Int. J. Environment and Pollution, Vol. X, Nos. X-X, 1999

Statistical evaluation of centerline concentration estimates by atmospheric dispersion models

John S. Irwin*
Atmospheric Sciences Modeling Division, Air Resources Laboratory,
National Oceanic and Atmospheric Administration, Research Triangle Park, NC 27711, USA

Abstract: Within the American Society for Testing and Materials (ASTM) a Standard Practice (Z6849Z [1]) is being developed to provide an objective statistical procedure for comparing air quality simulation modeling results with tracer field data. The practice is limited to steady-state local-scale transport from isolated point sources in simple terrain. Evaluation data having similar external conditions are grouped together, and comparisons are made of the model's ability to replicate without bias the average of the centerline maximum concentrations for each group. Centerline concentrations measured during three field studies are compared with estimates from three steady-state plume models (ISC, HPDM, and AERMOD). These results combined with those presented in Irwin and Rosu (1998) provide a complete examination of the draft ASTM Standard Practice under development. It is concluded that the evaluation methodology is capable of objectively discerning differences in skill between models in their ability to estimate the centerline maximum concentration.

Keywords: air dispersion modeling, statistical evaluation, model performance

Reference to this paper should be made as follows: Irwin, J.S. (1999) 'Statistical evaluation of atmospheric dispersion models', *Int. J. Environment and Pollution*, Vol. X, No. X, pp. 000-000.

1 Introduction

Statistical evaluation of model performance is viewed as part of a larger process that collectively is referred to as model evaluation. Two major considerations in developing the statistical comparison measures were that deterministic steady-state dispersion models provide estimates of the average concentration for the specified meteorological conditions, and that the large 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. The goal of the practice is to:

(1) determine which model's results are closest to the observed average result, and to

(2) determine which model's results are significantly different from the results of the model identified in (1) using an objective statistical significance test.

To compare simulation results with an observed average result, the practice begins by stratifying the experimental observations 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 similar 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). For each regime, we can compare the average of a model's estimates with the average derived from the group of observations. From a summary of these results across all the regimes, we determine (1) as the model with the smallest combined value of the average absolute fractional bias and variance. From a statistical test of the differences seen in the summary (absolute fractional bias and variance) across all regimes, we determine (2).

To illustrate how the evaluation methodology would work, the draft Practice describes how comparisons could be made of model performance in estimating the average maximum centerline concentration. In the discussion that follows, we investigate the results obtained when the suggested evaluation methodology is implemented.

* 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 author upon request.

Irwin and Rosu (1998) tested various aspects of the draft practice up to and including the manner in which the experimental data should be processed in order to select receptors suitably close to the observed center of mass such that one could assume the observed concentration is representative of the centerline concentration. It was determined that a robust way to combine observed concentrations along arcs within a given regime was to use the computed center of mass from each arc as a common reference point. They found that by expressing the receptor positions relative to the center of mass seen for each experiment, the results from all the experiments within the regime could be combined. Once grouped in this manner, a lateral dispersion can be computed for all the results in the regime. This regime lateral dispersion, σ_y, can then be used to define for each experiment which receptors are close (within $\pm 0.67\sigma_y$) to the center of mass. The statistical properties derived from concentration values from these receptors are considered representative of centerline concentration values.

Bootstrap resampling is used to compute the average statistical properties of the observed and simulated centerline concentration values (averages and standard deviation of the averages). Resampling is also used to compute the statistical properties of the differences to be seen between the simulated and observed average centerline concentration for each regime. From a numerical experiment, Irwin and Rosu (1998) concluded that bootstrap sample sizes ranging from 70 to 400 should be sufficient in developing stable estimates of the standard deviations of the averages. In developing the bootstrap samples, two replicate sampling methods were tested by Irwin and Rosu (1998): sample of one versus sample of a pair. This was done to assess whether observations from adjacent receptors could be treated as independent. Slight yet significant differences were seen in the variances computed, and it was concluded that further testing was needed to assess the effect of these differences when using these sampling techniques within the context of the draft American Society for Testing and Materials (ASTM) practice to assess model performance.

Irwin and Rosu (1998) left for further study several features of the draft practice: one, the summarization across all regimes such that one can conclude which of several models most consistently simulates the observed average result; two, the summarization to determine if the difference seen between models is significant, and three, the effect on conclusions reached in model performance of the differences seen in the resampling results when individual samples are taken versus a pair.

2 Tracer field data and defining regimes

2.1 Near-surface nonbuoyant releases

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. Non-buoyant near-surface releases of tracer typically are well characterized using surface layer similarity scaling. To group these data into regimes, 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 seven stability groups, which for each we have sampling results for five downwind arcs. This provides us with 35 regimes with results from approximately 11 experiments in each regime. The eleven most stable experiments were divided into two regimes, with seven experiments in one group and four experiments in a separate group. Irwin and Rosu (1998) determined that four experiments were dispersing in a 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 (Briggs, 1982).

2.2 Elevated buoyant releases

The sulfur-hexafluoride, SF_6, tracer experiments conducted at Kincaid (Bowne et al., 1983), involved a release from a 183-m stack with a buoyant plume rise on the order of 200 m. There were 171 experiments conducted during April, May and August of 1980, and May and June of 1981, with near-surface hourly concentrations with reasonably complete meteorology. There were twelve roughly-defined receptor arcs ranging from 0.5-km to 50-km from the release, but not all arcs were activated for each experiment. We divided the data into four stability ranges defined in terms of Zi/L (Zi is the modeled mixing height). This provided 29 regimes where centerline concentration values could be compared with modeling results.

The SF_6 tracer experiments conducted at Indianapolis (Murray and Bowne, 1988) involved a release from an 84-m stack with a buoyant plume rise. There were 170 experiments conducted during September and October of 1985, with near-surface hourly concentrations, with reasonably complete meteorology. There were twelve roughly-defined arcs ranging from 0.2-km to 12-km from the release, but not all arcs were activated for each experiment. We divided the data using Zi/L into four daytime stability ranges and one nighttime stability range. This provided 36 regimes where centerline concentration values could be compared with modeling results.

3 Generation of modeling results

Three plume dispersion models were run on the three field studies: ISC (U.S. EPA, 1979), HPDM (Hanna and Paine, 1989), and AERMOD (Cimorelli et al., 1996). ISC is a commonly used multiple source model that employs the Pasquill dispersion curves. It was developed in the late 1970's. HPDM was developed in the late 1980's and is one of several models designed in the 1980s that employed convective boundary layer parameterizations of dispersion. AERMOD was developed over the last five years and attempts to summarize current state-of-practice boundary-layer parameterizations of dispersion. The meteorological processor for AERMOD was employed with available on-site observations, hourly National Weather Service (NWS) weather observations, and twice-daily NWS upper air observations, to characterize the meteorological conditions for each of the three tracer dispersion sites (PES, 1998). To provide a common set of meteorology, these data were then converted to formats suitable for use by each of the three models. The Pasquill stability category for use by ISC was derived from the estimated Obukhov length and site roughness length using relationships discussed by Golder (1972). For each tracer release, the models were run so that the simulated centerline concentration was obtained for all possible downwind arcs for each field study. We then used these simulated centerline concentration values, C, divided by the emission rate, Q, for each experiment, for comparison with the average of the observed C/Q values selected as being representative of centerline concentration values for a particular regime.

4 Analysis of results

Following the draft ASTM procedures for each regime, an average and a standard deviation of this average was computed from the C/Q values observed and estimated by each model. For each regime, 500 bootstrap samples were generated, where the samples were developed using sampling by pairs. Figure 1 illustrates the regime averages obtained for each of the three studies as scatter plots and Quantile-Quantile (Q-Q) plots. In all these analyses, we are comparing average observed and estimated centerline concentrations from each regime as determined by the resampling analysis, so differences seen result from differences in the simulated and actual dispersion and plume rise (rather than differences in the direction of transport). The Q-Q plots are simple pairings of predicted concentrations; ranked highest to lowest, with observed concentrations, ranked in the same manner. The dotted lines shown in the figures depict a factor of two difference. The scatter plots provide a visual means for comparing values paired in time and space, but do not provide a means for assessing the ability of the models to simulate the frequency distribution of values. The Q-Q plots provide a means for assessing the differences to be seen between the observed and simulated frequency distributions, but do not provide a means for assessing whether the maxima being simulated are for similar meteorological conditions as when observed.

In Figure 1, it is evident that ISC and AERMOD best characterize the near-surface releases at Prairie Grass. It appears that HPDM tends to underestimate the concentration values on average. The results for Indianapolis and Kincaid exhibit considerable scatter, with large differences with observations evident in both the scatter plots and the Q-Q plots.

Figure 2 illustrates regime averages obtained for the Kincaid SF_6 comparisons for four stability ranges as a function of downwind distance. Results for the most unstable conditions are shown in Figure 2(A) and conditions progressively approach conditions that are more neutral as we proceed from Figure 2(A) to Figure 2(D). The error bars depict the bootstrap estimate of the standard deviation of the computed average, multiplied by two. Where the error bars overlap, the differences seen would test to be not significant at the 95% confidence limit.

In Figure 2(A), we see that the ISC model tends to overestimate the centerline concentration values for all distances downwind. The ISC model tends to underestimate the average observed centerline concentration values shown in Figures 2(B) through 2(D). It is only when we reach the most neutral results shown in Figure 2(D), that ISC begins to be in accord with the observations, and then only for transport distances beyond 7 km. The error bars for the HPDM results overlap the results shown for the observations until conditions become more neutral. As we approach neutral conditions (Figures 2(C) and 2(D)), HPDM tends to overestimate the centerline concentration values for transport distances of less than 5 km. There is also a tendency to underestimate concentrations for the nighttime regimes at Indianapolis by HPDM, see solid square symbols in Figures 1(C). The results for AERMOD generally are in accord with the observations except for the most unstable cases shown in Figures 2(A) and 2(B), where for transport distances of order 4 to 6 km, AERMOD tends to underestimate the observed average centerline concentration values.

5 Overall performance comparisons

We use a pooled average absolute fractional bias and its standard deviation computed across all regimes using inverse variance weighting, for summarizing model performance across all regimes for a given experiment. This avoids being misled by offsetting bias between regimes, where an overestimation in one regime might be offset by an underestimation in another regime. Pooling the results together using inverse variance weighting devalues regimes in which the computed standard deviation of the absolute fractional bias is large in comparison to that computed in other regimes.

To assess which model compares best with observations, we form a model comparison measure, MCM = Avg(AFB) + 2 Std(AFB), where Avg(AFB) is the computed pooled absolute fractional bias over all regimes, and Std(AFB) is the computed pooled standard deviation of the absolute fractional bias over all regimes. The model with the smallest value of MCM is deemed the model that on average compares best with observations. The differences between the Avg(AFB) values computed for each model are tested using a null hypothesis that the differences seen in the computed values for the Avg(AFB) are not statistically significant. Here we assume the values of Avg(AFB) possibly have different standard deviations, thus the t-test is:

$$t = \frac{\left| Avg(AFB)_i - Avg(AFB)_{base} \right|}{\sqrt{Std(AFB)_i^2 + Std(AFB)_{base}^2}}$$

where *base* refers to the statistics for the model having the lowest value for the MCM (base) and *i* refer to the model being compared with the base case model. If t is greater than a Student-t value of 1.96 (assuming degrees freedom are greater than 30), we reject the null hypothesis that the two means are the same with 95% confidence (for further discussion on hypothesis testing concerning two means see a standard text as Miller and Freund, 1965).

Table 1 summarizes the results obtained for the three models tested on these three tracer dispersion studies. Here we have excluded the suspect cases noted above for Indianapolis. It is seen that AERMOD has the lowest value of the MCM for Prairie Grass and Kincaid. HPDM has the lowest MCM for Indianapolis. The differences between AERMOD and HPDM for Kincaid are seen to be not statistically significant. The average performance by the AERMOD model is roughly the same over all three field studies. The performance by the HPDM model is roughly the same for the Kincaid and Indianapolis field studies, and is seen to have the largest differences with observations for the low-level releases for Project Prairie Grass. The ISC model has large differences with observations for all three field studies. For Indianapolis, AERMOD's performance is essentially unchanged when the nighttime regimes are included. HPDM's performance is degraded significantly for Indianapolis, if the nighttime regimes are included.

When onsite Zi and L values were used to sort the Kincaid experiments into regimes (instead of using results from AERMOD's meteorological preprocessor), variations in the statistical results for the MCM of order 35% were seen, but these variations did not alter the conclusions reached by the summary statistics.

The above results were based on sampling by pairs to develop the bootstrap samples. When individual samples were drawn, the main effect was to reduce the computed variances for the averages and differences between model and observed averages. These smaller variances meant the differences seen were more likely to be deemed significant. Positive correlation between adjacent concentration observations is likely, and such correlation would cause larger variances to be computed to the regime averages. It is concluded that samples should be drawn by pairs, otherwise differences will be deemed significant when if fact they likely are not significant.

6 Conclusions

The results presented here combined with those presented in Irwin and Rosu (1998) provide a complete examination of the draft ASTM Standard Practice under development. These results suggest that the evaluation methodology is capable of objectively discerning differences in skill between models in their ability to estimate the centerline maximum concentration at the surface downwind from a point source release. Next steps includes revising the practice to be consistent with the findings presented here and in Irwin and Rosu (1998), and developing a numerical algorithm that can generate pseudo modeling results with known statistical properties. This last step will allow direct investigation of how discerning the developed evaluation procedures are, and could provide a means for testing whether proposed future changes to the methodology are significant and therefore should be adopted. Two concluding remarks are warranted. The first is on the grouping of data for analysis, and the second is on how the modeling results were created for the analyses discussed in this paper.

The grouping of the data is a valuable and important feature of the draft Practice, and yet it is inherently subjective and will likely cause concern. Stratifying the data into groups is a standard statistical technique to provide greater discernment in statistical significance tests (Miller and Freund, 1964). The grouping of the data allows us to compute an average characteristic (such as the centerline maximum concentration), which is the proper statistic to compare with what the steady-state model is attempting to simulate. By making the basic statistical comparisons within each regime, we ensure that we are comparing model estimates and observations that have ostensibly similar meteorological conditions. Since we know how the models perform in each regime, it becomes obvious under what conditions the dispersion characterizations need improvement. By employing bootstrap resampling within each regime, we provide summary statistics for each regime. These summary statistics provide a means for performing an objective test of whether the differences seen between models in one or more of the regimes are statistically significant. Preliminary testing of alternative grouping criteria suggests that the relative ranking of performance between models remains the same. As promising as these preliminary results are, there may always be concern raised regarding the criteria used in selecting and grouping the data for analysis. For instance, the current methodology assumes that average characteristics of data grouped together are representative of a steady-state meteorological condition. If conditions were not steady-state over the sampling time of one or more of the experiments, should these experiments be used? It is currently believed that as the methodology receives broader use, experience will provide guidance on acceptable practices in grouping data for analysis.

With respect to the data comparison results shown, it is important to note that meteorological processors developed for ISC and HPDM were not used in developing the model input meteorology. Hence, the results shown here can not be interpreted as being what would be expected if one were to operate ISC or HPDM with their respective meteorological processors. Furthermore, HPDM was designed for elevated releases, not surface releases, and thus the performance seen for the Prairie Grass experiments was not unexpected. The focus of this analysis was on whether the draft evaluation methodology was capable of discerning differences in skill, not on defining the actual relative performance of the three models if they were used as designed.

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.

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.

Cimorelli, A.J., S.G. Perry, R.F. Lee, R.J. Paine, A. Venkatram, J.C. Weil, and R.B. Wilson, (1996) *Current progress in the AERMIC model develop program*, Preprints, 89th Annual Meeting and Exhibition of the Air and Waste Management Association, Air and Waste Management Association, Nashville, TN, June 24-28, 1996, Preprint No. 96-TP24B.04, 27 pp.

Golder, D.G., (1972) A comparison of stability parameters, *Boundary-Layer Meteorology*. (3):47-58.

Hanna, S.R., and R.J. Paine, (1989) Hybrid Plume Dispersion Model (HPDM) development and evaluation, *Journal of Applied Meteorology*, (28):206-224.

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.

Irwin, J.S., and M-R. Rosu, (1998) *Comments on a draft practice for statistical evaluation of atmospheric dispersion models,* Proceedings of the 10th Joint Conference on the Application of Air Pollution Meteorology with the A&WMA. Phoenix, AZ., pp. 6-10.

Miller, I., and J.E. Freund, (1964) *Probability and Statistics for Engineers*, Prentice-Hall Inc., Englewood Cliffs, NJ, pp. 165ff.

Murray, D.R., and N.E. Bowne, (1988) *Urban Power Plant Plume Studies*, EPRI Report No. EA-5468, Research Project 2736-1, Electric Power Research Institute, Palo Alto, CA.

Pacific Environmental Services, (1998) *The development of centerline concentrations from AERMOD, ISCST3, and HPDM for the Prairie Grass, Indianapolis, and Kincaid Field Experiments*, EPA Contract No. 68D70069, Work Assignment No. I-04. Technical Memorandum February 4, 1998, 11pp.

U.S. Environmental Protection Agency, (1979) *Industrial Source Complex (ISC) Dispersion Model User's Guide*, EPA-450/4-79-030 (Volume I) and EPA-450/4-79-031 (Volume II), U.S. Environmental Protection Agency, Research Triangle Park, NC, pp. 360 (Volume I) and pp. 452 (Volume II).

Table 1. Summary statistics for model performance. N is the number of regimes, and the Student-t compares the difference between the Avg(AFB) value with the model having the smallest value of MCM, 'Base'. A value of 1.96 for the Student-t statistic indicates the difference is significant with 95% confidence.

		N	Avg(AFB)	Std(AFB)	MCM	t-test value
Prairie	AERMOD	35	0.40	0.010	0.42	Base
Grass	ISC	35	0.94	0.009	0.96	39.77
	HPDM	35	1.86	0.002	1.86	137.14
Kincaid	AERMOD	29	0.35	0.025	0.40	Base
	ISC	27*	1.99	0.001	1.99	65.19
	HPDM	29	0.38	0.023	0.42	0.79
Indianapolis	AERMOD	30	0.36	0.031	0.42	2.06
	ISC	30	0.87	0.021	0.91	18.12
	HPDM	30	0.27	0.025	0.32	Base

* In two regimes ISC's predictions were essentially zero, when observed concentrations were otherwise. In such cases, the computed variance of the AFB is zero, which precludes including results in the pooled Avg(AFB) as this requires use of the inverse of the variance.

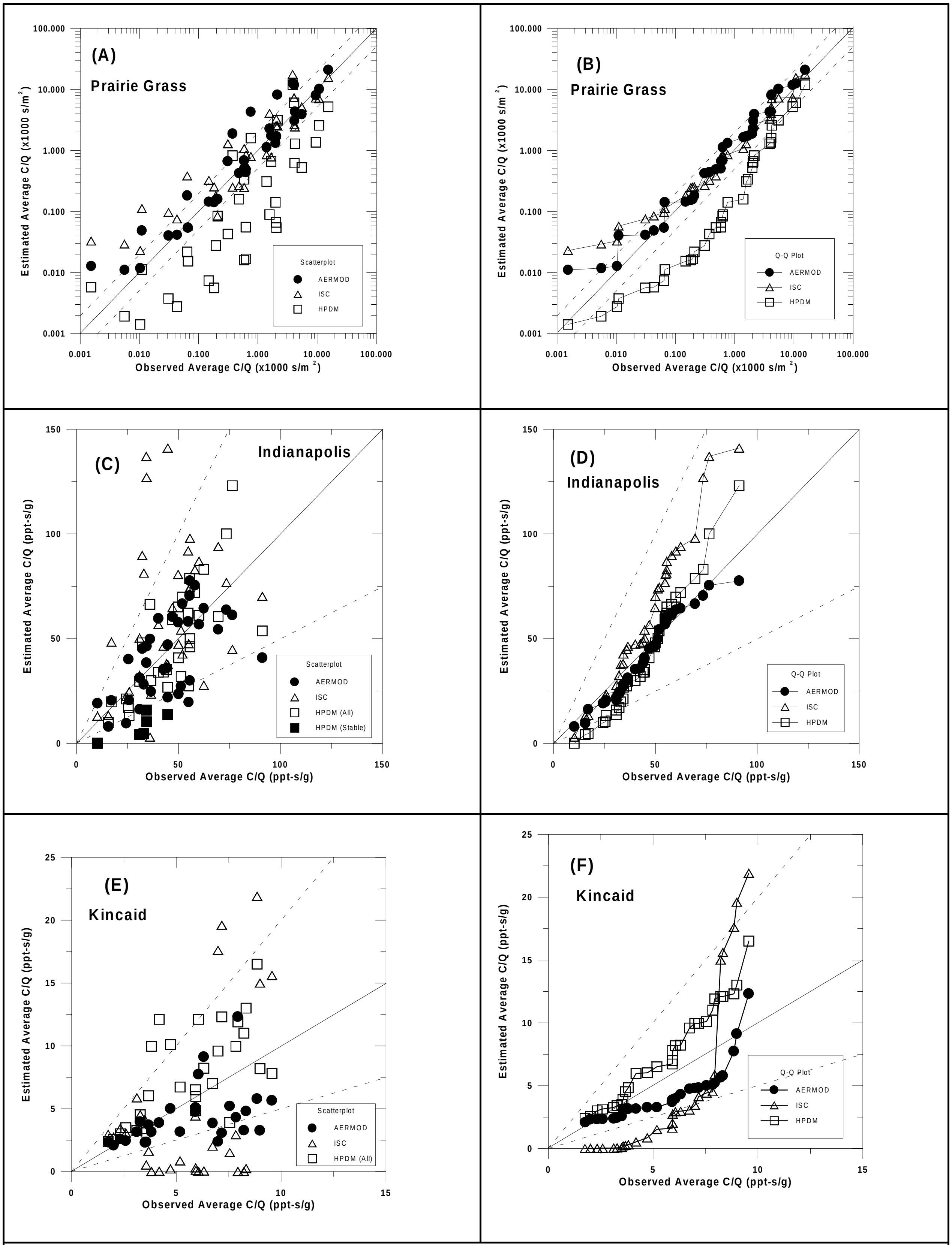

Figure 1. Scatter and Quantile-Quantile plots, comparing modeling results from three models on tracer dispersion data collected at three field studies. Dotted lines indicate a factor of two over or under prediction.

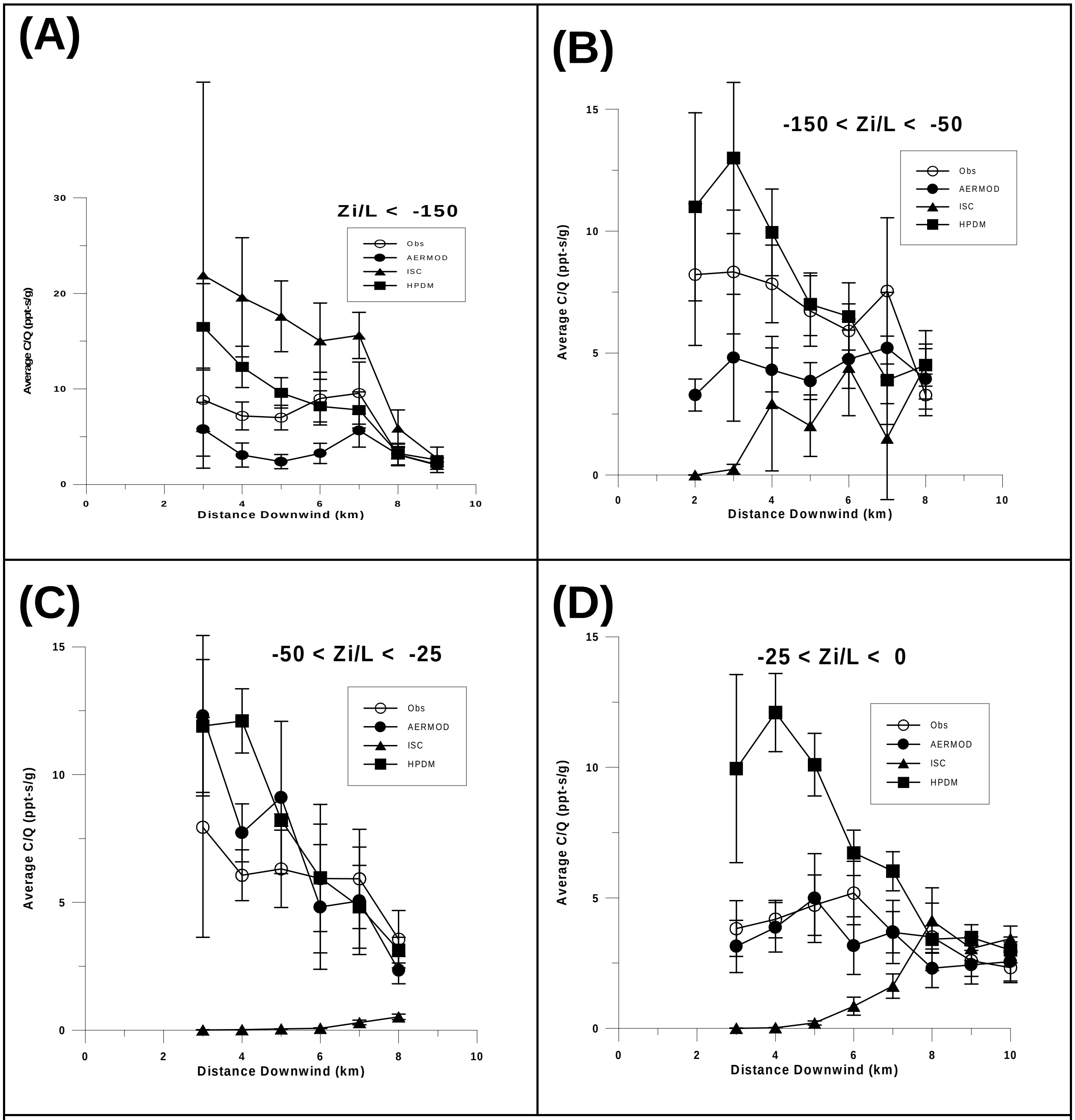

Figure 2. Comparison of modeled and observed average C/Q values as a function of downwind distance for four stability ranges defined in terms of Zi/L for Kincaid tracer results. Modeling results are shown for three models.