7.3

USE OF URBAN 2000 FIELD DATA TO DETERMINE WHETHER THERE ARE SIGNIFICANT DIFFERENCES BETWEEN THE PERFORMANCE MEASURES OF SEVERAL URBAN DISPERSION MODELS

Steven Hanna and Patricia Fabian, Harvard School of Public Health, Boston, MA

Joseph Chang, George Mason University, Fairfax, VA

Akula Venkatram, University of California, Riverside, CA

Rex Britter and Marina Neophytou, University of Cambridge, UK

David Brook, Defense Science and Technology Laboratory, Porton Down, UK

1. INTRODUCTION

Dispersion models are often evaluated by comparison with field data. The Hanna et al. (1993) BOOT model evaluation software is widely used to evaluate the performance of individual models. The software also allows the differences in the statistical performance measures between two models to be assessed to determine whether there are significant differences at some confidence level (usually 95 %). The methodology is demonstrated using six urban dispersion models applied to the Salt Lake City Urban 2000 field data (Allwine et al, 2002). As a preliminary exercise, the comparisons focus on the maximum predicted and observed concentrations on the seven monitoring arcs during each of the 18 field experiments.

A statistical test such as that in BOOT is needed to compare the performance measures for two models because of the possibility that the two models' predictions may be correlated. For example, if the concentrations predicted by model B are always exactly two times the concentrations predicted by model A, then there will always be a significant difference between the performance measures calculated for the two models. Otherwise, it would be possible to look solely at whether the 95 % confidence ranges on the performance measures for Models A and B overlap.

The urban dispersion models included so far are:

Urban Hazard Prediction Assessment Capability (HPAC) (with Urban Dispersion Model, UDM) (DTRA, 2001 and 2004)

Urban baseline dispersion model (Hanna et al., 2003)

UDM (Hall et al, 2002)

Corresponding author: Steven Hanna, 7 Crescent Ave., Kennebunkport, ME 04046 207 967 4478 (fax: 5696) E-mail: hannaconsult@adelphia.net

AERMOD urban (EPA, Cimorelli et al., 2004)

Barrio Logan Model (BLM) developed by Venkatram et al. (2004)

Simple Urban Dispersion Correlation (SUDC) developed by Neophytou and Britter (2004)

2. STATISTICAL METHODS

The following equations define the statistical performance measures that are most often used in the BOOT evaluation software (Hanna et al., 1993, Chang and Hanna, 2004). These performance measures include the fractional bias (FB), the geometric mean bias (MG), the normalized mean square error (NMSE), the geometric variance (VG), and the fraction of predictions within a factor of two of observations (FAC2):

$$FB = \frac{\left(\overline{C_o} - \overline{C_p}\right)}{0.5\ \left(\overline{C_o} + \overline{C_p}\right)} \qquad (1)$$

$$MG = \exp\ \left(\overline{\ln C_o} - \overline{\ln C_p}\right) \qquad (2)$$

$$NMSE = \frac{\overline{\left(C_o - C_p\right)^2}}{\overline{C_o}\ \overline{C_p}} \qquad (3)$$

$$VG = \exp\ \left[\overline{\left(\ln C_o - \ln C_p\right)^2}\right] \qquad (4)$$

FAC2 = Fraction of predictions that are within a factor of two of observations (5)

where
C_p: model predictions of concentration,
C_o: observations of concentration,
overbar ($\overline{C}$): average over the dataset, and
σ_C: standard deviation over the dataset.

A perfect model would have MG, VG, and FAC2 = 1.0; and FB and NMSE = 0.0. Of course, because of the influence of random atmospheric processes, there is no such thing as a perfect model in air quality modeling. In addition to the standard performance measures defined above, which use data from a large number of experimental trials, the simple ratio of the overall maximum observed concentration to the overall maximum predicted concentration on each arc can be listed. These two maxima may occur during different experiment trials.

Typical magnitudes of the above performance measures and estimates of model acceptance criteria have been summarized by Chang and Hanna (2004) based on extensive experience with evaluating many models with many field data sets. It was concluded that, for comparisons of maxima concentrations on arcs, and for research-grade field experiments, "acceptable" performing models have the following typical performance measures. The fraction of predictions within a factor of two of observations is about 50% or greater (i.e., FAC2 > 0.5). The mean bias is within ± 30% of the mean (i.e., –0.3 < FB < 0.3 or 0.7 < MG < 1.3). The random scatter is about a factor of two of the mean (i.e., NMSE < 4 or VG < 1.6). However, these are not firm guidelines and it is necessary to consider all performance measures in making a decision concerning model acceptance. Since most of these criteria are based on research grade field experiments, model performance would be expected to deteriorate as the quality of the inputs decreases.

Depending on scenario, agency, and questions asked of models, there is a variety of additional performance measures that are considered. For example, the EPA is interested in maximum C at any time or place on the domain. Therefore they use Quantile-Quantile (Q-Q) plots where only the separately ranked observed and predicted C's are plotted. For emergency response, DHS, DOD and others are interested in predicting area coverage properly for hazard impact, thus using Figure of Merit and other measures of predicted area coverage. Different weights can be given to false positives and false negatives.

The statistical comparison method can be used for ANY performance measure, as long as it can be expressed quantitatively. BOOT is general and can be used for types of models other than atmospheric transport and dispersion models. Alternate performance measures can easily be added. The general statistical comparison procedure is independent of the performance measure.

To estimate confidence limits on differences in a performance measure (PM) between two or more models, start with a master table containing sets of observations and predictions by the models. Use bootstrap or jackknife resampling to resample from the master table N times. There are M rows with one observation and L model predictions. In each resample, M of the rows are randomly chosen (with replacement) and the PM is then recalculated for each model. For each resample, the difference in PM between two models (i and j) is calculated, ΔPM(i,j) . The N values of ΔPM(i,j) can be ranked and used to define 95 % confidence limits. If the 95 % confidence limits on ΔPM(i,j) overlap 0.0, then it can be concluded, with 95 % confidence, that the difference in PM between the two models is not significantly different from 0.0.

3. OVERVIEW OF SLC URBAN 2000 STUDY

This preliminary demonstration of the comparison methodology focuses on the maximum hourly-averaged concentrations on seven arcs for the 18 Urban 2000 trials in Salt Lake City. The Urban 2000 experiment is described in great detail by Allwine et al. (2002) and a general analysis of the concentration data and the meteorological data is given by Hanna et al. (2003). The observed concentrations used in this demonstration are listed in Table 1. The numbers are the maximum hourly-averaged concentrations, C, divided by the source term, Q, observed during each experiment trial on each monitoring arc. Figures 1 and 2 contain maps where the locations of the 66 individual concentration monitors are indicated and the seven monitoring arcs can be seen. Figure 1 shows the outer three arcs, at distances of about 2, 4, and 6 km from the source. Figure 2 shows the four inner arcs, at distances of 156, 394, 675, and 928 m from the source. The SF_6 source location is shown as a star on the figures, and was at a height of 1.5 m near a large downtown building. During each of the six nights of the experiment, tracer gas was released continuously in three one-hour periods, separated by one-hour period where no gas was released.

Figure 1 also contains the locations of the meteorological sites. Observed winds were relatively light, averaging about 1.5 m/s. Even though all experiments were at night, nearly-neutral conditions prevailed in the built-up downtown area due to mechanical mixing around the buildings and due to heat inputs from man's activities. Several of the urban dispersion models (e.g., Venkatram et al., 2004, Hanna et al., 2003, and Cimorelli et al., 2004) assume that stabilities over built-up urban areas are always nearly-neutral or convective.

As seen in Figure 2, the closest monitoring arc to the source was at a distance of about 156 m, or about one block from the source. Consequently, the concentration observations are at locations where the plume has grown to an extent that the turbulence and dispersion is due to the combined influence of many buildings. All six models are intended to be applicable to this scenario (i.e., downwind distances beyond a "few" buildings) and are not applicable to distances where the tracer plume would be strongly influenced by a single building or a single street canyon.

4. OVERVIEW OF SIX MODELS

As stated earlier, since this is a demonstration exercise, no attempt was made to cover the complete spectrum of urban dispersion models that are currently in use. The six models that are included represent the authors' own models plus the HPAC-Urban model, which has been the subject of evaluations by Chang et al. (2004). The sole criterion for being an author and for having one's model included was that one could provide predictions in the desired format by the time deadline for this manuscript. It is important to note that all six models are intended for distances beyond a few building heights. The models are not intended for application to flow and dispersion around the "first" building encountered. A brief overview of the models is given below:

AERMOD (Cimorelli et al., 2004) is the U.S. EPA's updated Gaussian-type plume model for application to industrial sources and other local sources at downwind distances less than about 10 or 20 km. It automatically handles urban areas through inputs of surface roughness length and building geometries. It assumes nearly-neutral or convective conditions in urban areas. For this application, it was run using observed average wind speeds in the urban area and a southeasterly wind direction. Geometry information was input only for the buildings near the source. We had many discussions with the developers about the current runs, since the model is usually applied to stack sources and not to sources near street level in an urban canopy. Clearly more sensitivity runs are needed for this type of application.

Baseline Urban Dispersion Model (Hanna et al., 2003) is also a Gaussian-type model, using observed average wind speeds in the urban area and assuming nearly-neutral to convective stabilities at all times. Turbulence is parameterized using urban boundary layer relations developed by the authors. A key assumption is that, even at low wind speeds, the lateral turbulent velocity does not drop below 0.5 m/s. The latter assumption results in wide plume spread during light winds.

BLM (Barrio Logan Model) (Venkatram et al., 2004) is a Gaussian-type model based on simple parameterizations of the urban boundary layer and dispersion in the urban area. It assumes an initial size to the plume due to mixing behind buildings near the source. It assumes that observed wind speeds and turbulence are available from some height above the average building height. For the current exercise, BLM uses winds from the M08 anemometer (see Figure 1), at a height of 23 m on top of a warehouse building.

HPAC (Hazard Prediction Assessment Capability) (DTRA, 2001, 2004) is a multipurpose code that could be described as a Gaussian puff model. It is widely used in U.S. Department of Defense (DOD) applications and has been previously evaluated for several non-urban field experiments (Chang and Hanna, 2004). Two of the authors (Hanna and Chang) were involved in an extensive evaluation of HPAC with the Urban 2000 data base (Chang et al., 2004). At that time, the urban HPAC that was evaluated was not officially released. However, the latest Version 4.04.011 (DTRA, 2004) automatically includes the urban algorithms. HPAC incorporates UDM (Hall et al. 2002) for $x < 2$ km in Salt Lake City, after which it switches to its standard puff dispersion algorithm (SCIPUFF). For the current exercise, the wind inputs are from the Raging Waters (RGW) site, denoted as N01 in Figure 1. This option, which uses winds from a site about 5 km upwind of the urban area, was shown by Chang et al., 2004) to produce optimum performance.

SUDC (Simple Urban Dispersion Correlation) (Neophytou and Britter, 2004) is a one-line urban model formulation developed by the authors using observations from several urban field and laboratory experiments. The relation, which states that C/Q is proportional to x^{-2}, has been developed for the range x/H_b less than about 60, where H_b is average building height. For Urban 2000, where H_b is about 15 m (Hanna et al., 2003), this would imply that SUDC is recommended only for the closest four monitoring arcs, at x less than about 1000 m. Nevertheless, for the statistical demonstration exercise in this paper, we also include the SUDC predictions at arcs 5 through 7 (x = 2, 4, and 6 km). Future studies should focus only on the nearest four arcs.

UDM (Urban Dispersion Model) (Hall et al., 2002) is a widely-used model developed by the UK Defence Science and Technology Laboratory (DSTL) based on assumptions of a Gaussian shape and assumptions concerning empirical parameterizations developed from special field and laboratory experiments involving obstacle arrays. UDM is also included in the Urban HPAC model discussed earlier. However, it was also tested as a "stand-alone" model using its original software and applied to all seven monitoring arcs. These UDM runs were made by David Brook and originally included several options for wind inputs. The UDM outputs evaluated here are using the so-called "all winds" option, where all wind monitoring sites in Figure 1 are included. UDM predictions varied somewhat depending on which wind option was used, as expected.

5. RESULTS

As recommended by Hanna et al. (1993), it is useful to first "look" at the data before applying the statistical performance measures. Figure 3 contains scatter plots of the observed versus model-predicted C/Q values for each of the six models. The points represent maximum hourly-averaged C/Q for each of the 18 trials and seven monitoring arcs. Because higher concentrations occur on the closest arcs, the points in the upper right corner of each plot represent the closest arcs. Points below the line of perfect agreement represent underpredictions. It is seen in Figure 3 that the six models do a fairly good job of matching the observed C/Q values, although there are some indications of slight underpredictions or overpredictions for some models and some arc distances. Also, the amount of scatter is similar in the six plots. Some of the biases seen on the plots are: The Baseline model underpredicts at the higher values; HPAC slightly overpredicts at most distances; SUDC has little bias except at the lower values (i.e., the most distant arcs), where it is stated to not be applicable; UDM does better at the higher values (i.e., the closest arcs) and underpredicts at the lower values; AERMOD tends to underpredict by a factor averaging 2 or 3 at all values; and BLM has little bias.

Continuing with the first step where the results are "looked at", Figure 4 contains Quantile-Quantile plots in the same format as Figure 3. The difference is that, in Figure 4, the observations and predictions are separately ordered from lowest to highest C/Q, and then the points that are plotted represent the N^{th} rank of the observations and predictions. Again, the models seem to have approximately equivalent performance, although the BLM model may be closest to the "perfect agreement" line. The general tendency towards slight underpredictions of UDM and AERMOD can be clearly seen, as well as the slight overprediction tendency of HPAC.

The quantitative performance measures defined in Section 2 are listed in Table 2 for the six urban dispersion models. It is seen that there is a variation in which models perform better for certain performance measures. Also, in many cases, the performance measures for two or more models appear to be close (e.g., VG is 1.55 for Baseline and 1.70 for BLM). The main purpose of the current paper is to determine whether these values are not significantly different from zero.

Figure 5 is a plot of MG versus VG for the individual models. This plot is suggested by Chang and Hanna (2004) as a way to quickly compare model performance with a single diagram. MG is a measure of the relative mean bias and VG is a measure of the relative scatter. A perfect model has MG = VG = 1.0. Therefore, the closer a model is to the bottom and middle of the plot, the better its performance. Figure 5 suggests that the BLM and Baseline models are closest to the point of optimum performance. However, since MG and VG are based on logarithms of C_o/C_p, equal weighting is given to high and low concentrations. Therefore the underprediction by the Baseline model at high C/Q does not receive any different weight than an equal underprediction at very low concentrations.

Tables 3 and 4 contain the result of the main focus of this paper. An "X" is entered if the values of MG (in Table 3) or VG (in Table 4) for the first model are significantly different from the values for the second model, at the 95% confidence level. Most of the model pairs do show a significant difference in the tables. The model pairs that do not show a significant difference could be guessed by looking at Table 2 and Figure 5. For MG, only UDM (with MG = 3.15) and AERMOD (with MG = 3.25) do not show a significant difference. For VG, there are three pairs of models that do not show a significant difference – UDM (17.6) and AERMOD (9.73), HPAC (2.94) and SUDC (2.40), and Baseline (1.55) and BLM (1.70).

With M = 18 trials times 7 arcs per trial = 126 values of pairs of predicted and observed max C/Q, M is large enough that the differences between models are usually significant at the 95 % confidence level for the performance measures MG and VG. It is easier to show significant differences if the numbers of data points are large (i.e., the central limit theorem approximately applies).

6. CONCLUSIONS

This paper has outlined an approach for determining whether there are significant differences in the performance of urban dispersion models. Six models have been included in the demonstration exercise, which has made use of tracer data from the Salt Lake City Urban 2000 field study. It is shown how the methodology can identify pairs of models whose quantitative performance measures are or are not significantly different from 0.0 at the 95 % confidence level.

In the future, there are plans to include additional urban dispersion models, additional outputs such as all data paired in time and space, additional performance measures (e.g., MG expressed as separate false negative and false positive components), and additional urban field data sets. The methodology can also be applied to CFD models being used to predict flow and dispersion around individual buildings and street canyons.

Finally, it is clear from this preliminary model evaluation exercise that all six urban dispersion models used somewhat different options for wind inputs, and that all these wind inputs can be considered "on-site" data for the Urban 2000 field data. As a result, it is important to better understand the sensitivity of these urban models to different wind inputs, and for model developers to provide clear recommendations of appropriate wind inputs.

Acknowledgements: Drs. Hanna and Chang appreciate the support of the Defense Threat Reduction Agency (DTRA), with Richard Fry as program manager. In addition, Dr. Hanna acknowledges the support of the Department of Homeland Security (DHS) through a subcontract from Los Alamos National Laboratory, where Dr. Michael Brown is the technical program manager and has made many suggestions for improvements.

References

Allwine, K.J., J.H. Shinn, G.E. Strait, K.L. Clawson and M. Brown, 2002: Overview of Urban 2000. *Bull. Am. Meteorol. Soc*. 83, 521-536.

Chang, J.C., S.R. Hanna, Z. Boybeyi and P. Franzese, 2004: Use of Salt Lake City Urban 2000 field data to evaluate the Urban HPAC dispersion model. To appear in *J. Appl. Meteorol.*

Chang, J.C., Hanna, S.R., 2004: Air quality model performance evaluation. To appear in *Meteorol. Atmos. Phys.*

Cimorelli, A.J. et al., 2004: AERMOD: A dispersion model for industrial source applications. Part I: Model formulation and boundary layer characterization. To appear in *J. Appl. Meteorol*.

DTRA, 2001: The HPAC User's Guide, Version 4.0. report HPAC-UGUIDE-02-U-RAC0, Prepared for DTRA by SAIC, 10260 Campus Point Drive, San Diego, CA 92121.

DTRA, 2004: HPAC Version 4.04.011 (DVD containing model and accompanying data and document files). Available from DTRA, 6801 Telegraph Rd., Alexandria, VA.

Hall, D.J., A.M. Spanton, I.H. Griffiths, M. Hargrave and S. Walker, 2002: The Urban Dispersion Model (UDM): Version 2.2 Technical Documentation. DSTL/TR04774, Porton Down, Salisbury SP40JQ UK.

Hanna, S.R., R. Britter and P. Franzese, 2003: A baseline urban dispersion model evaluated with Salt Lake City and Los Angeles tracer data. *Atmos. Environ*. 37, 5069-5082.

Hanna, S.R., Chang, J.C., Strimaitis, D.G., 1993: Hazardous gas model evaluation with field observations. *Atmos. Environ*. 27A, 2265-2285.

Neophytou, M.K. and R.E. Britter, 2004: A simple correlation for pollution dispersion prediction in urban areas. DAPPLE Cambridge Note 1, January 2004, http://www.dapple.org.uk

Venkatram, A., V. Isakov, D. Pankratz, J. Heumann and J. Yuan, 2004: The analysis of data from an urban dispersion experiment. *Atmos. Environ*. 38, 3647-3659.

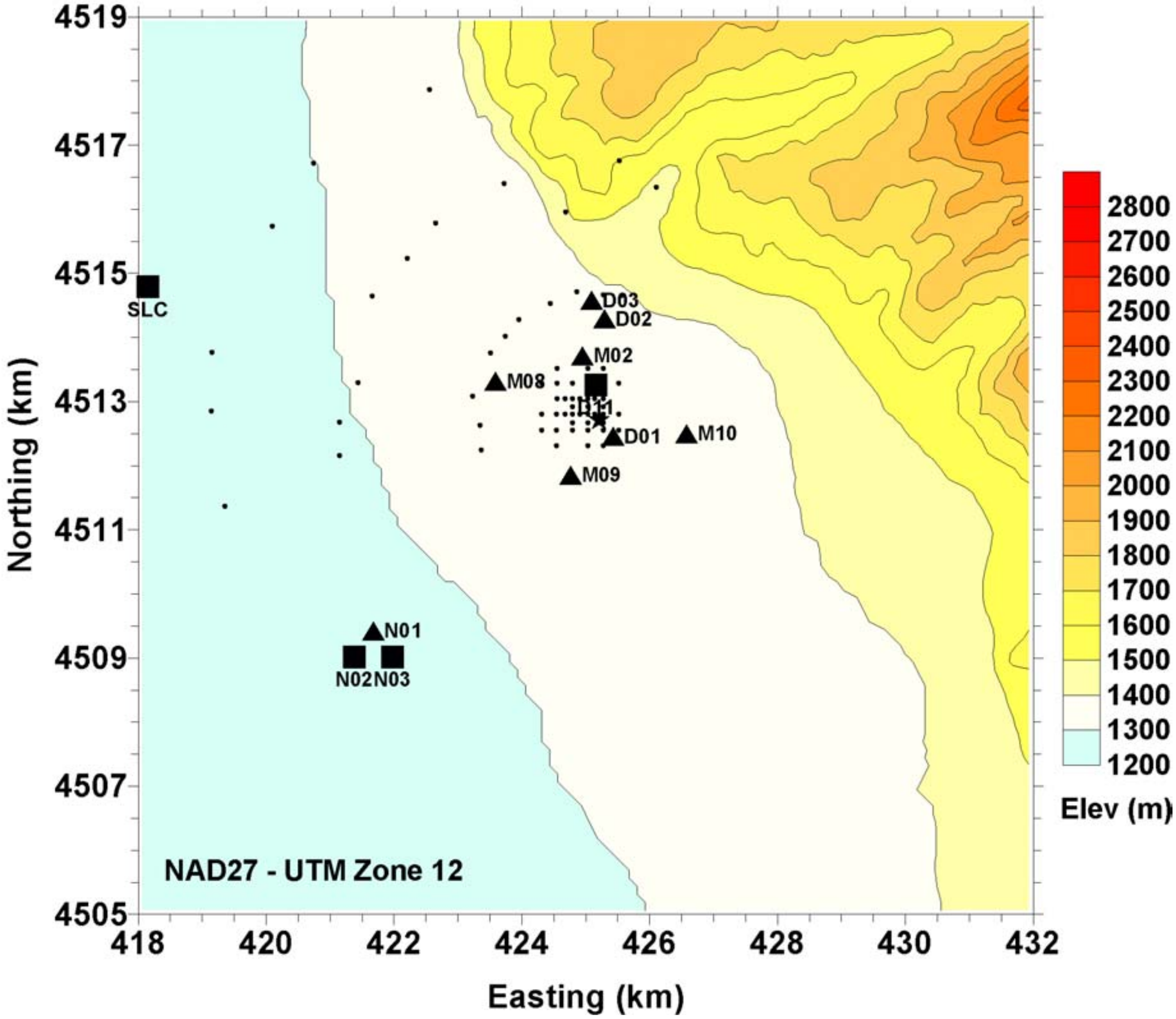

Figure 1. Map of the Salt Lake City Urban 2000 domain, showing the release location (star), terrain elevations (m), and locations of tracer samplers (small dots) and meteorological measurement sites (triangles indicate surface sites; and squares indicate vertical profile sites, where D11 and N02 are sodar sites, N03 is a profiler site, and SLC is a radiosonde site).Map of Salt Lake City Urban 2000 domain. The monitoring arcs at distances of 2, 4, and 6 km can be clearly seen to the northwest of the release location.

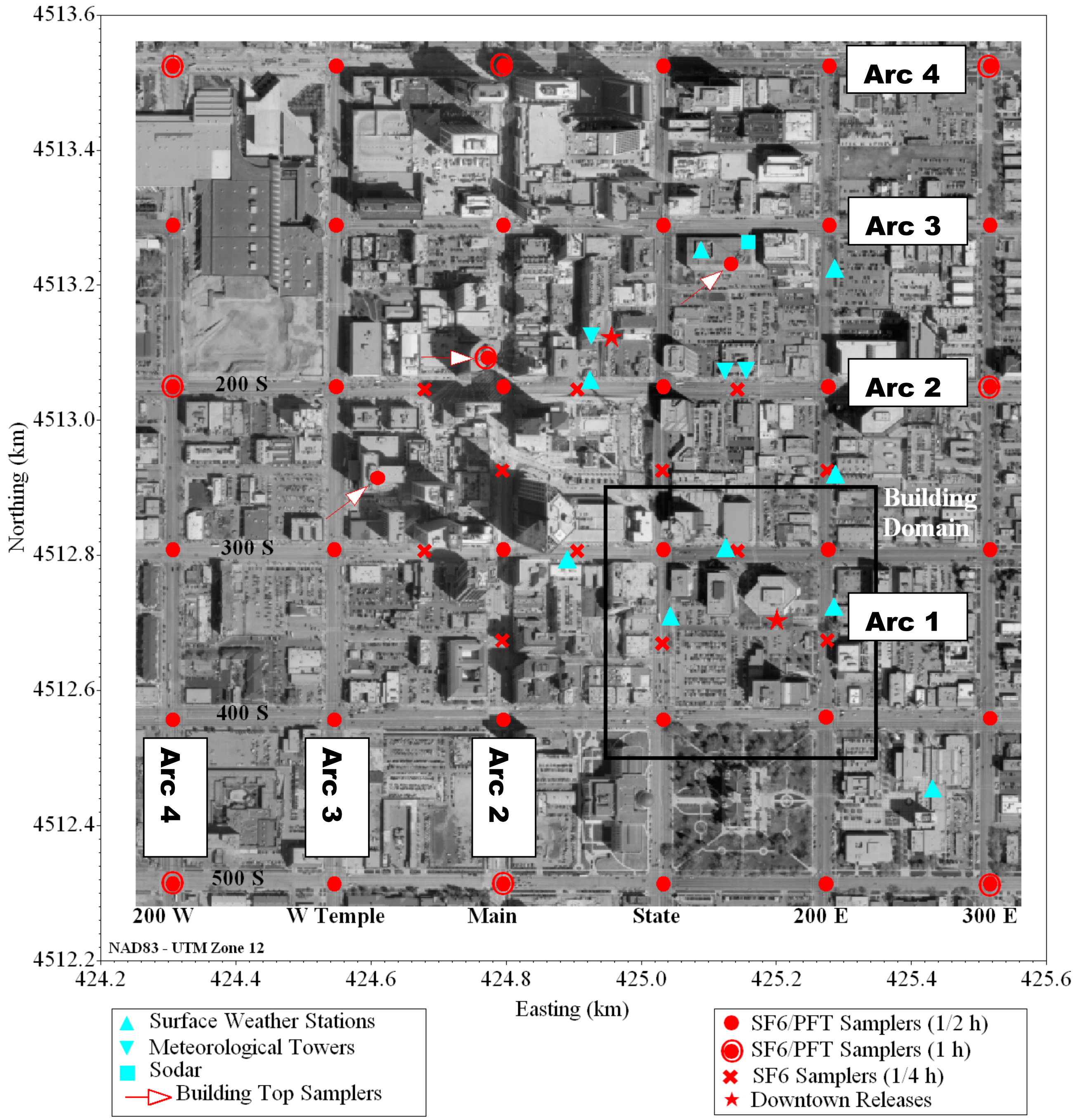

Figure 2. Map of downtown region of Salt Lake City Urban 2000 domain. The four inner monitoring “arcs” can be seen.

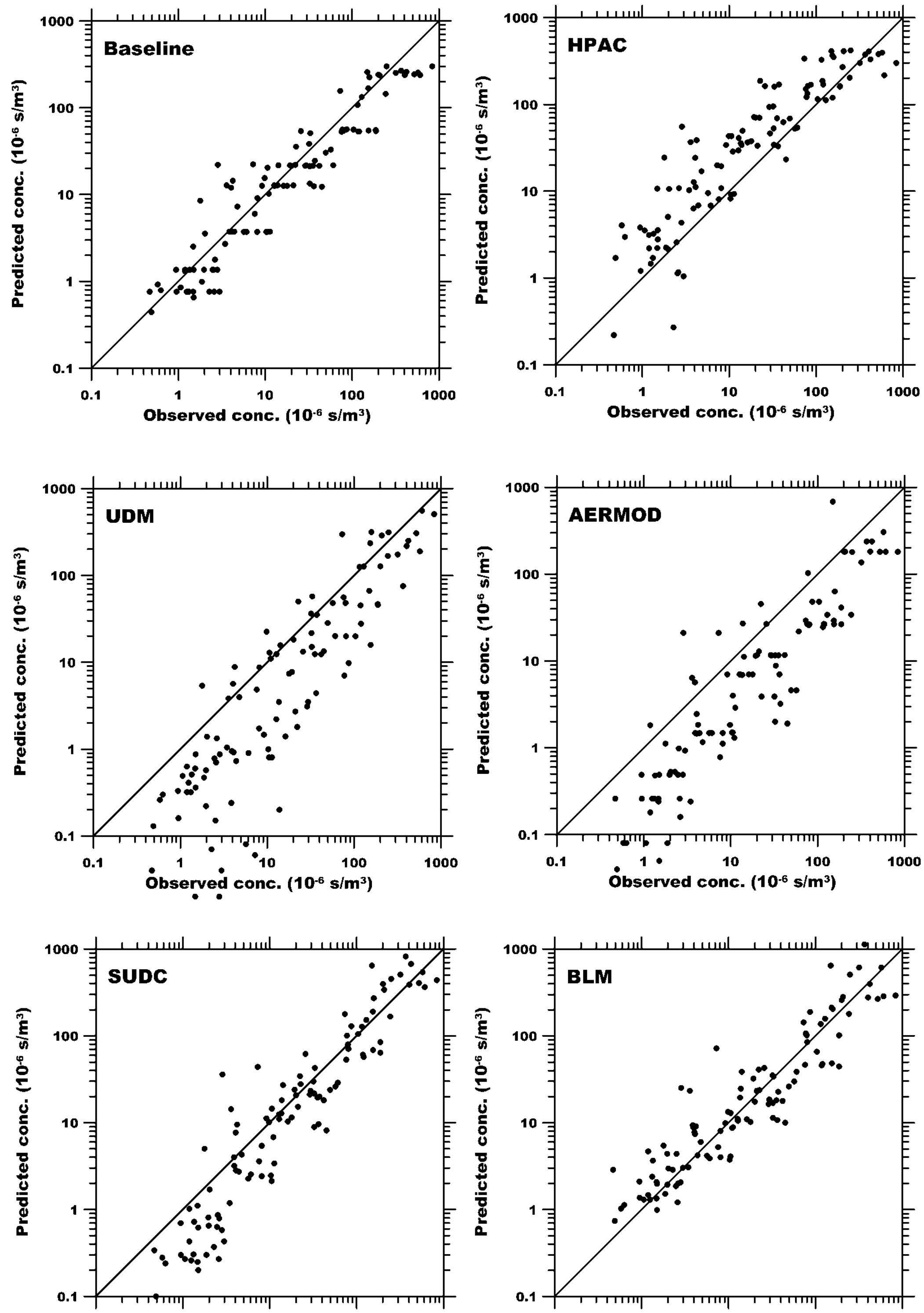

Figure 3. Scatter plots of predicted versus observed maximum hourly-averaged C/Q for each trial and monitoring arc. In general, the largest C/Q values are on the closest arcs.

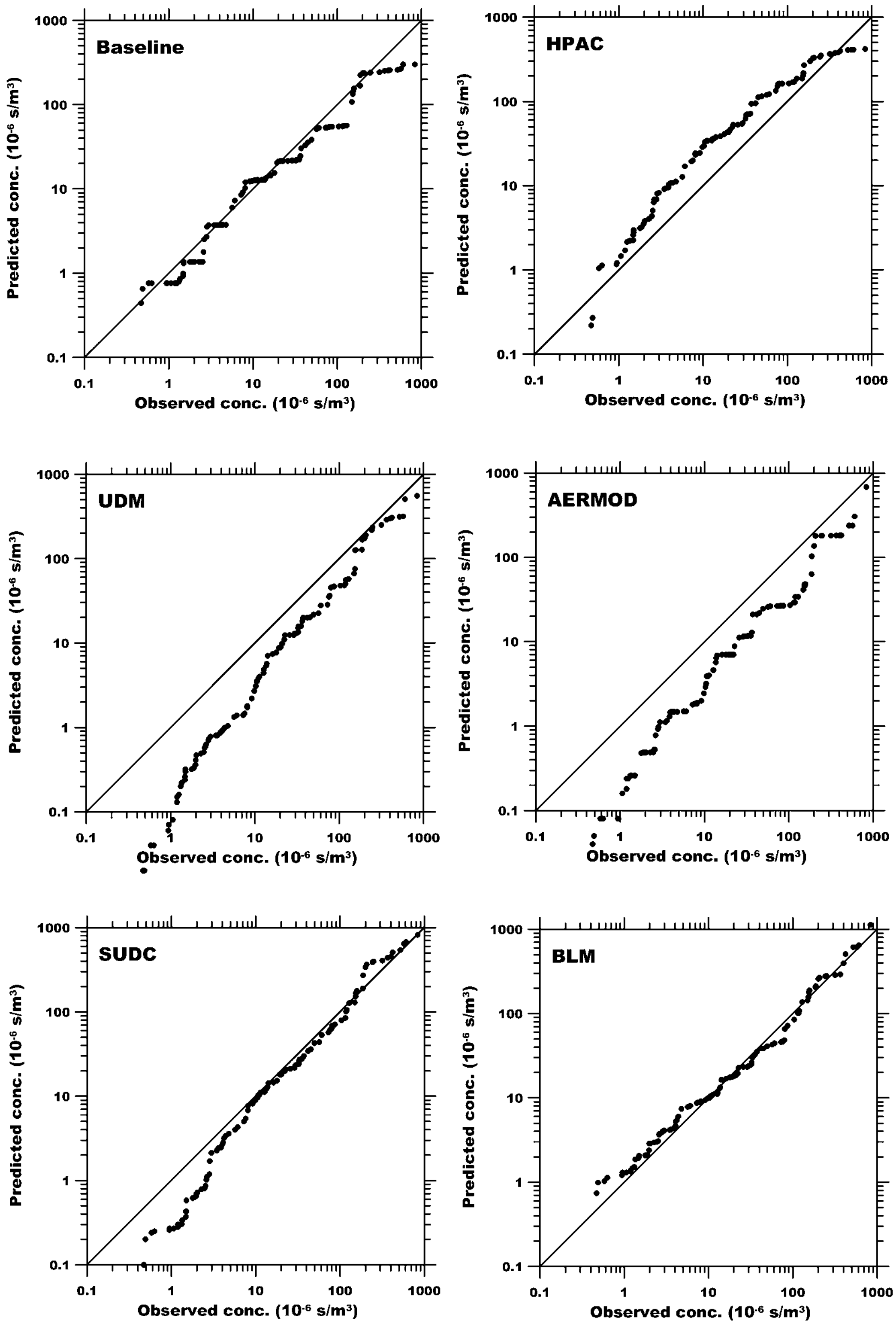

Figure 4. Quantile-Quantile plots of predicted and observed maximum hourly-averaged C/Q for the 18 trials and seven monitoring arcs. In general, the largest C/Q values are on the closest arcs.

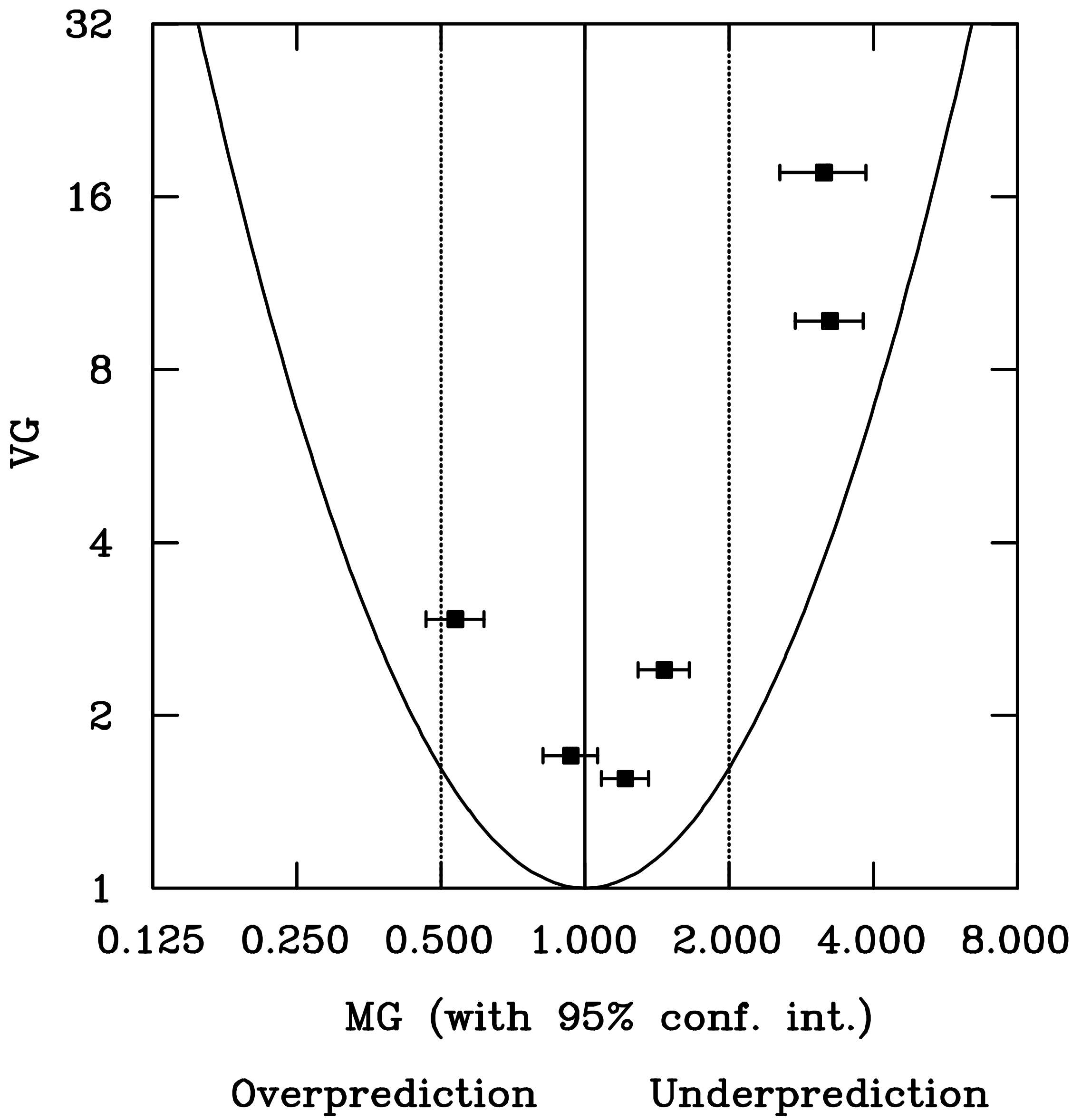

Figure 5. Plot of the performance measures MG versus VG, using the results in Table 2. MG is a measure of the relative bias and VG is a measure of the relative scatter. A perfect model would have MG = VG = 1.0. Lateral lines indicate 95% confidence intervals on MG. The parabola indicates the minimum VG for a value of MG. Dashed lines indicate factor-of-2 bias.

Table 1. Listing of observed C/Q (10^{-6} s/m^3) for each of the 18 trials and seven monitoring arcs. These are the data used in the evaluations of the six models' predictions of C/Q.

			Sampling Arc and Distance (m) From Source						
IOP	Trial	u (m/s)	Arc 1 156	Arc 2 394	Arc 3 675	Arc 4 928	Arc 5 1974	Arc 6 3907	Arc 7 5998
2	1	0.81	317.7	79.6	14.2	3.58	3.91	1.97	0.47
2	2	0.61	421.2	103.7	2.85	n/a	n/a	n/a	n/a
2	3	0.5	366.1	86.8	7.26	1.39	n/a	n/a	n/a
4	1	1.13	606.3	120.1	35.4	16.2	5.67	2.81	1.48
4	2	0.94	836.1	154.7	29.5	12.7	4.39	0.94	0.95
4	3	0.76	573.1	186.7	60.6	21.1	11.3	2.53	2.28
5	1	0.64	149.6	77.3	22.2	13.8	3.87	1.19	2.98
5	2	0.91	249.4	80.5	19.3	13.6	4.09	1.35	1.32
5	3	1.06	402.2	118.3	20.1	12.8	8.09	1.5	1.25
7	1	1.01	520	187.7	32.4	17.9	6.08	1.98	n/a
7	2	1.04	200.6	25.8	28.9	9.16	10.3	2.47	2.59
7	3	1.21	207.6	75.8	41.8	36.4	10.5	3.13	1.85
9	1	2.69	129.3	49.6	44.9	4.77	n/a	0.63	0.49
9	2	2.47	243.7	56.8	32.5	n/a	n/a	1.06	1.01
9	3	3.23	115.3	37.3	11	7.56	2.63	1.5	N/a
10	1	1.51	158	33	10.6	4.05	2.03	1.19	n/a
10	2	2.16	153.4	31.9	9.84	8.1	3.45	1.87	n/a
10	3	2.31	72.9	22.7	4.22	1.78	1.48	0.58	n/a
Averaged over all IOPs and trials		1.39	317.9	84.9	23.8	11.6	5.56	1.67	1.52

Table 2. Listing of statistical performance measures (defined in equations 1 through 5) for the six urban dispersion models. The focus is on maximum hourly-averaged C/Q observed and predicted on each monitoring arc for each of the 18 trials.

	Obs.	Baseline	HPAC	UDM	AERMOD	SUDC	BLM
Highest (10^{-6} s/m^3)	836	299	418	557	684	822	1134
2nd Highest (10^{-6} s/m^3)	606	299	410	508	305	674	646
FB	n/a	0.373	-0.226	0.414	0.684	-0.090	-0.054
NMSE	n/a	1.90	1.32	1.76	4.87	1.48	2.50
MG	n/a	1.22	0.54	3.15	3.25	1.46	0.93
VG	n/a	1.55	2.94	17.6	9.73	2.40	1.70
FAC2	n/a	0.720	0.477	0.346	0.234	0.542	0.710

Table 3. Results of significance tests for the relative mean bias, MG. An "X" means that there is 95% confidence that the values of MG calculated for the two models are significantly different from 0.0.

MG	Baseline	HPAC	UDM	AERMOD	SUDC	BLM
Baseline		×	×	×	×	×
HPAC			×	×	×	×
UDM					×	×
AERMOD					×	×
SUDC						×

Table 4. Results of significance tests for the relative variance, VG. An "X" means that there is 95% confidence that the values of VG calculated for the two models are significantly different from 0.0.

VG	Baseline	HPAC	UDM	AERMOD	SUDC	BLM
Baseline		×	×	×	×	
HPAC			×	×		×
UDM					×	×
AERMOD					×	×
SUDC						×