1.36 COMPARISONS OF TRANSPORT AND DISPERSION MODEL PREDICTIONS OF THE URBAN 2000 FIELD EXPERIMENT

S. Warner, N. Platt and J.F. Heagy
Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311-1882 USA

INTRODUCTION

Within the U. S. Department of Defense, interest in understanding transport and dispersion in an urban environment stems from the need to reliably estimate the population effects resulting from releases of chemical or biological agents. Such estimates require knowledge of the concentrations of dispersed material as a function of time and location. The Defense Threat Reduction Agency (DTRA) has begun to augment their current Hazardous Prediction and Assessment Capability (HPAC – *DTRA* 2001) to include urban effects. In response to this difficult challenge, field studies in which tracer gases are released within an urban environment have been conducted to study flow and dispersion. In this study, we compare the predictions of Urban HPAC to the observations of the Urban 2000 field trial.

BRIEF DESCRIPTION OF URBAN 2000

A series of sulfur hexafluoride (SF_6) releases were carried out in the Salt Lake City area in October 2000 – referred to as "Urban 2000" (*Allwine et al.,* 2002). Meteorological and tracer measurements were conducted throughout the Salt Lake City urban region with an outermost arc of SF_6 samplers located 6 km downwind of the release. For each of six intensive operating periods (IOPs), 3 separate short-duration SF_6 releases were monitored for two hours. For each of these 18 separate 2-hour monitoring periods, this study considers the SF_6 samples that were collected at 30-minute intervals at 66 ground locations. Therefore, one can consider 72 (18×4) 30-minute average concentrations $\times$ 66 locations, or 4752 (30-minute average) concentrations, for "paired in space and time" comparisons to the predictions of a given model.

BRIEF DESCRIPTION OF URBAN HPAC

HPAC is composed of a suite of software modules that can generate source terms for hazardous releases, retrieve and prepare meteorological information for use in a prediction, model the transport and dispersion of the hazardous release over time, and plot and report the results of these calculations. By using HPAC to provide predictions in an urban environment, one can conveniently capture some of the effects of the urban canopy on transport and dispersion by setting the surface type to "urban." In addition to this baseline Urban HPAC predictive capability described above, Urban HPAC offers an urban dispersion model (UDM) and an urban windfield module (UWM) either or both of which can be invoked. In order to use UDM and UWM, Urban HPAC provides a building database that provides the locations, planar geometries, and heights of buildings to support the calculation of flows in the urban regime. The U. K.'s Defense Science and Technology Laboratory developed the UDM component of Urban HPAC (*Hall et al.,* 2002). The UWM predicts steady-sate winds inside the urban boundary layer using a canopy parameterization (*Lim et al.,* 2002).

PREDICTIONS CONSIDERED

Four types of Urban HPAC predictions were examined: HPAC with surface type entered as "urban," denoted baseline or "UC" (for urban canopy); HPAC with the UDM toggled on, denoted "DM"; HPAC with the UWM toggled on, denoted "WM"; and HPAC with both the UDM and the UWM toggled on, denoted "DW." For each model type, 5 weather input options were examined: surface measurements from the Salt Lake City airport located about 10 km from the downtown area (SLC), wind profile measurements from the Raging Waters site located 5 km upwind of the release (RGW), measurements from the top of the Latter Day Saints' administration building located within the urban area (LDS), all available meteorological measurements (ALL), and an Operational Multiscale Environment Model With Grid Adaptivity (*Bacon et al.,* 2000) forecast (OMG). Therefore, a total of 20 model configurations were created – 4 dispersion configurations × 5 meteorological input options. A more detailed description of the preparation of these predictions can be found in *Boybeyi et al* (2003).

PROTOCOL FOR PAIRED IN SPACE AND TIME COMPARISONS

For this analysis, predictions and observations paired in space and time were compared, referred to here as "point-to-point" comparisons. For each release, 30-minute average concentrations in parts per trillion (ppt) – 4 per release – and 2-hour integrated concentrations or dosages (ppt min) were examined. Unlike analyses based on derived quantities (e.g., "plume width," arcmax), this examination explicitly considers the size, shape, and specific location (not just the downwind distance) of the cloud. Therefore, it is expected that the ability of the model (to include the input wind field information) to match the wind speeds and directions will be of vital importance for a point-to-point analysis.

The following procedures were followed for these point-to-point comparisons. First, an assumed background value of 3 ppt was added to all predictions (*Allwine et al.,* 2002). Next, observations that were less than 3 ppt were set to 3 ppt – the estimated background. Some values in the distributed set of observations were denoted with a "-999" which indicated that they were unusable for any of several reasons. These observations, and their paired predictions, were removed from these comparisons. No other changes to the observations or predictions were made. The above procedure was applied directly for 30-minute average concentration comparisons. Two-hour dosages were considered valid for comparison only when all four 30-minute average concentration observations existed. For the 30-minute average concentration comparisons, 94 percent of all possible observation/prediction pairings were included via the above protocol. Similarly, 81 percent of 2-hour dosage pairings were included via this protocol.

NON-PARAMETRIC HYPOTHESIS TESTING FOR DIFFERENCES

In addition to the previously described user-oriented two-dimensional measure of effectiveness (*Warner et al.,* 2001), 13 one-dimensional statistical measures were computed for each model configuration. Confidence intervals were computed for all measures using a bootstrap procedure (*Warner et al.,* 2004a). Definitions for two of the one-dimensional measures discussed in this paper – Fractional Bias (FB) and Normalized Absolute Difference (NAD) – are given below:

$$FB = \frac{\left(\overline{C_p} - \overline{C_o}\right)}{0.5\left(\overline{C_o} + \overline{C_p}\right)} \qquad (1) \quad \text{and} \quad NAD = \frac{\sum_{i=1}^{n}\left|C_p^{(i)} - C_o^{(i)}\right|}{\sum_{i=1}^{n}\left(C_o^{(i)} + C_p^{(i)}\right)}. \qquad (2)$$

where C = observation/prediction of interest (e.g., 30-minute average concentration), C_p corresponds to model prediction, C_o corresponds to observation, a bar above the quantity (e.g., $\overline{C}$) denotes the average, n = number of data points used in the comparisons, $C_o^{(i)}$ refers to the ith observed concentration (or dosage), and similarly, $C_p^{(i)}$ refers to the ith predicted concentration (or dosage). A perfect prediction will have a two-dimensional user-oriented MOE score of (1,1), no scatter (NAD = 0), and no average bias (i.e., FB = 0).

Given the very large set of comparisons generated from the above (over 25,000 were examined, *Warner et al.,* 2004b), an objective way for logically identifying statistically significant differences between sets of model predictions was needed. We computed p-values for the hypothesis testing of differences between estimated metrics associated with different model configurations. Significant differences between models were sought by comparing the statistic of interest for two models (at a time) for each of the 18 releases. The statistic of interest, for example, FB, is paired by release (e.g., IOP 4, Release 4) for each prediction set. For the one-dimensional metrics, the Permutation test with general scores (*Cytel,* 1998) – a non-parametric test similar to the Sign and Wilcoxon Signed-Rank tests (*Sprent,* 1998) – was used to compute p-values. For suitably low p-values, one can reject the null hypothesis of equivalence between the values being compared (*Daniel,* 1997 and *Sprent and Smeeton,* 2001). The Permutation test uses the absolute magnitude of the paired differences and is generally more powerful than the Wilcoxon Signed-Rank test, which only uses the rankings of the paired data. Figure 1 illustrates the value of the above hypothesis testing procedure for detecting significant differences. The point estimates and 0.95 approximate confidence intervals for FB are shown in Fig. 1a for the four model configurations that used the ALL weather input option. With respect to FB, variations between model performances can be seen in Fig. 1a but the statistical significance of these differences is not readily apparent. Next, Fig. 1b shows the estimated mean FB *difference* for the six comparisons that result from the consideration of the four model configurations. Now, if two sets of predictions with similar FB performance are compared, one expects the mean FB difference to be zero. In fact, one of these mean differences ("DM-DW") straddles zero, given its 0.95 confidence interval. But it is also clear that for some of these comparisons, the mean FB difference is different from zero and a more formal test reveals a quantitative result associated with that difference – the p-value. For the example shown in Fig. 1b, the p-values strongly suggest that the differences in FB that involve comparisons to the WM model configuration are statistically significant and that the WM_ALL predictions led to larger ("worse") FB values relative to the DM_ALL, UC_ALL, and DW_ALL predictions.

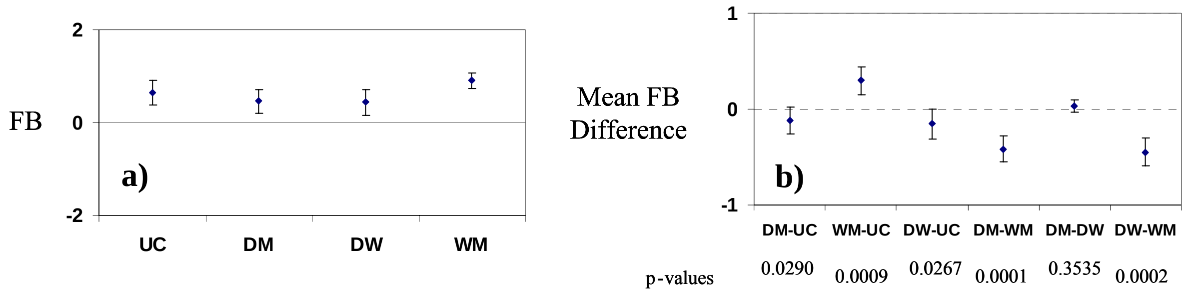

Figure 1. a) FB values for the four dispersion model configurations using the ALL weather input option. These values are based on model predictions of 30-minute average concentrations. b) Mean FB differences for the six model configuration comparisons.

For the two-dimensional MOE, the chosen hypothesis test procedure starts by computing vector differences between various paired (by release) model predictions. If two sets of model predictions were identical, then the 18 vector differences would be (0,0). For this study, the null hypothesis is that the two models being compared are equivalent. Under this hypothesis, any MOE vector difference is expected to be equally likely to reside in any of the four quadrants, defined as positive x, positive y (++); positive x, negative y (+-); negative x, positive y (-+); and negative x, negative y (--). Given this null hypothesis, one tests how unlikely the observed result is by simulating the appropriate permutations. For example, consider a comparison of predictions of model A and model B ("A-B") for the 18 separate and independent releases of Urban 2000. If 10 of 18 differences were found in the ++ quadrant, the implication would be that model A led to an improved MOE value (closer to (1,1)) relative to model B and with a computed p-value of 0.0216. For the following discussions, 0.0216 is the level of significance assumed for statistically based statements. The 2-dimensional, 4-quadrant hypothesis test described here is a natural extension of the one-dimensional Sign test.

EXAMPLE RESULTS: COMPARISONS OF MODEL CONFIGURATIONS

Based on examinations of the MOE and statistical measures of scatter, it was found that the "best" predictions were, in general, associated with the Urban HPAC model that included UDM ("DM") and the specific UC_OMG configuration. The words "best" and "better" as used in this discussion of comparative model configuration performance are based on the finding of less scatter and/or an MOE value closer to the perfect value of (1,1). Similarly, the words "worst" and "worse" imply more scatter and/or an MOE value further from the perfect value of (1,1). These findings relate only to these Urban HPAC predictions of a single field experiment – Urban 2000. The predictions that included only changing the surface type to "urban" – that is, "baseline" Urban HPAC – or included only UWM typically performed worst (again with the exception of the UC_OMG configuration). In general, there were no significant improvements due to including UWM with UDM relative to using UDM alone. For the exceptional case in which the OMG weather input option was used, the baseline Urban HPAC predictions that did not include UDM and UWM were best. These predictions were especially unusual for these Urban HPAC predictions of Urban 2000 in that they were relatively unbiased (that is, they were neither over- nor under-predictions) whereas the other 19 sets of predictions had FB values implying over-prediction. This particular UC_OMG-related finding may be the result of counteracting effects or compensating errors. Further investigation is required.

Figure 2 illustrates the above findings by providing a relative ranking of model performance based on quantitative statistical testing of predictions (based on all sampler locations). For each weather input option, the 6 comparisons that result from the four model configurations (UC, DM, WM, and DW) were examined using the previously defined hypothesis test procedures. If a comparison of two sets of predictions led to a p-value ≤ 0.0216 for either the MOE or NAD and for either the 30-minute average concentration or 2-hour dosage predictions, then the better performing model configuration is ranked above the worse performing configuration in Fig. 1. For four of the five weather input options (SLC, RGW, LDS, and ALL), the DM predictions were ranked best or tied for best and the UC predictions were ranked worst or tied for worst. For the predictions run with the OMG weather input option, the exceptional result is that the UC_OMG configuration is ranked as best.

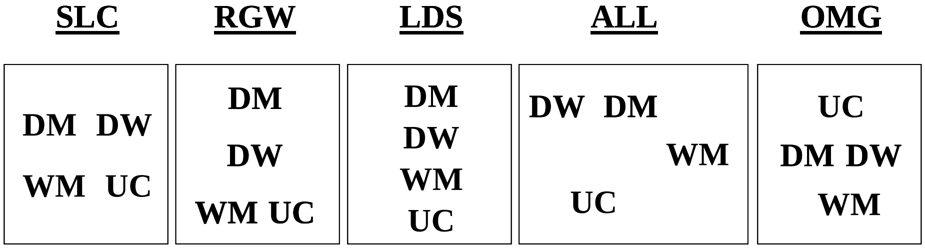

Figure 2. Rankings of Urban HPAC dispersion model configuration performance as measured by the MOE and NAD. For the ALL weather input option comparisons, the WM predictions could be ranked as below DM, above UC, and not statistically different from DW.

This study also revealed results associated with model performance as a function of time integration (30-minute versus 2-hour averages), model performance as a function of downwind distance, model performance when used to predict hazard locations (vice actual amounts of material), and the relative performance of the five weather input options that were examined.

CONCLUSIONS

An inclusive, paired in space and time – point-to-point – comparative protocol and quantitative hypothesis testing were used to compare the predictions of the Urban 2000 field experiment by 20 Urban HPAC model configurations. Robust findings were reported that included the statistical discernment of differences between configurations and general trends of Urban HPAC predictions of Urban 2000.

ACKNOWLEDGEMENTS

The effort is supported by the Defense Threat Reduction Agency with Mr. Richard Fry as project monitor. The views expressed in this paper are solely those of the authors.

REFERENCES

Allwine, K. J., J. H. Shinn, G. E. Streit, K. L. Clawson, and M. Brown, 2002: Overview of URBAN 2000, A multiscale field study of dispersion through an urban environment. *Bull. Amer. Meteor. Soc.*, **83**, 521-536.

Bacon, D. P., N. N. Ahmad, Z. Boybeyi, T. J. Dunn, M. S. Hall, P. C. S. Lee, R. A. Sarma, M. D. Turner, K. T. Waight III, S. H. Young, and J. W. Zack, 2000: A dynamically adapting weather and dispersion model: the operational multiscale environment model with grid adaptivity (OMEGA). *Monthly Weather Review*, **128**, 2044-2076.

Boybeyi, Z., J. C. Chang, S. R. Hanna, P. Franzese, N. Platt, and S. Warner, 2003: Independent evaluation of Urban HPAC with the Urban 2000 field data. Prepared for DTRA by the Institute for Defense Analyses and George Mason University, 341pp.

Cytel, 1998: StatXact 4 for Windows, User Manual, Cytel Software Copyright, 941 pp.

Daniel, W. W., 1997: Applied Nonparametric Statistics. Wadsworth, 656 pp.

DTRA, 2001: The hazard prediction and assessment capability (HPAC) user's guide version 4.0, HPAC-UGUIDE-02-U-RAC0, prepared for DTRA by Science Applications International Corporation (SAIC), 10260 Campus Point Drive, San Diego, CA 92121, 598 pp.

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, 106 pp.

Lim, D., D. S. Henn, and R. I. Sykes, 2003: UWM Version 0.1, Technical Documentation, Titan Corporation, ARAP Group, P.O. Box 2229, Princeton, NJ, 08543-2229, 35 pp.

Sprent, P., 1998: Data Driven Statistical Methods. Chapman and Hall, 406 pp.

Sprent, P. and *N. C. Smeeton*, 2001: Applied Nonparametric Statistical Methods, 3d ed. Chapman and Hall, 461 pp.

Warner, S., N. Platt, and J. F. Heagy, 2001: User-oriented measures of effectiveness for the evaluation of transport and dispersion models, *Proc. of the Seventh Int. Conf. on Harmonisation Within Atmospheric Dispersion Modelling for Regulatory Purposes*, Belgirate, Italy, 24-29.

Warner, S., N. Platt, and J. F. Heagy, 2004a: User-oriented two-dimensional measure of effectiveness for the evaluation of transport and dispersion models. *J. Appl. Meteor.*, **43**, 53-73.

Warner, S., N. Platt, and J. F. Heagy, 2004b: Comparisons of transport and dispersion model predictions of the Urban 2000 field experiment. In press, *J. Appl. Meteor.*