March 8, 1999

Normalized Centerline Concentration Listings

I have been working for the last several years to test, develop and demonstrate a method for comparing modeled and observed centerline concentration values. To facilitate those who wish to 'play in the game', but who have no desire to try to use my admittedly non-friendly software, I have created three special files.

1. File Names

Filename	File Used to Define the Regimes	Number of Model's Results Listed
Kinlist.out	Khpdm3.txt	7
Indlist.out	Indpes.txt	5
Pgrlist.out	Prgpes.txt	5

(You can add one (1) to the number of 'models', if you believe, as I do, that the maximum concentration along the arc is a biased estimate of the true centerline average.)

2. Structure Of Information In Files

There is a one line header at the top of each of the files, which defines the output listed. The header labels are:

YYMMDDHHMM= year, month, day, and ending time (hours and minutes) of sampling time,
Regime = Group of data (assumed to have similar meteorological conditions), always at same (similar) distance from source,
Exp = Experiment number used in storing the data,
Arc = Arc number (increases in sequence as distance downwind increases),
n = number (count) of this value for this experiment and arc combination,
ni = total number of values from this experiment and arc combination,
Grp_Sy = lateral dispersion (m) computed using all receptors from all arcs in this regime,
Xbar = average distance downwind (km) for this arc for this experiment,
y = +/- distance (m) receptor is located from center of mass of all observed concentration values for this regime,
y/sy = y divided by the regime lateral dispersion (dimensionless),
q = emission rate (g/s)
obs/q = observed concentration divided by emission rate [units of measurement/(g/s)], the units of measurement was PPT for Kincaid and Indianapolis, and was milligrams per cubic meter for Prairie Grass,
max/q = maximum concentration along the arc divided by emission rate,

AERMOD1 = AERMOD (version 97064) concentration divided by emission rate,
AERMOD2 = AERMOD (version 97248) concentration divided by emission rate,
AERMOD3 = AERMOD (version 98022) concentration divided by emission rate,
ISC3 = Industrial Source Complex concentration divided by emission rate,
HPDM1 = HPDM (met as delivered to PES, used onsite Zi and onsite sigma-v) divided by emission rate,
HPDM2 = HPDM (met as delivered to PES, used calculated Zi and sigma-v) divided by emission rate,
HPDM3 = HPDM (AERMOD meteorology and used calculated Zi and sigma-v) divided by emission rate.

3. Philosophy Of Model Evaluation

The draft ASTM standard espouses a philosophy (or statistical model) in the treatment of the observations and the model estimates. This philosophy is fundamentally different than that employed in current methodologies, such as Cox and Tikvart (1990) or Model Evaluation Kit (Olesen, 1996).

For most operational plume (puff) models of today, the model formulations employ average characterizations of plume (puff) rise, and the vertical and lateral growth of the plume (puff) as it disperses downwind, all of which are often empirically derived. The meteorology is often assumed to be horizontally homogeneous, hence we can envision it as the average meteorological condition to be experienced as the plume (puff) disperses downwind. With this basis of the physics within the models, it is my opinion (and that espoused in the draft ASTM standard) that the models are attempting to estimate the average concentration one should see (if you were lucky enough to see these conditions many times and average all the observations together). Whereas, the observations are individual realizations, that belong to a population of realizations, the average of which is what the models are attempting to simulate.

If you think about this for a moment, this means that if you compare a collection of observations and modeling results, the only point we can hope they will agree (with zero bias) is a comparison of the averaged observed and average modeled result. **It means that the observations and the modeling results come from two different statistical populations, whose means are (hopefully for an unbiased model) the same**. The variance seen in the modeled values results from differences between *realizations of averages*. The variance seen in the observations results from that seen in the modeled values plus an additional variance caused by stochastic variations between *individual realizations*, which is not accounted for in the modeling. If the observed and modeled variances agree, this has to be happenstance, since the model does not have the physics (or empirically based statistical models) for characterizing the fluctuations to be seen in the dispersion from one realization to the next[1].

[1] There are a few models that do have added physics (or empirical algorithms) to allow the models to attempt to simulate the fluctuations to be seen in the dispersion from one

4. **Consequences - How Do You Create 'Averages' From Observations?**

If you agree with the above paragraphs, there are some interesting problems that have to be solved in order to conduct a 'proper' comparison of observed and modeled ambient concentration values. First of all, the observations must be grouped together in some fashion so that you can create from this grouped data an 'average' for comparison with the modeling result. What criteria should be used to group the data? How do you group the data, as the emission rates may be different, the direction of transport may be different, etc?

4.1 Beware of direct comparisons of observations and modeling results

There are other consequences as well. If the observations and modeling results come from different populations, than you will arrive at erroneous conclusions if you make statistical comparisons by directly pairing the modeling results with the observations. To illustrate this, I created a program called Gauss09.for. It is included in this package with the data listings. To run this program, extract gauss09.exe to an empty directory. You can either run the program from the dos prompt (by typing gauss09 <return>), or you can run the program by double clicking gauss09.exe in Windows Explorer. It will prompt you for several inputs. The output from the run is capture in a file called gauss09.dat. If you run this program you can create samples of observations and modeling results (from log-normal distributions). You can control how many samples will be in each group and the number of groups. At the end of the file that captures all the output (Gauss09.dat), there is a summary table which compares Fractional Bias and Absolute Fractional Bias (AFB), computed using observation and model pairs and computed using the observed and modeled averages for a group. Notice that even when the observations and the model have exactly the same mean (or distributions), the AFB computed using paired observations with modeling results is deceptive, as it says there is a greater difference in the population averages than really exist. This has to do with how an AFB is computed, but a similar result can occur with most any standard statistical measure of agreement.

This means, by the way, that directly comparing frequency distributions of the observed and model concentrations is very misleading. You may have to think about this a while, but if you had a perfect model, then the maximum centerline concentration simulated would have to be less than observed (because the observation has scatter not in the model's simulation), and the minimum centerline concentration simulated would have to be greater than observed! The upper precentiles of the two cumulative frequency distributions should not agree - so why define model performance based on a comparison of the upper percentile values of the respective cumulative frequency distributions?

4.2 Compare averages with aveages.

This means that in order to best represent how well a model is estimating the 'true' average of the observations, we should be comparing averages of observations and modeled

realization to the next - ADMS and SCIPUFF are two of which I am aware. But such models are few in number, and have never seen service (as yet) in routine air pollution assessments in the USA.

results. Once you have the average observed and average modeled result for each regime, the standard statistical methods for comparison work! Scatter plots, regressions, correlations, comparisons of cumulative frequency plots, etc. But first you have to compute the averages. Are you beginning to see the consequences of this different philosophy?

5. Pooling (Summarizing) Over Several Regimes

The draft ASTM standard adds one embellishment to the computation of the averages, it employs bootstrap sampling to estimate a variance to be associated with the averages computed for each Regime (group). It then computes a pooled average AFB over all regime using a weighted average. The weighted average employed was designed to combine estimates of averages where the uncertainty associated with each of the averages is possibly different. The weight assigned to each regime is the inverse of the variance computed for the average for the regime. As the variance increases, the weight decreases. Regimes with large uncertainty in the estimate of the average are discounted.

As discussed by Meier (1953), an inverse-variance weighted average provides a best estimate of an average, $\overline{\mu}$, over all regimes (or groups), and also provides an estimate of the variance, Var($\overline{\mu}$), of $\overline{\mu}$. Defining the following notation:

$$w_i = \frac{1}{\sigma_i^2}, \quad w = \sum w_i, \quad \theta_i = \frac{w_i}{w} \qquad \textbf{(1)}$$

then:

$$\overline{\mu} = \sum_{i=1}^{R} \theta_i \, u_i, \quad Var(\overline{\mu}) = 1/w \qquad \textbf{(2)}$$

where $\overline{\mu}$ is the average over all R regimes of some measure of performance (like a fractional bias or an absolute value of the fractional bias), and u_i is the average computed for regime i. When the variances, σ_i^2, of u_i are known, then Equation (2) provides the best estimate of the variance of $\overline{\mu}$. When the variances of u_i for each regime are estimated from the data (hence we have s_i^2 versus σ_i^2), then Meier (1953) provides an approximately unbiased estimate of Var($\overline{\mu}$) as:

$$Var(\overline{\mu}) = \frac{1}{\hat{w}}[1 + 4\sum_{i=1}^{R} \frac{1}{n_i}\hat{\theta}_i(1 - \hat{\theta}_i)] \qquad \textbf{(3)}$$

where "^" denotes replacement of σ_i^2 by s_i^2 in Equations (1) through (2), and n_i is the number of values in regime (or group) i.

The current draft of the ASTM practice employs Equation (2) rather than (3) in estimating $\mathrm{Var}(\overline{\mu})$. For large n_i, Equation (3) asymptotically approaches (2). I recently have reconsidered the use of (2), and now am considering (3) as perhaps a wiser choice. For example, for the Kincaid experiments, Equation (2) estimates for AERMOD3 an absolute fractional bias averaged over all 29 regimes of 0.35, and a standard deviation of this average of 0.025. The standard deviation based on (3) is 0.026 (about 4% larger). The average value of n_i for Kincaid was 58. I worry that for other experiments, the differences between (2) and (3) may prove more significant.

Using this inverse-weighted average is fine, if your statistic for assessing performance is fairly stable. But the AFB is not well behaved, if the observation is nonzero and the model predicts zero. In such circumstances the AFB locks itself to a value of -2, its associated estimated variance is nearly (if not) zero, and this regime dominates the assessment of performance when it is pooled together with the other regimes' results. Rather than create a new statistic for comparison, I suggest a different course of action. I would exclude this regime when assessing this model's performance with other models. Yes, the other models may have not estimated zero for this regime in question. But our purpose was to see which models are performing badly, when discerning differences in performance is not obvious (estimating zero when a nonzero average value is seen, is obvious).

6. Conclusion

The current draft ASTM practice argues for a change in philosophy of how atmospheric dispersion modeling results should be compared with observations. There has been some misunderstanding that the draft ASTM practice was attempting to provide a comprehensive evaluation of dispersion model performance. This is not so. Its purpose was only to argue for a change in philosophy, and to help readers understand the consequences of such a change in philosophy, it provides an example of how one might compare estimates of centerline concentration values.

I would like first to see whether the dispersion modeling community endorses the proposed change in philosophy. If this be the case, then we can investigate and test various ways modeling results can be compared with observations, and over time, come to a consensus on a minimum set of evaluation measures (and some kind of summarization procedure over all measures) that provides an objective means for selecting those dispersion models that are performing best.

I hope you join in the discussion of the relevancy of this proposed change in philosophy for comparison of dispersion modeling results with observations. Please participate at the Rouen Harmonization Symposium in October 1999. I hope to see you there.

Best Regards,

John S. Irwin
email: irwin.john@epa.gov

References:

Cox, W.M. and Tikvart, J.A., "A Statistical Procedure for Determining the Best Performing Air Quality Simulation Model," Atmospheric Environment, Vol. 24A, 1990, pp. 2387-2395.

Meier, P., "Variance of a weighted mean." Biometrics, Vol. 9, 1953, pp. 59-79.

Olesen, H.R., "Toward the establishment of a common framework for model evaluation," Air Pollution Modeling and its Application XI, Edited by S.E. Gryning and F.A. Scheirmeier, Plenum Press, N.Y., 1996, pp. 519-528.