UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
REGION 10
August 14, 1985

REPLY TO ATTN OF: M/S 329

MEMORANDUM

SUBJECT: Dispersion Modeling of Fugitive Emissions Sources at Bunker Limited - Protocol for Model Evaluation

FROM: David C. Bray, Acting Chief and Robert B. Wilson
Air Programs Development Section Regional Meteorologist

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

In our previous memo to you (July 29, 1985), we explained an alternative methodology we developed for modeling fugitive lead emissions from the Bunker Limited lead smelter in northern Idaho. In this memo we propose a protocol to evaluate whether or not this approach improves the model performance for predicting measured ambient concentrations.

The goal of this evaluation is to determine which model methodology, the LONGZ model with multiple volume sources or the LONGZ with a single volume source for fugitive sources, is best for assessing compliance with the National Ambient Air Quality Standard (NAAQS) for lead: 1.5 micrograms per cubic meter, 3-month (calendar quarter) average. We must test the relative spatial and temporal accuracies of the methods since we will use the better method for determining the appropriate design concentration and the effectiveness of proposed control strategies. Thus, all comparisons of predicted and observed concentrations will be paired in time and space.

Under this goal we have two objectives of equal significance. The first objective is to determine which technique is better able to predict the highest measured concentrations. Since the highest predicted concentration in ambient air becomes the design concentration and will ultimately determine how much control of emissions will be required, it is critical that the magnitude of the highest measured concentrations be predicted accurately. Furthermore, from a regulatory perspective, significant underprediction of maximum concentrations can not be tolerated. We can eliminate from this evaluation the concern for underprediction by imposing the criterion that if either of the methods underpredict the maximum measured concentrations by more than 10%, the method will be rejected.

Since the design concentration may occur at a different time and location than the highest measured concentration, the second objective is to determine which method is a better predictor of all concentrations measured over the entire monitoring network. We need to have confidence that we will be using the better technique for accounting for temporal and spatial variations of concentration. In the point scoring scheme, the 100 possible points will be divided evenly between the two objectives.

For the first objective we propose to compare the bias and variablity of the methods for predicting the maximum measured concentration for each of the four quarters for which we have acceptable emissions, meteorology, and air quality data bases. The highest predicted and measured concentrations occurred at the Silver King School monitor (SKS) every quarter. The test of bias is the average residual, $\overline{d}$ (observed minus predicted). We will use a two-tailed t test to first determine if the two average residuals are significantly different. If they are not different at the 95% confidence level, neither method can be considered better in terms of bias, and no points will be awarded. If the average residuals are significantly different, then the method with the minimum average residual will be assigned 35 points.

The test of variability is the variance, s^2. As with the bias, we will first test to see if the variances are significantly different from one another at the 95% confidence level. For this test we will use the two-tailed F test. Again, if the variances are not significantly different, no points will be awarded. If they are different, the method with the lowest variance receives 15 points. It is recognized that in the tests for this first objective, the number of data pairs is low (only four pairs).

The same statistical tests can be made for the second objective. In this case we have 36 pairs of predicted and observed concentrations, four quarters at nine monitoring stations. In addition to $\overline{d}$ and s^2, we test correlation with the Pearson correlation coefficient, r. The 95% confidence intervals for the r values are determined using the Fisher z test. If the confidence intervals overlap, the correlations can not be considered significantly different, and no points will be assigned. However, if they are significantly different, the method with the higher correlation will receive 10 points.

(Note: all statistical tests are referenced to Statistical Methods, Snedecor and Cochran, Sixth Edition, 1967.)

The evaluation scheme can be summarized as follows:

First, if $(\overline{C}_p / \overline{C}_o)_{SKS} < 0.90$, the method will be rejected. If both methods pass this criterion, then the following tests will be performed.

OBJECTIVE	TEST	PARAMETER	SIGNIFICANCE TEST	POINTS*
1	Bias for maximum measured concentrations	$(\overline{d})_{SKS}$	Two-tailed t test (95%) 3 degrees of freedom	lower gets 35
	Variability for maximum measured concentrations	$(s^2)_{SKS}$	Two-tailed F-test (95%) 3 degrees of freedom	lower gets 15
2	Bias for all measured concentrations	$(\overline{d})_{All}$	Two-tailed t test (95%) 35 degrees of freedom	lower gets 25
	Variability for all measured concentrations	$(s^2)_{All}$	Two-tailed F test (95%) 35 degrees of freedom	lower gets 15
	Correlation	$(r)_{All}$	Fisher z test - 95% confidence intervals	higher gets 10

* Points are awarded only if parameters are significantly different.

The method with the greatest number of points will be used in the modeling analysis for control strategy evaluation. However, the single volume source method must receive at least 35 points to be selected over the multiple source method since the latter corresponds with current modeling guidance concerning fugitive source emissions. In this evaluation we are comparing one method against another, rather than comparing a model against an absolute set of performance standards. Thus, points are awarded on a relative basis, one model doing significantly better than the other, rather than based on achieving a certain level of performance. Also, the confidence level (95%) is only a rough approximation to "truth" for comparing the performance of the two methods, since the performance measures are not independent of one another.

We would appreciate your concurrence with our proposed evaluation protocol as soon as possible. We will also generate the complete set of statistics which we used to compare the performances of the LONGZ Urban/Rural, LONGZ Rural/Rural, and the ISCLT Models in our original model selection process. These statistics which were presented in Rob Wilson's memo to you of May 23, 1984, are for information to be consistent with our previous evaluation work.

cc: G. Abel