1.2 Comments on a Draft Practice for Statistical Evaluation of Atmospheric Dispersion Models

John S. Irwin[* **]
Atmospheric Sciences Modeling Division, Air Resources Laboratory,
National Oceanic and Atmospheric Administration, Research Triangle Park, North Carolina
and
Dr. Mihail-Radu Rosu
Atmospheric Modeling Division, National Exposure Research Laboratory,
U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

1. INTRODUCTION

Within the American Society for Testing and Materials (ASTM), there is currently a draft Standard Practice (Z6849Z[1]) under consideration which describes an objective statistical procedure for comparing air quality simulation modeling results with tracer field data. The practice is limited to local scale (first tens of kilometers) transport and dispersion from isolated point sources in simple terrain situations. The practice describes how comparisons might be made of simulated centerline concentration values with observed concentrations from receptors near the observed center of mass along sampling arcs. The goal of the practice is to define which of several dispersion models has the least bias in estimating the centerline maximum concentration, and whether the differences seen between models is statistically significant. As discussed in the practice, statistical evaluation of model performance is viewed as part of a larger process that collectively is referred to as model evaluation. It is assumed that through use and experience, the practice can be extended to assess performance for other features in the concentration pattern (crosswind integrated concentration, lateral dispersion, etc.).

A major consideration in developing the statistical comparison measures was that operational dispersion models provide estimates of the average concentration for the specified meteorological conditions. Another major consideration in developing the statistical comparison measures was that differences seen in comparisons of model predictions and observations of atmospheric air concentrations may largely reflect an inherent uncertainty caused by the stochastic nature of turbulence within the atmosphere. This component of the variance was considered inherent because it cannot be reduced significantly by improving the physics of the air quality models.

To address these considerations, the practice stratifies the evaluation data into regimes, where one can reasonably argue that the physical processes affecting the dispersion are similar. A regime is an estimate of an ensemble. Here ensemble refers to the infinite population of all possible realizations and is developed from a set of experiments having fixed external conditions, which in practice we have but a small sample to work with. Model performance is then assessed by the ability of the model to replicate without bias the regime's characteristics (such as the average maximum, average lateral extent, or average crosswind integrated concentration). The practice attempts to assess the significance of differences seen in alternative models' results. The goal of the practice is to (1) determine which model's estimates of the centerline concentration values are least biased, and (2) determine which model's results are significantly different from the results of the model identified in (1). For each regime, we can compare the model's estimates of centerline concentration values with observed values. And from a summary of these results across all the regimes, we can determine (1). For each regime, we can determine the relative difference between models, in their estimates of centerline concentration values. And from a summary of these results across all regimes, we can determine (2).

The conceptual summary provided in the last two paragraphs conceals various problematical issues. For a given experiment and arc, what is an observed centerline concentration? How is bias to be determined? How do we summarize over all regimes? Given the summary statistics over all regimes, how do we accomplish (1) and (2)? In this presentation, examples will be presented of results achieved in testing some of the procedures of the draft practice with tracer field data.

2. DEFINING REGIMES

The practice requires stratification of the evaluation data into regimes, where one can reasonably argue that the physical processes affecting the dispersion are similar. Prairie Grass (Barad, 1958, and Haugen, 1959) included sampling along five arcs downwind from a near-surface point source release of sulfur-dioxide, SO_2. We determined an Obukhov Length, L, from the onsite meteorology for each release. We sorted the 70 experiments from most unstable to most stable, where the stability is defined in terms of 1/L (1/L < 0 being unstable and 1/L > 0 being stable). We then divided the data into six stability groups, which for each we have sampling results for five downwind arcs. This provides us with 30 regimes with results from approximately 11 experiments in each regime. Table 1 list the experiments for each stability group. As analyses progressed, it was determined that four experiments (3, 4, 13, 14) were dispersing in a

* *Corresponding author's address*: John S. Irwin, U.S. Environmental Protection Agency (MD-14), Research Triangle Park, NC 27711
** On assignment to the Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency.
[1] A review copy of the draft practice is available from the corresponding author upon request.

manner significantly different from the rest of the experiments in the most stable stability group. These four experiments had some of the lowest surface winds, and smallest positive L values. The dispersive conditions were such that very little vertical growth occurred. Much of the SO_2 (released 46 cm above the ground) flowed beneath the samplers on the 50- and 100-m arcs, which were 1.5 m above the ground.

3. DEFINING PLUME CENTERLINE POSITION

Given experimental field data, how do we define the transport centerline? Standard plume models available today do not attempt to characterize the serpentine characteristics of actual plume dispersion. These standard plume models attempt to characterize the average lateral and vertical dispersion, assuming steady-state meteorology throughout the network of receptors. The modeled plume centerline position at the surface is also the center of mass at the surface at each distance downwind. In the draft practice the assumption made is that the observed center of mass along an array of receptors at a fixed distance downwind, is a useful approximation of the actual plume centerline position (for this experiment's results). For experiments with a dense network of receptors and for dispersing plumes that are relatively contiguous, using the observed center of mass as an estimate of the plume centerline position is adequate, but as we will discuss later, observations do not always fulfill even these simple constraints

4. OBSERVED CENTERLINE CONCENTRATIONS

If we have evenly spaced concentration values along an arc, and the lateral concentration distribution is Gaussian, then the average of all concentration values within ±0.67Sy, would be 0.93Cmax, where Sy is the computed lateral dispersion and Cmax is the centerline concentration maximum. In the draft practice, all observed concentration values within ±0.67Sy of the center of mass are deemed to be reasonable observations of the centerline maximum. And in the draft practice, Sy is computed using observed concentration values for a given arc for a given sampling period.

There are alternatives to the procedure outlined above. Available evidence suggests that large stochastic variations in concentration values are likely at any position within the dispersing plume. This argues for large sample sizes, in order to be able to discern whether differences seen between observed and modeled average centerline concentrations are significant. If we select one value (say the value nearest the computed position of the center of mass), we will need many experiments in each regime. By allowing more than one value to be selected from each arc, the draft procedure attempts to strike a compromise, where fewer numbers of experiments can be used for each regime, and yet attain reasonable sample sizes. If we attempt to use averaging or fitting techniques to derive Cmax, we may lose the ability to characterize the stochastic variations to be seen in the centerline concentration values. Retaining the ability to characterize the variations in centerline concentrations, means we can extend the practice to allow evaluation of fully stochastic models that attempt to simulate these variations.

The 10-minute concentration values seen along the arcs during the Prairie Grass experiments are from evenly spaced receptors, and the dispersing plume appears reasonably contiguous. In such cases, the computation of the center of mass and Sy is straightforward. However, the 60-minute concentration values seen along the arcs during the Kincaid experiments (Bowne et al., 1983) suggest that the dispersing plume, resulting from a 183-meter stack with a buoyant plume rise on the order of 200 m, presents a difficult sampling problem. From a visual inspection of all of the Kincaid sulfur-hexafluoride, SF6, tracer concentration values, it was decided to test whether Sy would best be computed by analyzing all the data for a regime as a whole.

The lateral dispersion computed over all of the experiments within a regime will differ somewhat from the lateral dispersion computed individually for each arc. To investigate the effects of alternative methods for computing Sy and defining centerline concentration values, we used the Project Prairie Grass data. Figure 1 illustrates one of the effects to be seen in the two alternative methods for defining which concentration values are deemed to be near the centerline (center of mass). In this illustration, Group 4 in Table 1, the open circles are the concentration values for which $|y/Sy| \leq 0.67$, when Sy is defined individually for each experiment. The solid circles are the concentration values for which $|y/Sy| \leq 0.67$, when Sy is defined for the group of 12 experiments. The lateral dispersion for Test 24 is a bit broader than the average for the group, and the lateral dispersion for Test 38 is a bit narrower than the group.

Table 2 summarizes the average and standard deviation of the near centerline concentration values, C, using the alternate definitions of crosswind position. The slight differences seen are not statistically significant. If we combine Group 6 and 7, then much larger differences are seen. This reinforces that care must be exercised to only group together results that are comparable. History suggest that new comprehensive experiments will not be generated rapidly. Rather than attempting to devise objective criteria to cover all contingencies, we suggest experience and working with the data will provide effective ways to sort available data into useful regimes.

5. AVERAGES VERSUS PERCENTILES

The draft practice tests a model's performance to simulate without bias the median of the observed near centerline concentration values. The measure of bias is a fractional bias computed as FB = 2(E-O)/(E+O), where E and O are the estimated and observed median centerline concentration values for each regime. The practice relies on bootstrap resampling to estimate a standard deviation for the

computed FB. Summarizing results across all regimes involves using computed values of the average fractional bias, AFB, and its standard deviation. For summarizing over all regimes, this involves computing a pooled variance from the results obtained for AFB for each regime. This final step of summarizing across all regimes is a critical step, and fundamental to providing an objective means for testing whether differences seen are statistically significant. The practice relies on standard methods for pooling results over all regimes. These methods in turn rely on the assumption that the bootstrap samples for the AFB are well described by an average and a standard deviation, i.e. they are approximately normal distributions.

The middle panel of Figure 2 illustrates the distribution of median values generated by bootstrap resampling from a collection of centerline concentration values. The median is determined by creating a sample (with replacement) of N values of centerline concentration values. N is the number of centerline values from all experiments in the regime. Each experiment in the regime has equal probability of being selected. If more than one value is available for a selected experiment, one value is chosen at random. The examples shown in Figure 2 are for 1000 samples of N values each. It is obvious that the generated distributions for the 50-th and 90-th percentile values are decidedly not Gaussian. The lower panel of Figure 2 illustrates the distribution of average values generated by bootstrap resampling from the same collection of centerline concentration values. These results are typical of results obtained for other arcs and stability groups of Project Prairie Grass data.

It was concluded from these results that development of confidence bounds using sample standard deviations would not be appropriate for the comparison of percentile values. Whereas, the distribution of averages is well characterized using sample standard deviations.

6. NUMBER OF SAMPLES

In the above analysis, distributions were generated of sample percentile values and of sample averages. In these investigations 1000 samples were used. The question arises, how many samples is enough? Efron and Tibshirani (1993) suggest that bootstrap samples in the range of 50 to 200 usually provide good standard error estimates. We choose to address this question by conducting a numerical experiment.

Using bootstrap sampling as described above, we computed the standard deviation, Std, of the centerline concentrations for the Prairie Grass data for each regime (arc and stability group) using a sample of 10 values. Ten bootstrap samples of size ten were developed for each regime. We then computed the average, Avg(Std), and standard deviation, Std(Std), of the sample standard deviations, Std, from the 10 values generated for each regime. The ratio of Std(Std)/Avg(Std) for samples of 10 values ranged from 0.118 to 0.325. This ratio is expected to decrease as $1/\sqrt{n}$, where n is the bootstrap sample size. We can then solve for the value of n such that the ratio of Std(Std)/Avg(Std) is less than some desired tolerance, say 0.05. This is found to be satisfied if $n \geq 10*(0.325)^2/(0.05)^2 = 423$.

7. WITHIN ARC CORRELATION

In the draft practice, each experiment within a regime is considered equally probable. But should we consider the concentrations selected from each experiment (for which in most cases there are more than one) to be independent? They may be correlated in space? As discussed by Young (1994), there is very limited empirical study of bootstrap procedures for dependent data. To test for such correlation we conducted the following numerical experiment.

We computed using two sampling methods the average and the standard deviation of the average for each regime using 500 bootstrap samples. In the first method, we used a sample of one centerline concentration value at random from each selected experiment within a regime In the second method, we used a sample of two centerline concentration values (a pair) at random from each selected experiment within a regime. A pair was defined as two centerline concentration values that were adjacent to one another (in position) on the arc. In method one, random samples of one are selected until the number of samples equaled N, where N represents the number of centerline concentration values from which to select. In method two, random samples of two are selected until the number of samples equaled 2*INT(N/2), where INT refers to the integer value.

7.1 Individual versus group Sy

There were 30 regimes (six stability groups and five arcs). For these 30 regimes, we compared the averages generated by method one sampling for each regime, when individual Sy values were used to define the centerline concentration values (see Section 3) versus group Sy values. None of the differences seen in the computed averages were statistically significant at the 5% confidence limit. The same conclusion was reached when we compared the averages generated using method two (pairwise) sampling, individual versus group Sy values.

We tested to see if the bootstrap computed variances were different. First we compared whether the differences seen, using individual versus group Sy values, in the method one variances were significant. For four of the 30 regimes the differences were significant at the 5% confidence limit. Then we compared whether the differences seen, using individual versus group Sy values, in the method two (pairwise) variances were significant. For two of the 30 regimes the differences were significant at the 5% confidence limit.

We concluded that computed averages and variances were not significantly affected by whether the position of the receptors was defined using an individual or a group determined Sy value. These

results confirmed conclusions reached in Section 4.

7.2 Sample of one or a pair

We compared method one versus method two averages generated for each regime when individual Sy values were used. In general, the differences seen in the averages generated by method one and method two sampling were of order 2% or less. None of the differences seen were significant at the 5% confidence limit.

We compared the variances generated by method one versus method two sampling for each regime when individual Sy values were used to define the centerline concentration values. Differences seen in 15 of the 30 comparisons were significant at the 5% confidence limit. Most of these (9) were for results obtained for the stability groups 1 and 2. We compared the variances generated by method one versus method two sampling for each regime when group Sy values were used to define the centerline concentration values. Differences seen in 14 of the 30 regimes were significant at the 5% confidence limit. Again, most of these (10) were for results obtained for stability groups 1 and 2. In general, the differences seen in the standard deviations generated by method one and method two sampling were of order 35%. There is a trend in the relationship between the method two (pairwise) and method one standard deviations. For the unstable stability groups (1-3), method two sampling (pairwise) generates larger standard deviation values. And for the more stable stability groups (4-6), method two (pairwise) sampling generates smaller standard deviations. These differences may reflect correlation effects being preserved by sampling pairs of values (method two), instead of sampling individual values (method one). At this time it is not known, whether the differences seen by method one and method two sampling would alter conclusions reached in assessing differences in performance between alternative dispersion models. This will be a topic for assessment as we further test the draft ASTM practice.

8. CONCLUSIONS

We conclude that in defining which receptor positions are to be deemed to provide representative observations of the centerline concentration maximum, the receptor's position relative to the plume center of mass can be defined using a lateral dispersion derived for the regime. This should be robust to slight inadequacies in sampling. The average of the centerline maximum concentration values and bootstrap derived standard deviation was seen to be well behaved, versus results obtained for individual percentile values of the frequency distribution of centerline concentration values. We concluded that the average is a better statistic to use in the evaluation procedures in comparison to the current draft's suggestion to use percentile values. We illustrated a numerical experiment that can be used to assess how many bootstrap samples may be needed. Our results suggested bootstrap sample sizes ranging from 70 to 400. In developing the bootstrap samples, we tested two replicate sampling methods: sample of one versus sample of a pair. This was done to assess whether observations from adjacent receptors could be treated as independent. We saw significant differences in the results obtained, and conclude that further testing is needed to assess the effect of these differences when using these sampling techniques within the context of the draft ASTM practice to assess model performance.

Next steps includes testing how results obtained from each regime can usefully be summarized over all regimes. And once this is accomplished, we will test how well the procedures assess differences between models. Once these tests are completed, the practice can be redrafted and resubmitted for ASTM committee review and balloting.

9. DISCLAIMER

The information in this document has been funded in part by the United States Environmental Protection Agency under an Interagency Agreement (DW13937039-01-06) to the National Oceanic and Atmospheric Administration. It has been subjected to Agency review for approval for presentation and publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

10. REFERENCES

Barad, M.L. (Editor), 1958: Project Prairie Grass, A Field Program In Diffusion. Geophysical Research Paper, No. 59, Vol I and II, Report AFCRC-TR-58-235, Air Force Cambridge Research Center, 439 pp.

Bowne, N.E., R.J. Londergan, D.R. Murray and H.S. Borenstein, 1983: Overview, Results, and Conclusions for the EPRI Plume Model Validation and Development Project: Plains Site. EPRI EA-3074, Project 1616-1, Electric Power Research Institute, Palo Alto, CA. 234 pp.

Briggs, G.A., 1982: Similarity forms for ground-source surface-layer diffusion. *Boundary-Layer Meteorology*. Vol. 23, pp. 489-502.

Efron, B., R.J. Tibshirani, 1993: *An Introduction to the Boostrap.* Chapman & Hall, New York, NY. 436 pp.

Haugen, D.A. (Editor), 1959: Project Prairie Grass, A Field Program In Diffusion, Geophysical Research Papers, No. 59, Vol III, AFCRC-TR-58-235, Air Force Cambridge Research Center, 673 pp.

Young, G.A., 1994: Bootstrap: More than a stab in the dark? *Statistical Science*. Vol. 9(3):382-415.

TABLE 1. Groupings of experiments used in analyses. Group 1 has 11 experiments with most unstable stability defined by 1/L, where L is the Obukhov length. Groups 1 - 3 have L < 0, and Groups 4 - 7 have L > 0. Groups 6 and 7 have 11 experiments with most stable stability defined by 1/L. Group 7 are four experiments for which vertical dispersion was very limted, such that much of the SO_2 passed beneath the 1.5 m above-ground samplers along the 50-m arc. These four experiments have been noted by other researchers as being unique (see Briggs, 1982).

Stability Group	Number	Experiments
1	11	1, 2, 7, 10, 15, 16, 25, 43, 47, 52, 48S
2	11	5, 8, 9, 19, 26, 27, 44, 49, 50, 51, 62
3	12	6, 11, 12, 20, 30, 31, 33, 34, 45, 48, 57, 61
4	12	21, 22, 23, 24, 37, 38, 42, 46, 55, 56, 67, 35S
5	11	17, 18, 28, 29, 41, 54, 59, 60, 65, 66, 68
6	7	32, 35, 36, 39, 40, 53, 58
7	4	3, 4, 13, 14

TABLE 2. Summary of statistics computed for each stability group (see Table 1) for centerline concentration values, C/Q, computed for Project Prairie Grass 100-m data, where Avg = average, Std = standard deviation, N = number of values, L = Obukhov length (m), and Sy is group's lateral dispersion (degrees). Concentrations, C, are divided by emission rate, Q, and have units of $X10^3$ s/m^3. Individual results have receptor positions defined using lateral dispersion values computed for each experiment and arc. Group results have receptor positions defined using one lateral dispersion computed for all experiments within the group.

		STABILITY GROUP							
		1	2	3	4	5	6	7	6+7
Individual	Avg (C/Q)	0.26	0.54	0.56	1.89	3.59	8.46	3.32	5.17
	Std (C/Q)	0.19	0.20	0.17	0.64	2.32	5.38	3.31	4.85
	N	125	70	60	30	27	17	30	47
	Avg (L)	-9	-32	-72	140	30	7	4	5
Group	Avg (C/Q)	0.30	0.55	0.58	1.93	4.02	9.86	3.78	5.32
	Std (C/Q)	0.19	0.19	0.17	0.66	2.44	4.71	3.57	4.36
	N	125	70	63	33	31	15	37	70
	Sy	16.9	9.4	7.9	4.2	4.3	3.4	14.0	9.7

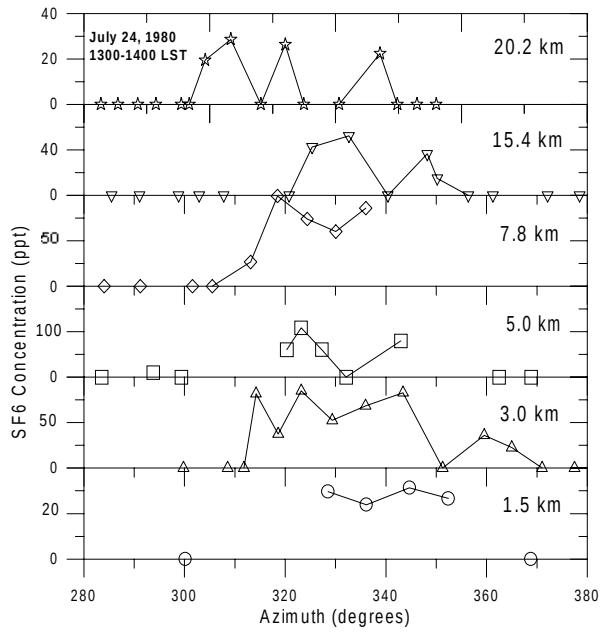

Figure 1. Observed 1-hour SF6 concentration values for July 24, 1980 of the Kincaid experiments.

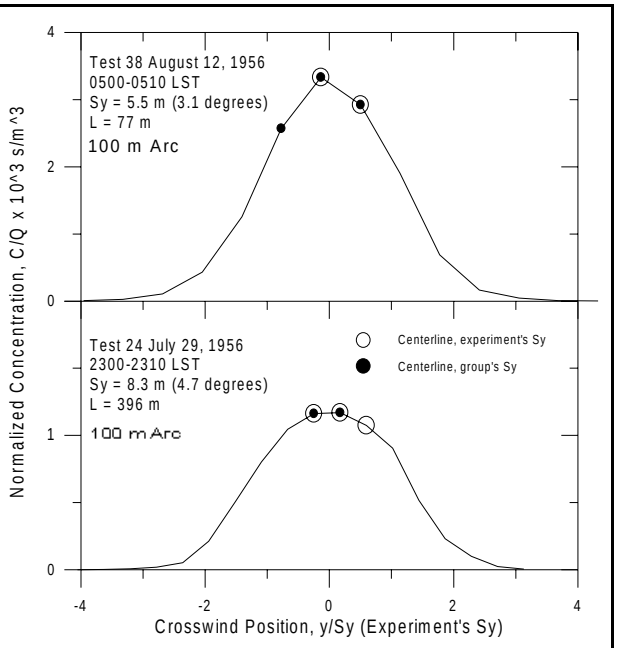

Figure 2. Illustration of how use of different methods to compute lateral dispersion will affect definition of centerline concentration values.

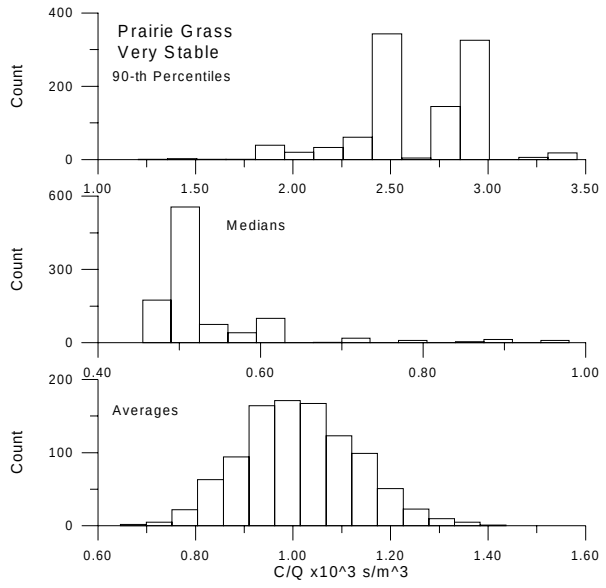

Figure 3. Examples of historgrams generated by resampling.

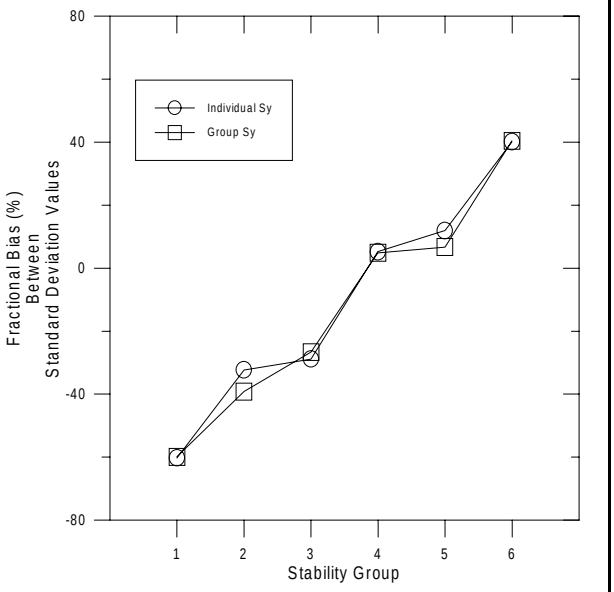

Figure 4. Average frational bias between method one and method two standard deviations as a function of stability group.