 50 percent for a model that overpredicts. More importantly, the procedure may be used to establish an emission limit that results in achievement of the ambient standards with a predefined probability level.

Ratio of MPTER Based to Measurement Based Design Concentration
For Rural Power Plants

Data Base	3-Hour Averages Bias Ratio	 Std. Dev.	Correlation Coefficient	24-Hour Averages Bias Ratio	 Std. Dev.	Chi-Square
Clifty Creek (1975)	1.21	0.17	0.4	0.86	0.15	3.4
Clifty Creek (1976)	1.45	0.16	0.5	1.09	0.15	6.3 ***
Muskingum River (1975)	1.25	0.15	0.4	1.12	0.19	2.3
Muskingum River (1976)	0.90	0.12	0.5	0.67	0.13	10.4 ***
Paradise (1976)	1.07	0.15	0.5	0.78	0.18	3.7
Kincaid (1980/81)	0.65	0.29	0.7	0.74	0.27	2.2
Composite Results	1.09[1]	0.06[2]	---	0.85[1] ***	0.07[2]	---

Note: Results are based on 100 Bootstrap Replications. Statistics are computed using the logarithm of bootstrap ratios. The design concentration is calculated by fitting an exponential distribution to the highest 25 values to minimize the effect of outliers. The maximum design value from among all the monitoring stations in the network was used to determine the ratio.

[1] Computed as the weighted sum of the individual ratios--weights equal to the reciprocal of the square of the standard deviations

[2] Computed as the square root of the reciprocal of the sum of the reciprocals of the square of the standard deviations

*** Significant at the 5 percent level, indicating overprediction or underprediction. The composite bias for 24-hour averages (0.85), indicates modest underprediction since the "z" score falls below -1.96. $z = (\log(0.85))/0.07 = -2.32$.

TABLE 2

Confidence Limits (95%) for the Ratio of MPTER Based to Measurement Based Highest Concentration for Rural Power Plants

	3-Hour Averages			24-Hour Averages		
Data Base	Lower	Ratio	Upper	Lower	Ratio	Upper
Clifty Creek (1975)	0.86	1.21	1.69	0.64	0.86	1.16
Clifty Creek (1976)	1.07	1.45	1.97	0.81	1.09	1.47
Muskingum River (1975)	0.94	1.25	1.67	0.77	1.12	1.63
Muskingum River (1976)	0.71	0.90	1.13	0.52	0.67	0.86
Paradise (1976)	0.79	1.07	1.44	0.55	0.78	1.10
Kincaid (1980/81)	0.37	0.65	1.15	0.44	0.74	1.26
Composite	0.97	1.09	1.23	0.74	0.85	0.98

Note: Confidence limits are based on 100 Bootstrap Replications. The limits are centered on the actual ratio of MPTER-based to monitor-based design concentrations assuming a lognormal distribution of the bootstrap ratios. The design concentration is calculated by fitting an exponential distribution to the highest 25 values to minimize the effect of outliers. The maximum design concentration from among all the monitoring stations in the network is used to determine the ratio.

Table 3

Probability of Attainment of Ambient Standards for Large SO_2 emitting Power Plants

Best Model Estimate (DV)	Probability of Attainment (Percent)				
	Bias Ratio (Model/Monitor)				
	0.50	0.85	1.00	1.09	1.50
0.7	5	83	96	99	100
0.8	1	62	87	94	100
0.9	0	39	70	83	99
1.0	0	21	50	67	98
1.1	0	10	32	48	94
1.2	0	4	18	32	87
1.3	0	2	9	19	76

Note: Assumes that the ratio of model based to measurement based design concentrations is lognormally distributed with log mean equal to the bias ratio and log standard deviation = 0.20. The "best estimate" (DV) is expressed as the ratio of the model based design value to the level of the ambient standard.

Table 4

Model Based Design Values Required to Achieve A Given Probability of Attainment For Large SO_2 Emitting Power Plants

Probability of Attainment (percent)	Design Value (DV)				
	Bias Ratio (Model/Monitor)				
	0.50	0.85	1.00	1.09	1.50
50	0.50	0.85	1.00	1.09	1.50
55	0.49	0.83	0.98	1.06	1.46
60	0.48	0.81	0.95	1.04	1.43
65	0.46	0.79	0.93	1.01	1.39
70	0.45	0.77	0.90	0.98	1.35
75	0.44	0.74	0.87	0.95	1.31
80	0.42	0.72	0.85	0.92	1.27
85	0.41	0.69	0.81	0.89	1.22
90	0.39	0.66	0.77	0.84	1.16
95	0.36	0.61	0.72	0.78	1.08

Note: Assumes that the ratio of model based to measurement based design concentrations is lognormally distributed with log mean equal to the bias ratio and log standard deviation = 0.20. The "best estimate" (DV) is expressed as the ratio of the model based design value to the level of the ambient standard.