User-Oriented Measures of Effectiveness for the Evaluation of Transport and Dispersion Models

S. Warner, N. Platt and J.F. Heagy
Institute for Defense Analyses, 1801 N. Beauregard Street, Alexandria, VA 22311, USA

Keywords: atmospheric transport and dispersion, measure of effectiveness, model performance, qualitative evaluation, quantitative evaluation

1 Introduction

There is an enduring need for measures that communicate the validity of a model's predictions when applied in a given regime (e.g., short-range or longer-range, unstable or stable meteorological conditions). We desire measures that have the following properties: 1) clear-cut interpretation for consistent utility and relatively widespread communication potential, 2) sensitive to real changes in model performance (e.g., can detect performance differences between various models), and 3) robust with respect to small measurement errors and uncertainties.

2 Measure of Effectiveness

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* where hazard is observed but not predicted and *false positive* 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 as shown in Figure 1.

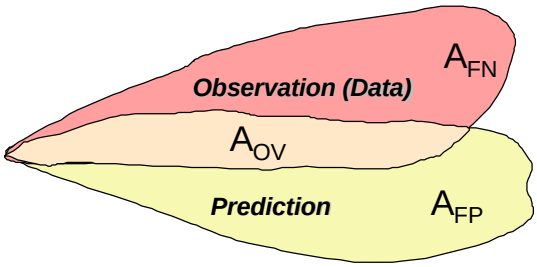


Figure 1 Conceptual View of 3 Comparative Dimensions.

The measure of effectiveness (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} = area of the prediction and A_{OB} = area of the observation. Importantly, this MOE considers the direction of the plume or location of the hazard when "scoring" a model's performance.

The various areas described above can be estimated from field trial observations. For example, we considered the distance traversed by samplers along an arc at which both the observations and predictions led to dosages above some selected threshold, as a natural analog of the overlap area. A similar methodology was used to estimate A_{FN} and A_{FP}. Figure 2 illustrates this procedure.

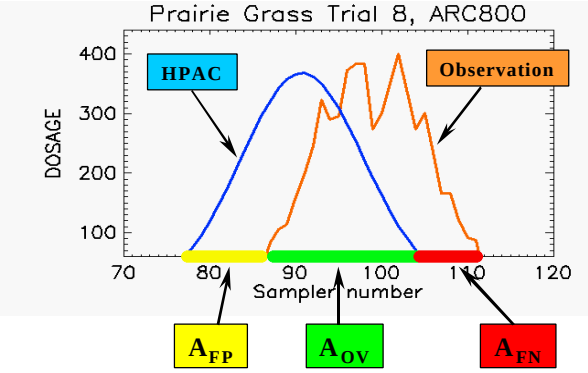


Figure 2 Illustration of Threshold-Based Estimates of A_{OV}, A_{FN}, and A_{FP}.

Another "area" estimate, illustrated in Figure 3, can be computed by integrating the areas under the associated curves and thus computing the summed dosages. This method has a natural extension to yet another area estimate of A_{OV}, A_{FN}, and A_{FP} in terms of a filter that considers the effects of the presumably toxic agent on an exposed population. First, the lethality or effects of the agent being studied and the dosages at a given sampler are examined. For example, probit curves might be used to assess the lethality/effects of exposure at a given sampler. Next, the dosages are converted into a fractional lethal/effects exposure, i.e., that fraction of an exposed population, at that dosage level, that would be expected to become a casualty. Then we apply a procedure similar to the one described by Figure 3 to calculate the marginal fractional lethal exposure obtained in the over- and under-prediction regions. In a sense, the above process filters our interpretation of the above areas through the appropriate lethality/effects "lens."[1]

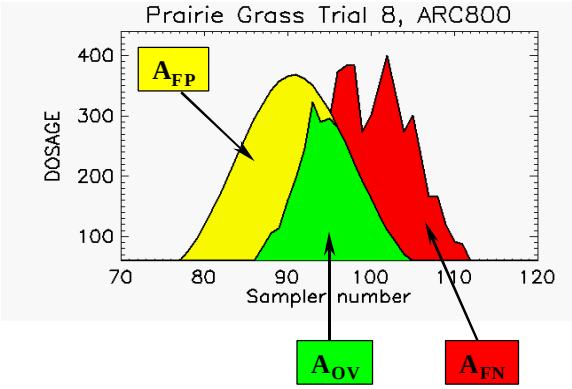


Figure 3 Illustration of Summed Dosage Estimates of A_{OV}, A_{FN}, and A_{FP}.

3 Some Comparative Results

These MOEs were applied to the comparison of Hazard Prediction and Assessment Capability (HPAC) and National Atmospheric Release Advisory Center (NARAC) predictions of the 1956 *Prairie Grass* field trial observations [Barad, 1958].[2]

---

1 In addition to the procedures described above, we also used a variety of interpolation techniques to estimate actual area sizes for A_{FP}, A_{FN}, and A_{OV} based on the *Prairie Grass* field trials. For the densely sampled *Prairie Grass* observations, the interpolation-based MOE results are similar to those described here.

2 For dispersion, HPAC uses the Second-Order Closure Integrated Puff (SCIPUFF) model, which is a Lagrangian model that employs the Gaussian puff numerical method – an arbitrary time-dependent concentration field is represented by a superposition of three-dimensional Gaussian distributions – and bases its turbulent diffusion parameterization on second-order closure theories. Alternatively, the dispersion component of the NARAC model evaluates the 3-D advection-diffusion equation by solving an appropriate stochastic differential equation (SDE) for many (typically hundreds of thousands) test particles via a Lagrangian Monte Carlo method.

Figure 4 compares 2D MOE estimates for HPAC and NARAC predictions. The x-axis corresponds to decreasing false negative fraction, from left to right, and the y-axis corresponds to decreasing false positive fraction, from bottom to top. Each colored cluster of points in the figure represents the estimated 95th percent confidence region for the given 2D MOE point estimate. The complete separation of these two regions implies that the differences between the HPAC and NARAC 2D MOE point estimates are statistically significant.[3]

The MOE estimates on the left (Figure 4a) are based on a threshold of 60 mg-sec/m^3, for instance, comparable to contouring an area at that threshold. We first notice that the MOE values for the HPAC and NARAC predictions are different in a statistically significant sense. That is, the 95th percent confidence regions for the two MOE estimates are separate. The MOE values suggest that, on average, the NARAC predictions led to smaller false negative fractions and larger false positive fractions than the HPAC predictions.

Variances, in a given model's prediction performance as a function of arc range and stability condition, were easily detected and typically led to statistically significant conclusions. Figure 4b presents confidence region estimates for the 2D MOE as a function of meteorological stability category grouping (SCG). For Figure 4b, three independent meteorological stability category groupings were examined: relatively unstable (SCG = 1, 2, 3); neutral (SCG = 4); and relatively stable (SCG = 5, 6, 7) [Irwin, 1998]. The estimates of Figure 3b are based on the application of a lethality/effects filter using a probit model with LD_{50} (that is, the value at which 50 percent of an exposed population would be expected to receive a lethal dosage) = 5,000 mg-sec/m^3 and slope = 12. The MOE clearly separates performance between models, and for a given model, between stability category groupings. In all cases, the NARAC model predictions led to the lower false negative and higher false positive fractions with the biggest differences associated with the neutral trials.

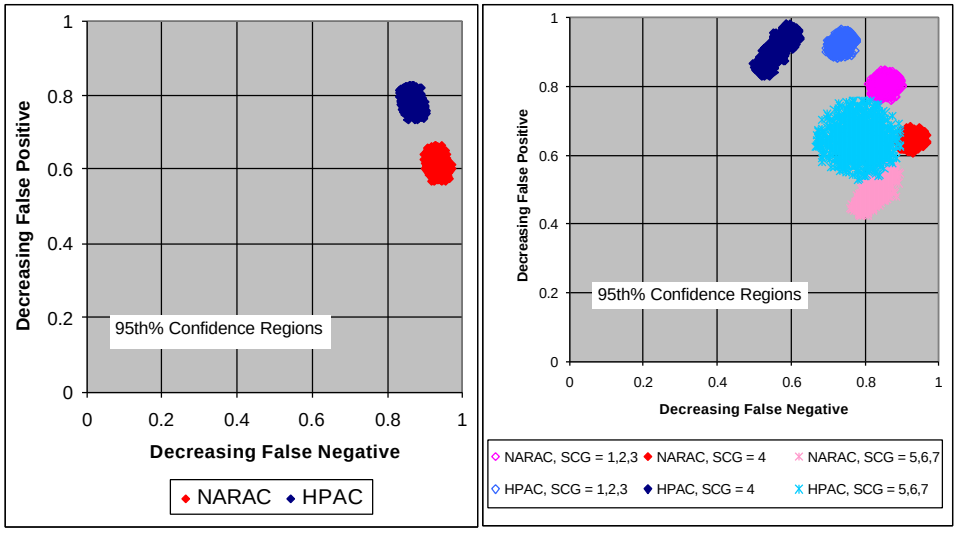


Figure 4. 2D MOE Comparisons of HPAC and NARAC Predictions of 51 *Prairie Grass* Trials 1) All Trials Based on a Threshold of 60 mg-sec/m^3, b) As a Function of Stability Category Grouping for a Probit Model with LD_{50} = 5,000 and Slope = 12.

In addition to providing predictions of the mean value of a plume's location and intensity, HPAC provides probabilistic outputs. HPAC can provide an area, outside of which, the probability of a hazard existing above a given threshold can be defined. For instance, HPAC can provide a plume where the probability of exceeding some dosage is less than, for example, 0.10.

3 Bootstrap (re-sampling) techniques were used to estimate the 95th percent confidence regions.

Figure 5 considers our MOE estimates of such probabilistic prediction outputs for the *Prairie Grass* field trials for a threshold of 60 mg-sec/m^3. The yellow region (centered at about 0.90, 0.76) corresponds to the MOE estimate based on the mean value prediction as we've seen previously (Figure 4a). Starting with the lower right region (at about 0.97, 0.44) we see the resulting estimate from considering the 0.01 plume prediction, that is, outside of that predicted area there would be a 1 percent chance of observing hazard above the threshold. The successive estimates correspond to HPAC predictions done at the 0.50, 0.80, 0.85, 0.90, 0.95, and finally, the 0.999 levels. Based on Figure 5, choosing a different probabilistic output can be seen as simply turning a "knob," that adjusts the tradeoff between false positive and false negative. One can imagine a model user considering the tradeoffs described by this MOE in selecting the output for his or her specific application.

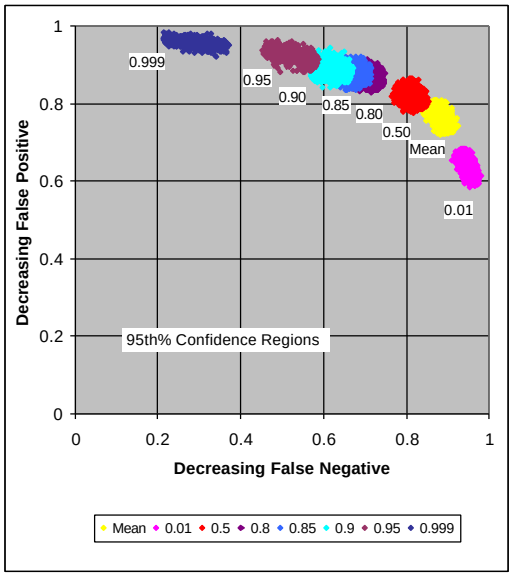


Figure 5 MOE Estimates for HPAC Probabilistic Predictions of 51 Prairie Grass trials at a Threshold of 60 mg-sec/m^3.

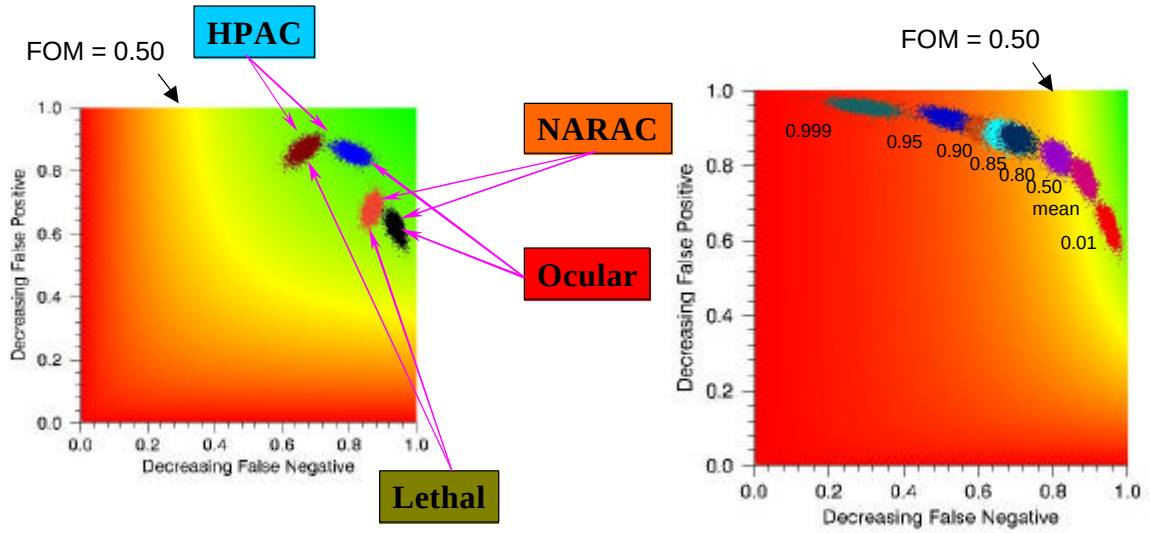
To make the above MOE truly user-oriented entails an additional step that *necessarily* requires user input. The "user" needs to make a decision that trades-off the false positive and false negative fractions in a way that he/she is willing to tolerate. Thus, the user needs to "color" the two-dimensional MOE space into acceptable and unacceptable regions. Figure 6 demonstrates this concept of "coloring" the user space. We take the function

$$User\ Figure of\ Merit\ (FOM) = \frac{A_{OV}}{\left(A_{OV} + C_{FN}A_{FN} + C_{FP}A_{FP}\right)} \qquad (2)$$

to represent a possible user trade-off between false positive and false negative regions. In this equation, C_{FN} and C_{FP} are user coefficients that weight the relative importance of false positive and false negative fractions.[4] The coefficient values chosen could be related to the specific model application being examined and the particular user's risk tolerance. Colors in Figure 6 correspond to isolines of the above FOM in 2D MOE space with the "bright" yellow line corresponding to a value of 0.5.

To produce Figure 6a, we used two notional models for the lethality/effects filter. We considered a model consistent with 50% of a population having ocular effects (threshold effects) and a probit model corresponding to 50% of a population dying from exposure to a chemical nerve agent. For Figure 6a, the coloring scheme is based on the user FOM described above with $C_{FN} = C_{FP} = 0.5$. MOE estimates for HPAC probabilistic predictions (as in Figure 5) are shown in Figure 6b overlaid on a coloring scheme that is based on the user FOM with $C_{FN} = 5$ and $C_{FP} = 0.5$.

4 A special case of this user figure of merit, with $C_{FN} = C_{FP} = 1$, has been previously considered [Mosca, 1998]



a) **b)**

Figure 6 Overlays of MOE Estimates on User Coloring: a) HPAC and NARAC Predictions (Based on Ocular and Lethal Effects) with User FOM Coloring ($C_{FN} = C_{FP} = 0.5$), b) HPAC Probabilistic Predictions (Based on 60 mg-sec/m^3 Threshold) with User FOM Coloring ($C_{FN} = 5$, $C_{FP} = 0.5$).

4 Conclusions

The two-dimensional measure of effectiveness, with false negative and false positive fractions considered orthogonal, consistently resolves important model performance features. Statistically significant resolution was seen between models and, for a given model, across conditions (for example, arc range and stability category grouping). Although not described in this extended abstract, we also found that changes to certain model input features could also be easily discerned with this MOE. For instance, as excursions, we examined model predictions that considered surface deposition of SO_2 and changes to the model dispersion parameterizations.

Also not described in this extended abstract, we computed several standard statistical measures that have previously been used to evaluate atmospheric transport and dispersion models. We found that the MOEs were consistent with these standard measures and, in some cases, may represent an improvement in terms of user interpretation [Warner, 2001].

We found that a lethality/effects filter can be used to compute MOE values, and can be characterized as a "tunable dial" that can relate the goodness of a prediction for agents of greatly varying toxicity. This feature may make this methodology of particular value with respect to user accreditation. For instance, the specific application and user will dictate the agent type and effects of interest; therefore defining how the lethality/effects dial should be tuned.

Finally, a quantitative description of a user's risk tolerance, within the context of the two-dimensional MOE space, has been presented. With such a description, decisions about the acceptability of a given model can be made for specific applications.

Acknowledgements

We would like to acknowledge the contributions of Allan Reiter (Defense Threat Reduction Agency – DTRA), Leon Wittwer (DTRA), Scott Bradley (Logicon), George Bieberbach (Logicon), Gayle Sugiyama (Lawrence Livermore National Laboratory – LLNL), John S. Nasstrom (LLNL), Kevin T. Foster (LLNL), and David Larson (LLNL). This work was sponsored by DTRA.

References

1. Barad, M.L. (Editor), "Project Prairie Grass, A Field Program in Diffusion," Geophysical Research Papers No. 59, Volumes I and II, DTIC #AD-152572/AFCRC-TR-58-235(I), Air Geophysical Laboratory, Hanscom Air Force Base, MA, 1958.
2. The meteorological stability category assignments (1 through 7) for the *Prairie Grass* field trials were based on a 1998 study. Irwin, J.S. and Rosu, M-R., "Comments on a Draft Practice for Statistical Evaluation of Atmospheric Dispersion Models," *Proceedings of the 10th Joint Conference on the Applications of Air Pollution Meteorology.* American Meteorological Society, Boston, pp. 6-10, 1998.
3. Mosca, S., Graziani, G., Klug, W., Bellasio, R., and Bianconi, R., "A Statistical Methodology for the Evaluation of Long-Range Dispersion Models: An Application to the ETEX Exercise," *Atmos. Environ.*, Vol. 32, No. 24, 4307-4324, 1998.
4. Warner, S., Platt, N., Heagy, J.F., Bradley, S., Bieberbach, G., Sugiyama, G., Nasstrom, J.S., Foster, K. T., and Larson, D., "User-Oriented Measures of Effectiveness for the Evaluation of Transport and Dispersion Models," Institute for Defense Analyses Paper P-3554, January 2001. This paper can be obtained electronically or on CD via an e-mail request to Steve Warner at swarner@ida.org.