1.25 APPLICATION OF USER-ORIENTED MOE TO TRANSPORT AND DISPERSION MODEL PREDICTIONS OF ETEX

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

INTRODUCTION

In October 1994, the inert, environmentally safe, tracer gas perfluoro-methyl-cyclohexane (PMCH) was released over a 12-hour period from a location in northwestern France and tracked at 168 sampling locations in 17 countries across Europe – hundreds of kilometers (*Graziani et al.,* 1998). This release, known as the European Tracer Experiment (ETEX), resulted in the collection of a wealth of data. IDA has obtained from the Joint Research Centre, European Commission 46 sets of transport and dispersion predictions associated with models from 17 countries as well as the observed PMCH sampling data associated with the October 1994 ETEX release (*Mosca et al.,* 1998a). This paper describes the extension of the previously developed user-oriented two-dimensional measure of effectiveness (MOE) methodology (*Warner et al.,* 2001) to evaluate the predictions of these 46 models against the long-range *ETEX* observations.

MEASURE OF EFFECTIVENESS (MOE)

A fundamental feature of any comparison of hazard prediction model output to observations is the overlap, over-, and under-prediction regions. We define *false negative* region where hazard is observed but not predicted and *false positive* region where hazard is predicted but not observed. Numerical values associated with estimates of the false negative region (A_{FN}), the false positive region (A_{FP}), and the overlap region (A_{OV}) characterize this conceptual view (Figure 1).

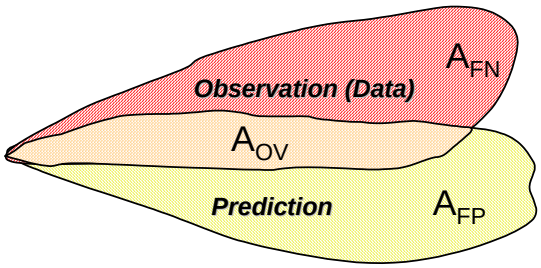

Figure 1. Conceptual View of 3 Comparative Dimensions

The MOE that we consider is two-dimensional (2D). The x-axis corresponds to the ratio of overlap area to observed area and the y-axis corresponds to the ratio of overlap area to predicted area. These mathematical definitions can be algebraically rearranged and we then recognize that the x-axis corresponds to 1 minus the false negative fraction and the y-axis corresponds to 1 minus the false positive fraction.

$$2D\ MOE = \left(\frac{A_{OV}}{A_{OB}}, \frac{A_{OV}}{A_{PR}} \right) = \left(\frac{A_{OV}}{A_{OV} + A_{FN}}, \frac{A_{OV}}{A_{OV} + A_{FP}} \right) = \left(1 - \frac{A_{FN}}{A_{OB}}, 1 - \frac{A_{FP}}{A_{PR}} \right) \quad (1)$$

where A_{PR} = region of the prediction and A_{OB} = region of the observation. Importantly, this MOE considers the direction of the plume or location of the hazard when evaluating via "scoring" functions a model's performance. The various regions described above can be estimated from field trial observations as previously described (*Warner et al.,* 2004).

"SCORING FUNCTIONS"

In this section we describe several notional scoring functions for the MOE space. Essentially, these scoring functions can be thought of as corresponding to the requirements of different possible model users. Such scoring functions can thus aid us in assessing if a model's MOE value, for a given set of field observations, is "good enough." In developing the MOE scoring functions, this section also describes and illustrates the mathematical relationships between the figure of merit in space, fractional bias, and a measure of scatter between observations and predictions. An objective scoring function associated with the 2D-MOE space is to consider the values closest to (1,1) as the best. This approach considers false negative and false positive fractions as *equally* undesirable. For such an objective scoring function (OSF) we define the "distance" to (1,1) – d_{OSF} as

$$d_{OSF} = \sqrt{\left(1 - \frac{A_{OV}}{A_{OB}}\right)^2 + \left(1 - \frac{A_{OV}}{A_{PR}}\right)^2} = \sqrt{\left(\frac{A_{FN}}{A_{OB}}\right)^2 + \left(\frac{A_{FP}}{A_{PR}}\right)^2} \,. \qquad (2)$$

Then, for different MOE values, OSF favors the smallest value of d_{OSF}. Next, the Figure of Merit in Space (FMS) is defined as the ratio of the intersection of the observed and predicted areas to the union of the observed and predicted areas at a fixed time and above a defined threshold concentration (*Mosca et al.,* 1998b). FMS is related to x and y components of the MOE as shown in the following equation:

$$FMS = \frac{xy}{x + y - xy} \,. \qquad (3)$$

A value of FMS = 1 implies complete overlap. Some users of hazardous material transport and dispersion models might consider false positives and false negatives quite differently. For many applications, false positives would be much more acceptable to the user than false negatives (which could result in decisions that directly lead to death or injury). Equation 4 is an example of a user scoring function that takes the above risk tolerance into consideration. We refer to this notional user scoring function as the Risk-Weighted FMS (RWFMS):

$$RWFMS = \frac{xy}{xy + C_{FN}\, y(1 - x) + C_{FP}\, x(1 - y)} \qquad (4)$$

where C_{FN}, $C_{FP} > 0$. Basically, this equation describes a modified FMS that includes coefficients, C_{FN} and C_{FP}, to weight the false negative and false positive regions, respectively. A hazardous material transport and prediction model might be applied to problems for which the actual location of the hazard or direction of the plume is of no particular importance. For example, such a model might be used to study potential future outcomes of an accidental or intentional release. In these cases, the actual weather (e.g., wind speed and direction) of the far future associated with the planning cannot be known with any certainty. For these applications it is desirable to have a scoring function that simply compares the sizes of the predicted and observed areas. In essence, model users in these cases would want a model that minimizes the overall model bias. Fractional bias (FB) has been used to evaluate transport and dispersion models under such circumstances and can also be related to the MOE:

$$FB = \frac{2(x-y)}{x+y}. \tag{5}$$

A scoring function that minimizes the absolute value of FB favors model predictions that lead to the least average bias – i.e., just the right amount of material or size of the hazard region.

Quite often, measures such as mean square error or normalized mean square error are used to characterize the differences between observed and predicted quantities – the scatter if you will. It is desirable to have a measure of scatter that can be mathematically related to the MOE space. For this purpose we define a specialized version of a measure of scatter – normalized absolute difference (NAD) – between observations and predictions:

$$NAD = \frac{\sum_{i=1}^{n} \left| C_p^{(i)} - C_o^{(i)} \right|}{\sum_{i=1}^{n} \left(C_o^{(i)} + C_p^{(i)} \right)} \tag{6}$$

where n = number of data points used in the comparisons and $C_o^{(i)}$ refers to the ith observed concentration, and similarly, $C_p^{(i)}$ refers to the ith predicted concentration. NAD can be expressed in terms of the false negative, false positive and overlap and, after substitution and algebraic simplification, NAD is related the summed concentration-based MOE components as

follows:
$$NAD = \frac{x+y-2xy}{x+y}. \tag{7}$$

A strictly monotonic relationship between NAD and FMS implies that scoring model predictive performance based on NAD or FMS will necessarily lead to identical rank orderings (*Warner et al.,* 2003). For both NAD and FB, the mathematical relationships with the summed concentration-based MOE can be generalized to the threshold-based MOE.

EXAMPLE COMPARATIVE RESULTS

Figure 2 presents MOE values for the 46 models when a threshold of 0.1 ng m^{-3} is applied. The model numbers shown in Figure 2 correspond to the Atmospheric Transport Model Evaluation Study (ATMES) II predictions that have been previously described (*Mosca et al.,* 1998a). The numbers in Figure 2 correspond to the model number with the unbolded labels referring to the "100 series" (e.g., the unbolded "12" implies model 112) and the bolded labels referring to the "200 series" (e.g., the bold "8" implies model 208). The 100 series models (101-135) used European Centre for Medium Range Weather Forecasts (ECMWF) analyzed meteorological data as input and the 200 series (201-214) used weather inputs selected by the modeler.

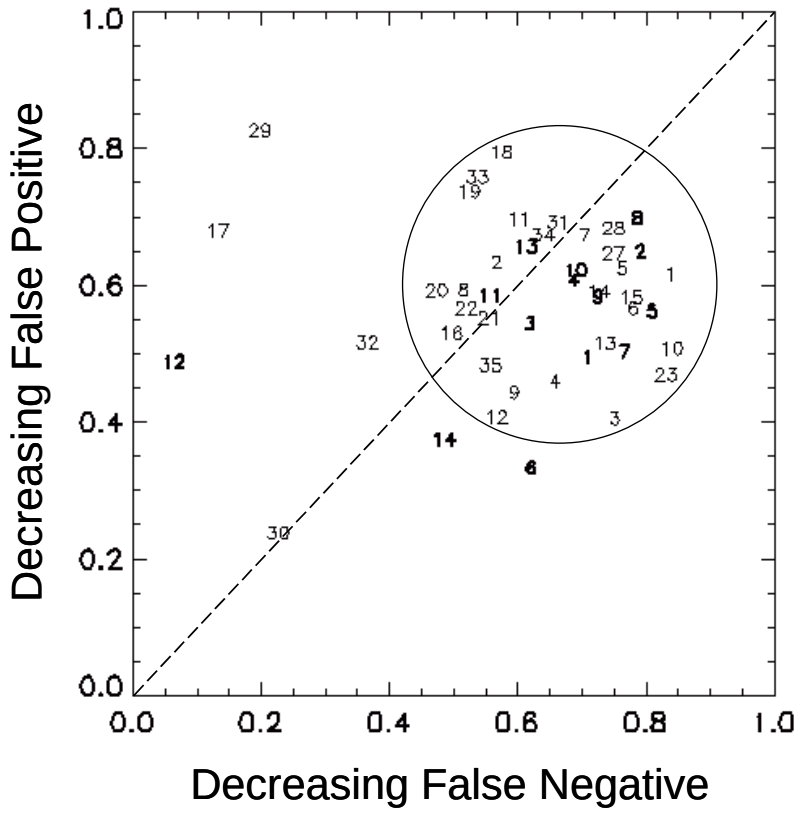

Figure 2. 3-Hour Average Concentration Threshold-Based MOE (0.1 ng m-3) Values for 46 ATMES II Participants. Unbold Labels Refer to Series 100 Models (e.g., "19" implies model 119) and Bold Labels Refer to Series 200 Models (e.g., "13" implies model 213).

Forty models led to MOE values within the (arbitrary) circular region highlighted in Figure 2 and the predictions of seven models led to MOE values outside this region (117, 129, 130, 132, 206, 212, and 214). An MOE value on the diagonal (the dashed light purple line) implies equal sizes of the observed and predicted regions – although not necessarily collocation. Twenty-seven of 46 MOE values lie below the diagonal, indicating over-prediction with respect to the number of locations with 3-hour average concentrations above 0.1 ng m^{-3}. MOE values were also computed at thresholds of 0.01 and 0.5 ng m^{-3} and for total summed concentrations (*Warner et al.,* 2003). Table 1 identifies the top ranked model predictions as judged by the OSF as well as the rankings (out of 46) of the Second-Order Closure Integrated Puff (SCIPUFF) and Atmospheric Release Advisory Center (ARAC) models, two models in which our sponsor had a particular interest. Rankings are identified for the three threshold-based and summed concentration-based MOE values. No single model dominated the top ranking. Complete rankings can be found in *Warner et al.,* 2003. The rankings described in this paper result from consideration of a single release and general inference about which model is "best" or ranked highest is not appropriate. Rather, these rankings describe performance in terms of this specific release only. In addition, for this single release field experiment, no direct measures of uncertainty associated with the computed MOE values or model rankings were readily available. However, variations in MOE values as a function of time after the release and sensitivities of the MOE values and rankings to the influence of a single sampler location were examined.

Table 1. Top-Ranked Models and Rankings of SCIPUFF and ARAC Based on MOE Values and the Objective Scoring Function.

Rank	0.01 ng m^{-3}	0.1 ng m^{-3}	0.5 ng m^{-3}	Summed Concentration
1	Canadian Meteorological Centre (202)	Swedish Meteorological and Hydrological Office (208)	ARAC (127)	German Weather Service (107)
Model	**0.01 ng m^{-3}**	**0.1 ng m^{-3}**	**0.5 ng m^{-3}**	**Summed Concentration**
SCIPUFF (121)	24	30	23	41
ARAC (127)	4	5	1	33

Past analysis has suggested that assessments of model performance could be sensitive to the results associated with a single sampling location – in particular, the location closest to the release where the concentrations would be highest. The sensitivity of MOE values to this phenomenon was examined by re-computing MOE values after the removal of a single sampling location. Each of the 168 sampling locations was removed (one at a time) generating 168 additional MOE values. For the MOE values based on summed concentrations (but not threshold exceedance), there was indeed a sensitivity associated with the sampling location closest to the release – at Rennes, France. While most of the models' MOE values were relatively unaffected by the removal of this location, a few were perhaps, overly influenced by this single sampler location. The two models of particular interest here, SCIPUFF and ARAC, were two of about 8 (of 46) model predictions that resulted in MOE values that were significantly influenced by the removal of the sampler location at Rennes. With the exclusion of the sampler location at Rennes, the OSF-based model rankings (summed concentration) for SCIPUFF and ARAC changed from 41 and 33 (Table 1) to 34 and 8, respectively.

Finally, variation in model predictive performance as a function of time, as judged by the MOE, was examined. Portions of the ETEX sampling network were monitored out to 90 hours after the release. Three-hour average concentrations (predictions and observations) were compared for 30 time periods and 12-hour running time window (i.e., 4 time periods in sequence combined) and 24-hour running time window (i.e., 8 time periods in sequence combined) MOE values were also computed. When judging model predictive performance using the MOE based on the 0.01 or 0.1 ng m^{-3} thresholds, one of two time-dependent behaviors was typically observed. For some models, an initial under-prediction of the number of locations that exceed the threshold is followed by a "correction" that leads to about the right number of locations predicted above the threshold, followed finally, by degradation that suggests a general missing of the locations at which the threshold is exceeded at the longest times (and distances). For other models, an initial over-prediction of the number of locations that exceed the threshold is followed by a "correction" that leads to about the right number of locations predicted above the threshold, followed again, by degradation that suggests a general missing of the locations at which the threshold is exceeded. SCIPUFF and ARAC both show this degradation (as judged by the 0.01 or 0.1 ng m^{-3} threshold-based MOE) at the longest times after the release, as do most of the examined transport and dispersion models.

FOLLOW-ON STUDY

This analysis has served as a base upon which to build future studies involving ETEX. First, MOE values based on actual areas (e.g., square kilometers) will be created. An important part of this follow-on effort will be to explore and understand potential sensitivities associated with interpolation given the underlying non-uniform sampler space across Europe. Given area-based MOE values, one can then include European population distributions and notional effects-levels of interest to place the MOE in its ultimate context – fraction of the population falsely warned and fraction of the population inadvertently exposed.

ACKNOWLEDGEMENTS

The authors thank Stefano Galmarini (Joint Research Centre – Environment Institute, Environment Monitoring Unit, Ispra, Italy) for providing access to the ATMES II model predictions and for useful discussions. 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

Graziani, G., W. Klug, and S. Mosca, 1998: Real-Time Long-Range Dispersion Model Evaluation of the ETEX First Release, Joint Research Center, European Commission, Office of Official Publications of the European Communities, L-2985 (CL-NA-17754-EN-C), Luxembourg.

Mosca, S., R. Bianconi, R. Bellasio, G. Graziani, and W. Klug, 1998a: ATEMS II – Evaluation of Long-Range Dispersion Models Using Data of the 1st ETEX Release, Joint Research Center, European Commission, Office of Official Publications of the European Communities, L-2985 (CL-NA-17756-EN-C), Luxembourg.

Mosca, S., G. Graziani, W. Klug, R. Bellasio, and R. Bianconi, 1998b: A statistical methodology for the evaluation of long-range dispersion models: an application to the *ETEX* exercise. *Atmos. Environ.*, **32** (24), 4307-4324.

Sykes, R. I., et al., 2000: PC-SCIPUFF Version 1.3 Technical Documentation, A.R.A.P Report No. 725, Titan Corporation, ARAP Group, December 2000, pages 221- 226.

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, 2004: 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, 2003: Application of User-Oriented MOE to Transport and Dispersion Model Predictions of the European Tracer Experiment, *IDA Paper P-3829*, 86 pp, November 2003. Available by e-mail request to nplatt@ida.org.