DISCUSSIONS AND SUGGESTIONS ON USING DIFFERENT POINT-TO-POINT PROTOCOLS FOR TRANSPORT AND DISPERSION MODEL EVALUATIONS

Nathan Platt, Steve Warner, James F. Heagy, and Jeffry T. Urban

Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, USA

Abstract: Historically, assessments of T&D models have involved comparisons to field trial data using quantities derived from observations – maximum concentrations, plume widths, or integrated concentrations over arcs at various downwind ranges. Recently several methodologies that compare observations and predictions paired in space and time have been developed. Two different protocols designed to deal with low observed or predicted values have emerged. One of these protocols requires that both the observation *and* the prediction must exceed a specified threshold before being considered in the comparison. The other protocol requires that either the observation *or* the prediction must exceed a specified threshold before being considered in the comparison. This presentation compares the potential effects of applying these two protocols to T&D model assessments.

Key words: *Atmospheric transport and dispersion, point-to-point protocol, model evaluations.*

1. INTRODUCTION

There is a continuing need to assess the accuracy of hazardous material transport and dispersion (T&D) models. Historically, assessments of T&D models have involved comparisons to field trial data using quantities derived from observations – maximum concentrations, plume widths, or integrated concentrations over arcs at various downwind ranges (Hanna, et. al, 1993). A significant limitation of using these derived quantities is that they, by construction, ignore directional (or point-to-point spatio-temporal) divergence between the observed and predicted plumes.

Recent improvements to T&D models and increased accuracy and availability of meteorological observations and numerical weather prediction tools have led to the proposed expansion in the use of T&D models as aids for "real time" emergency response. As a result, several methodologies that compare observations and predictions paired in space and time have been developed (Mosca et. al, 1998; Warner et. al, 2004a; Warner et. al, 2004b). These include two-dimensional user oriented MOEs; additionally, more traditional metrics (such as fractional bias (FB), normalized absolute difference (NAD), normalized mean square error (NMSE), geometric mean (MG), geometric variance (VG), and factor of 2 (FAC 2)) have been expanded to perform these point-to-point comparisons (Warner et. al, 2004b; 2004c).

Because of the inherent sensitivity of geometric and ratio metrics, coupled with instrumentation uncertainty for low observed measurements, special care needs to be taken when either predicted or observed values are near zero. Two different protocols designed to deal with low observed or predicted values have emerged. One of these protocols requires that both the observation *and* the prediction must exceed a specified threshold (typically based on the experimental limit of quantification (LOQ)) before being considered in the comparison – we refer to this as an "intersection" protocol (Chang et. al, 2005; Hanna et. al, 2008). The other protocol requires that either the observation *or* the prediction must exceed a specified threshold before being considered in the comparison – we refer to this as a "union" protocol (Warner et. al, 2004a; 2004b; 2004c). Figure 1 demonstrates differences between these two protocols based on notional observations and predictions modelled as Gaussians with different means and sigmas. Here, the observations are notionally represented by a dashed yellow-brown Gaussian ($\sigma = 1$, $\mu = 1$), the predictions are notionally represented by a solid dark-purple Gaussian ($\sigma = 2$, $\mu = -1$), and the critical threshold is set to 0.05. Two dotted vertical lines denote centrelines of the Gaussians which also correspond to the means. The solid thick blue line shows the "intersection" domain where observations and predictions are compared, while the solid thick red line shows the "union" domain where observations and prediction are compared.

This paper provides an initial comparison of the potential effects of applying these two protocols to T&D model assessments. First, idealized examples based on assumed plume profiles – either square wave or Gaussian – at a specified range are examined as a function of the "miss distance" between the centrelines of the assumed shapes (referred to as "difference in centrelines" in Figure 1). Second, *Project Prairie Grass* field trial data is compared with surrogate predictions as a function of the "miss angle" to demonstrate the effects associated with these two protocols.

EXAMPLE: SQUARE-WAVES

We begin our analysis of the effects of "union" and "intersection" protocols on T&D model evaluations with a pair of simple notional examples. We start with an assumption that both observations and predictions are modelled by a square wave with different amplitudes and widths as depicted in Figure 2. As in Figure 1, observations are represented by a dashed yellow-brown line (square wave with amplitude 1.0 and width 2.5) and predictions are represented by a solid dark-purple line (square wave with amplitude 0.5 and width 5), and the critical threshold is set to 0.05. This situation could be thought of as representing a field experiment consisting of lines of measurement samplers some distance from a continuous line source release. Our analysis examines changes in standard statistical measures (FB, FAC 2, NMSE, NAD, MG and VG) as a function of differences in centrelines of these square-waves.

It should be noted that, for the "intersection" evaluation protocol, differences between observations and predictions are constant and do not change when differences in centrelines are varied, thus yielding constant values for the statistical measures. Figure 3 demonstrates changes in the statistical measures as a function of a centreline difference of the two square waves. The horizontal centreline differences were discretized at 0.01 units. The blue line (with diamonds plotted at every 50th point) denotes statistical measures obtained using the "intersection" point-to-point

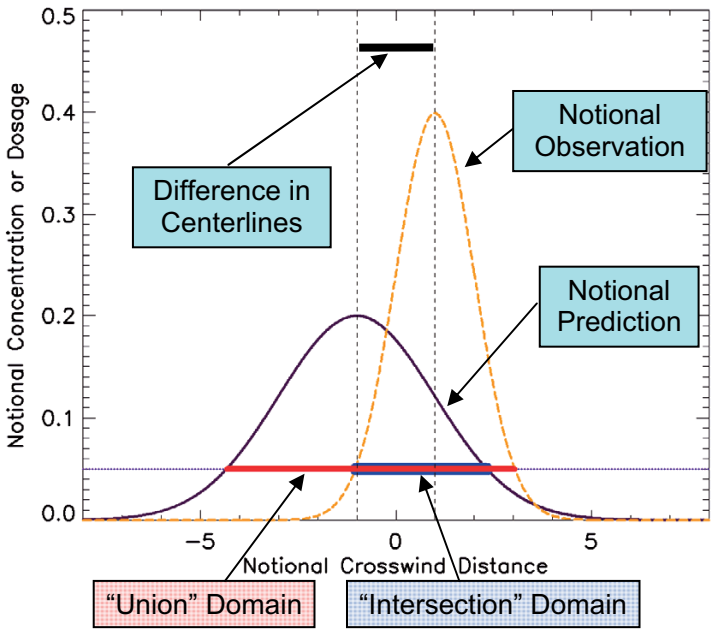

Figure 1. Demonstration of the "union" and "intersection" domains where evaluations take place based on notional observations and predictions modelled as Gaussian.

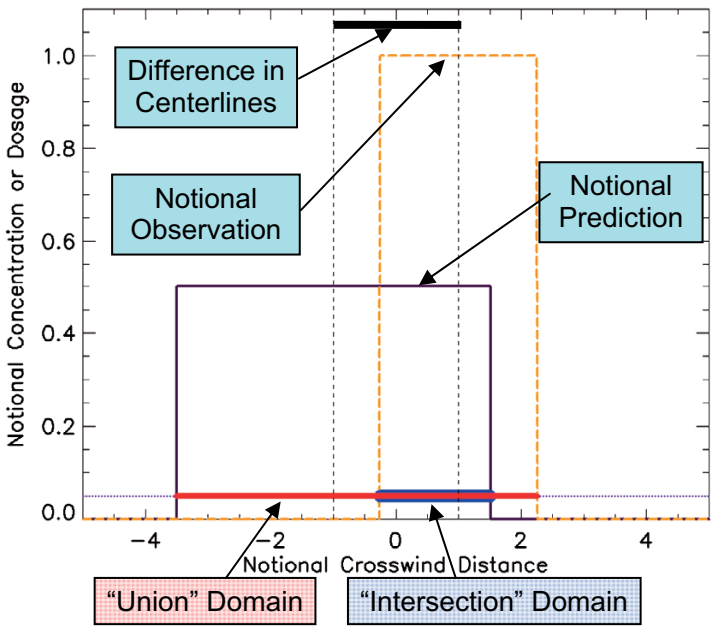

Figure 2. Demonstration of the "union" and "intersection" domains where evaluations take place based on notional observations and predictions modelled as square waves.

protocol while the red line (with "×" plotted every 50th point) denotes statistical measures obtained using the "union" point-to-point protocol. As expected, statistical quantities calculated using the "intersection" protocol do not vary as

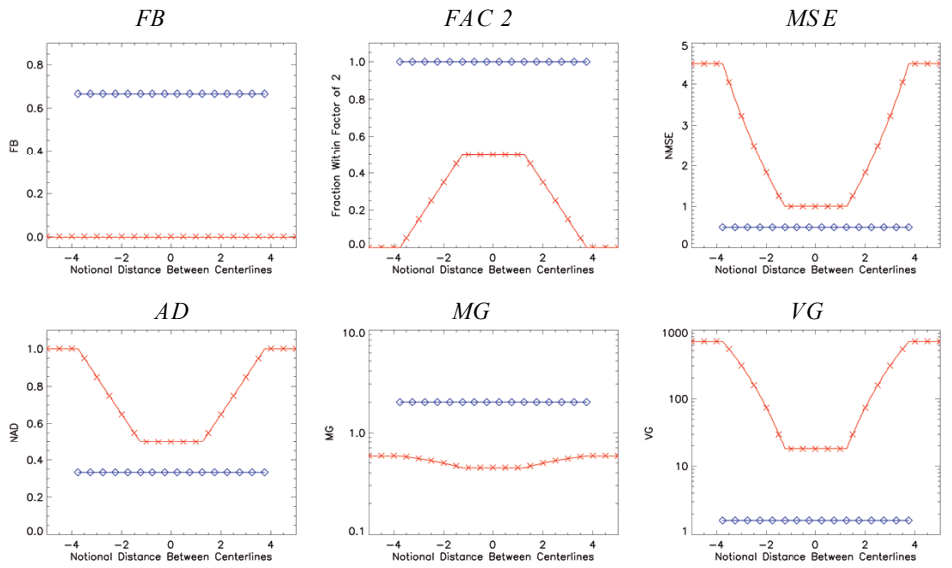

Figure 3. Standard statistical measures calculated as a function of centreline differences for the square waves. Blue line (with diamonds) quantities are calculated using the "intersection" protocol while red line (with "×") quantities are calculated using the "union" protocol.

a function of centreline difference, while most statistical quantities calculated using the "union" protocol show significant and monotonic degradation as the absolute centreline difference between the square waves is increased.

3. EXAMPLE: GAUSSIANS

We further illustrate expected potential differences in using "intersection" and "union" point-to-point protocols for evaluating T&D models, by considering predictions and observations notionally modelled as Gaussians as depicted in Figure 1. This situation could be thought of as representing a field experiment consisting of lines or arcs of measurement samplers some distance from a instantaneous or continuous point source release.

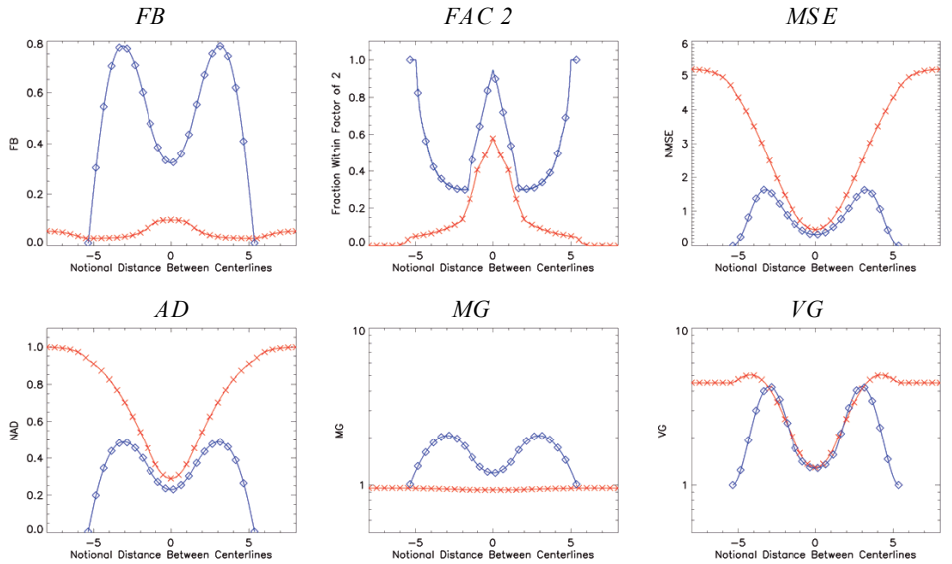

Figure 4. Standard statistical measures calculated as a function of centreline differences for the Gaussians. Blue line (with diamonds) quantities are calculated using the "intersection" protocol while red line (with "×") quantities are calculated using the "union" protocol.

It should be noted that the majority of field trials available for evaluation of T&D models fall into this category. Moreover, a large portion of atmospheric T&D models in use today are based on Gaussian parameterization of the plume, thus a predicted "plume" passing through a line or arc of samplers is expected to resemble a Gaussian shape. In fact, for a number of derived quantities used in T&D model evaluations (such as "calculated" observed plume

sigma at a particular distance from the release, etc.) it is assumed that the observed plume passing through a line or arc of samplers should resemble a Gaussian in some form. Figure 4 demonstrates changes in the statistical measures as a function of a centreline difference of the two Gaussians. As before, the horizontal centreline differences were discretized at 0.01 units. The blue line (with diamonds plotted at every 50th point) denotes statistical measures obtained using the "intersection" point-to-point protocol, while the red line (with "×" plotted every 50th point) denotes statistical measures obtained using the "union" point-to-point protocol. As was the case with square wave example, most of statistical measures calculated using the "union" point-to-point protocol show almost monotonic degradation as absolute centreline distance between the Gaussians is increased. It should be noted that the unexpected variations in FB and slight improvements in VG observed for large absolute centreline differences is due to small marginal contributions from the Gaussians that are below critical threshold for the "union" protocol. There is a significantly different behaviour observed in statistical measures as absolute centreline difference between the Gaussians in increased for the "intersection" protocol. For all statistical measures considered, the initial degradation in the performance for moderate differences in centrelines is followed by a sudden turn around and eventually very significant improvements in metrics for large differences in centrelines. In fact, when the two Gaussians just about miss each other at the critical threshold level, the comparison performance metric yields almost *perfect agreement* between observations and predictions.

4. PROJECT PRAIRIE GRASS FIELD TRIALS

While notional examples demonstrating potential differences in evaluation metrics resulting from the use of two point-to-point protocols seem to indicate that the "union" protocol is much more robust than the "intersection" protocol, it is not at all clear how these differences would manifest when actual field trial data with measurement noise and model predictions are compared. In this section we demonstrate the potential effects by comparing *Prairie Grass* field trial data (Barad, 1958) and predictions that we generated during an earlier study (Warner et. al., 2004a) using the two protocols. Figure 5 plots observations and HPAC predictions for *Prairie Grass* trial 42 at different arcs. The light-brown dashed line depicts observations (in mgsm^{-3}) at different samplers while the solid dark line depicts HPAC predictions. We simulate "shifted" HPAC predictions by shifting sampler indices to the left or to the right by a prescribed amount. Since there were 90 samplers at each of the closest four arcs and 180 samplers on an 800-meter arc, a minimum resolution of 2 degrees could be achieved during these shifts. The dotted dark lines in Figure 5 demonstrate the "shifted" predictions by ±30 degrees. It should be noted that this "mechanical" shift of predictions over sampler indices is theoretically equivalent to a "biased" shift of the winds used to create the HPAC predictions. The critical threshold of interest is set to 60 mg-s-m^{-3} which approximately corresponds to the sampler's LOQ.

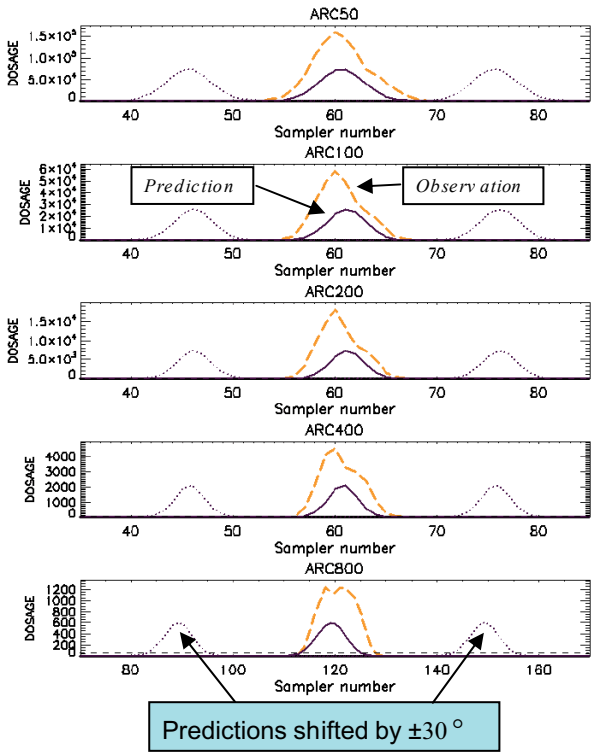

Figure 5. Plots of observations (dashed light-brown line) and HPAC predictions (solid dark line) for *Prairie Grass* field trial 42 for different arcs. Dotted dark lines represent HPAC predictions shifted by ±30 degrees. Dosage units are in mgsm^{-3}.

Figure 6 shows the differences in statistical performance metrics using the two protocols and resulting from shifting predictions to the right or to the left of the observations. Qualitatively, these plots are very similar to plots shown in Figure 4 when comparisons of the point-to-point protocols are done using two Gaussians sliding past each other. More fluctuations in the performance metrics are observed for different values of the "shift" due to fluctuations in observed values at the samplers (especially FAC 2), but the trends are similar. Most of the statistical metrics calculated using the "union" point-to-point protocol degrade gracefully as the absolute amount of the shift is increased, while statistical metrics calculated using the "intersection" point-to-point protocol seem to turn around at approximately a ±20 degree shift – below 20 degrees these metrics indicate degradation in the comparison of predictions with observations, while above 20 degrees these metrics imply significant improvements in the comparison of predictions with observations. We note that this type of behaviour for these two point-to-point protocols was typical for *Prairie Grass* field trial experiments – i.e., a large fraction of the trials showed significant improvements in most statistical metrics as predictions moderately to heavily shifted using the "intersection" protocol, while showing, as expected, significant degradation using the "union" protocol.

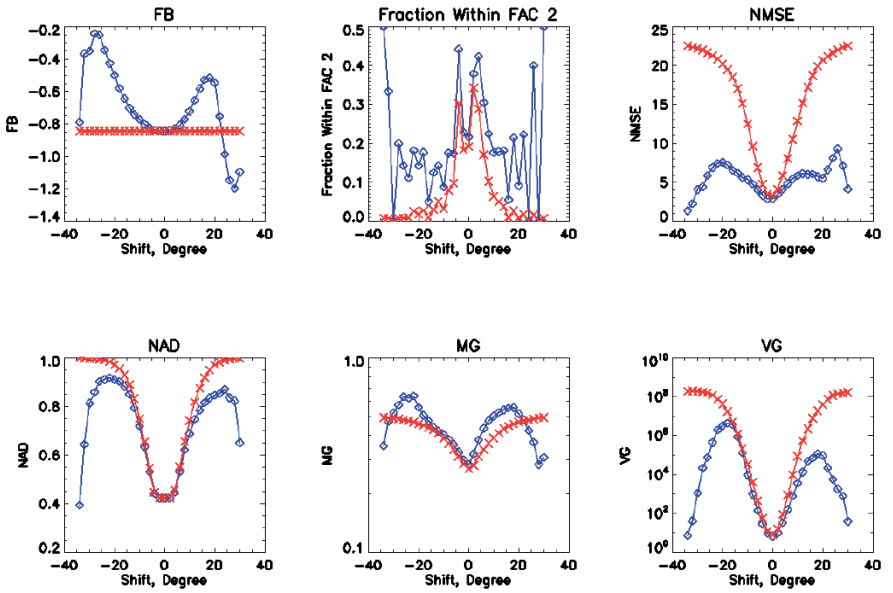

Figure 6. Standard statistical measures calculated as a function of HPAC predictions shift (in degrees) for *Prairie Grass* field trial 42. Blue line (with diamonds) quantities are calculated using the "intersection" protocol while red line (with "×") quantities are calculated using the "union" protocol.

5. CONCLUSIONS

Two different point-to-point protocols designed to deal with low observed or predicted values have recently emerged. One of these protocols requires that both the observation *and* the prediction must exceed a specified threshold (typically based on the experimental LOQ) before being considered in the comparison – referred here as an "intersection" protocol. The other protocol requires that either the observation *or* the prediction must exceed a specified threshold before being considered in the comparison – referred here as a "union" protocol. Two idealized examples (square wave and Gaussian) and *Prairie Grass* field trial data are used to examine the potential effects of using these two protocols for T&D model evaluations as the distance between observed and predicted "plume" centrelines is varied. While the "union" point-to-point protocol robustly captures expected degradation in statistical quantities as a function of increased distance between "plume" centrelines, the "intersection" point-to-point protocol fails to take into account both false negatives and false positives with respect to the threshold of interest. Thus using the "intersection" protocol yields statistical measures typically showing considerable improvement (or no change) as "plume" centerline difference distance increases or, in other words, the "intersection" protocol typically indicate significantly improved agreement between the predictions and observations when they almost miss each other.

We note that both the square wave and Gaussian examples used here are somewhat idealized – they did not take into account effects that either background (adding or subtracting) or measurement noise (additive or multiplicative) could have on the comparison of these two protocols. Future work will extend this analysis to include both of these affects.

Acknowledgments: *This effort was supported by the Defense Threat Reduction Agency with Dr. John Hannan as the project monitor and the Central Research Program of the Institute for Defense Analyses. The views expressed in this paper are solely those of the authors.*

REFERENCES

Barad, M.L., Ed., 1958: Project Prairie Grass, a field program in diffusion. Vols. I and II, *Geophysical Res. Papers* **59**, Rep. AFCRC-TR-58-235, 439 pp.

Chang, J.C., S.R. Hanna, Z. Boybeyi and P. Franzese, 2005: Use of Salt Lake City URBAN 2000 field data to evaluate the Urban Hazard Prediction Assessment Capability (HPAC) dispersion model. *J. Appl. Meteor.*, **44**, 485-501.

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

Hanna, S.R., E. Baja, J. Flaherty and K.J. Allwine, 2008: Use of tracer data from the Madison Square Garden 2005 field experiment to test a simple urban dispersion model, The 88th Annual American Meteorological Society meeting in New Orleans, 20-24 January 2008, online proceedings, 22 pp.

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

Warner, S., N. Platt and J.F. Heagy, 2004a: User-oriented two-dimensional measures of effectiveness for the evaluation of transport and dispersion models. *J. Appl. Meteor.*, **43**, 58-73.

Warner, S., N. Platt and J.F. Heagy, 2004b: Application of user-oriented measures of effectiveness to transport and dispersion model predictions of the European tracer experiment. *Atmos. Environ.*, **38**, 6789-6801

Warner, S., N. Platt and J.F. Heagy, 2004c: Comparisons of transport and dispersion model predictions of the URBAN 2003 field experiment. *J. Appl. Meteor.*, **43**, 829-846.