Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Science and Technology for Chem-Bio Information Systems (S&T CBIS) Conference, 25-28 October 2005, Albuquerque, New Mexico

Proposed Translation of Joint Effects Model (JEM) Accuracy Requirement Into a Measurable Acceptability Criterion

Steve Warner (swarner@ida.org)
Nathan Platt
James F. Heagy

Introduction and Outline

- **Introduction**
  - – Institute for Defense Analyses (IDA) is a non-profit research and development center, serving the Office of the Secretary of Defense, Defense Agencies, Unified Combatant Commands, and the Joint Staff.
  - – For this task, IDA provides independent technical analyses to support hazard prediction model evaluation efforts
    - » sponsors have included: DTRA, DATSD NCB, MDA, JEM PMO, IDA CRP/PD
- **Outline of Presentation**
  - – Joint Effects Model accuracy requirement
  - – Issues with the application of this requirement to evaluations
  - – Proposed solution
    - » user-oriented measure of effectiveness for hazard prediction models
    - » accuracy requirement interpreted as normalized absolute difference
  - – Example results
  - – Comments, caveats, summary

Joint Effects Model Accuracy Requirement

- **Mission Need**
  - "improved capability to portray chemical, biological, radiological, and nuclear (CBRN) effects in models and simulations…"
  - Joint MNS states, "…modeling and simulation does not adequately predict and track CBRN and Toxic Industrial Material (TIM) impact to support operational decisions and risk assessments."
- **Information Availability Requirement: Accuracy**
  - "JEM shall provide accuracy of ≥ 70%" threshold value (KPP)
    - » ≥ 85% objective
  - rationale states:"Accuracy ≥ 70% means that the error between the predicted and observed concentrations is no more than 30%"
- **Performance Requirement: Related to Accuracy**
  - JEM shall also predict hazard areas for CBRN and TIM events that reduce the likelihood of being Falsely Warned and Falsely Not Warned…"
  - rationale states: "For hazard prediction, operational risk management is a trade-off between the error of falsely warning personnel and the error of falsely not warning personnel."

***Operational Requirements Document for Joint Effects Model (JEM)*, 28 May 2004.**

Issues With the Application of the Accuracy Requirement to Evaluations

- **Issue 1: typically not achievable or even reasonable**
  - For state-of-the-art models, differences between observed and predicted average concentrations will typically be larger than 70%, e.g.,
    - » ≤ 40% within a *factor of 2* for the very short range MUST field experiment
    - » "accuracy" was 71% for the short-range, open field *Prairie Grass* field experiment which included "unrealistically" detailed meteorological inputs
    - » accuracies of 68%, 57% and 45% for the 1997 OLAD field trial (3 models)
    - » for longer ranges and more realistic input conditions, accuracies would be expected to be worse
  - One proposal has been to compare predictions and observations of the <u>maximum values only, regardless of their locations</u> and then require that the Fractional Bias (fraction over- or under-prediction) be ≤ 0.3

- **Issue 2: not what the user actually wants, needs, or requires**
  - In general, users are not interested in average or maximum concentrations
  - Rather, for most applications (planning & warning), users are interested in the locations and times at which a hazardous condition exists (or might exist)

Hazard region (hazard area) predictions are desired.

Proposed Solution

- Evaluate hazard prediction with a user-oriented measure of effectiveness (MOE) that allows for assessments of how well the prediction and observation of hazardous region “overlap” in time and space

This allows one to do the evaluation in the “space” that is of interest to the user.

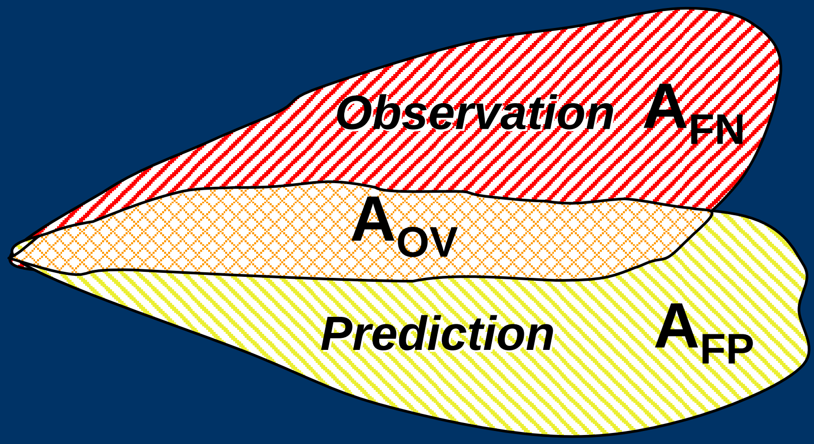
- Translate ORD accuracy requirement into an acceptability function – normalized absolute difference - that can be assessed based on the above MOE from field experiment observations and predictions

This allows the ORD accuracy requirement to be directly evaluated.

User-Oriented Two-Dimensional MOE Introduction

- **Fundamental feature of any comparison of model output to observations is over- and under-prediction**
  - False Negative (FN): hazard is observed but <u>not</u> predicted
  - False Positive (FP): hazard is predicted but <u>not</u> observed

These two "dimensions," FP and FN, define the space in which the user is most interested



A_{FN} = Region of False Negative

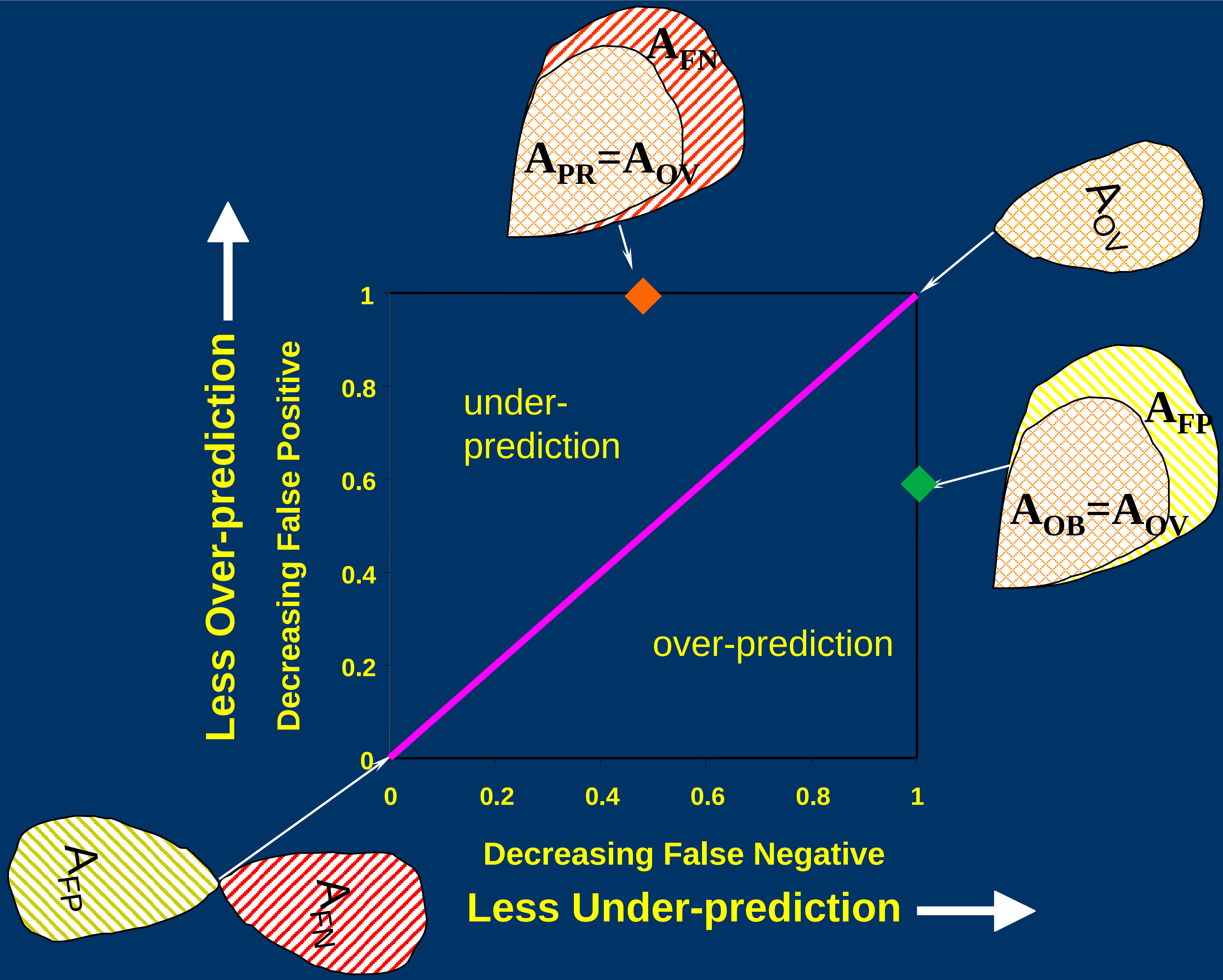
A_{OV} = Region of Overlap

A_{FP} = Region of False Positive

$A_{PR} = A_{OL} + A_{FP}$ = Region of Prediction

$A_{OB} = A_{OL} + A_{FN}$ = Region of Observation

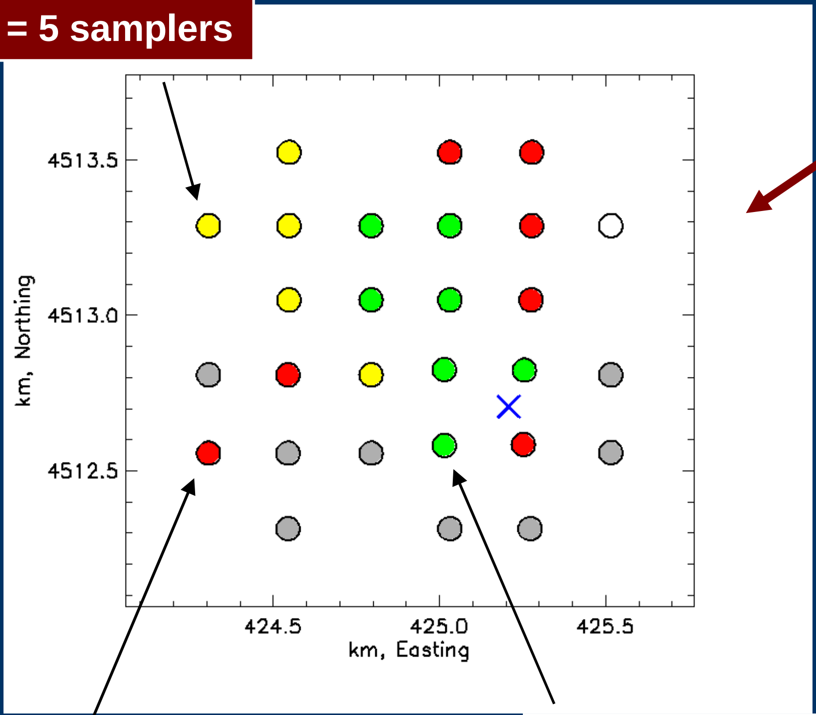
$$MOE = \left(\frac{A_{OV}}{A_{OB}}, \frac{A_{OV}}{A_{PR}}\right) = \left(1 - \frac{A_{FN}}{A_{OB}}, 1 - \frac{A_{FP}}{A_{PR}}\right)$$



Sample MOE Component Calculation: *Urban 2000,* Downtown Salt Lake City

Illustration of MOE threshold-based computation

A_{FP} = 5 samplers



A_{FN} = 7 samplers

A_{OV} = 7 samplers

for a 30 ppt concentration threshold

MOE = (7/14, 7/12) = (0.50, 0.58)

Several applications of this MOE have been published and include transformations of the MOE to account for:

(1) area interpolation,
(2) the underlying population distribution, and
(3) the expected human effects of notional agents.

Also, "scoring functions" have been developed, published, and applied to evaluate hazard prediction models with field experiment observations.

Accuracy ≡ Normalized Absolute Difference (NAD)

NAD ≤ 0.3 ⇒ Accuracy ≥ 70%

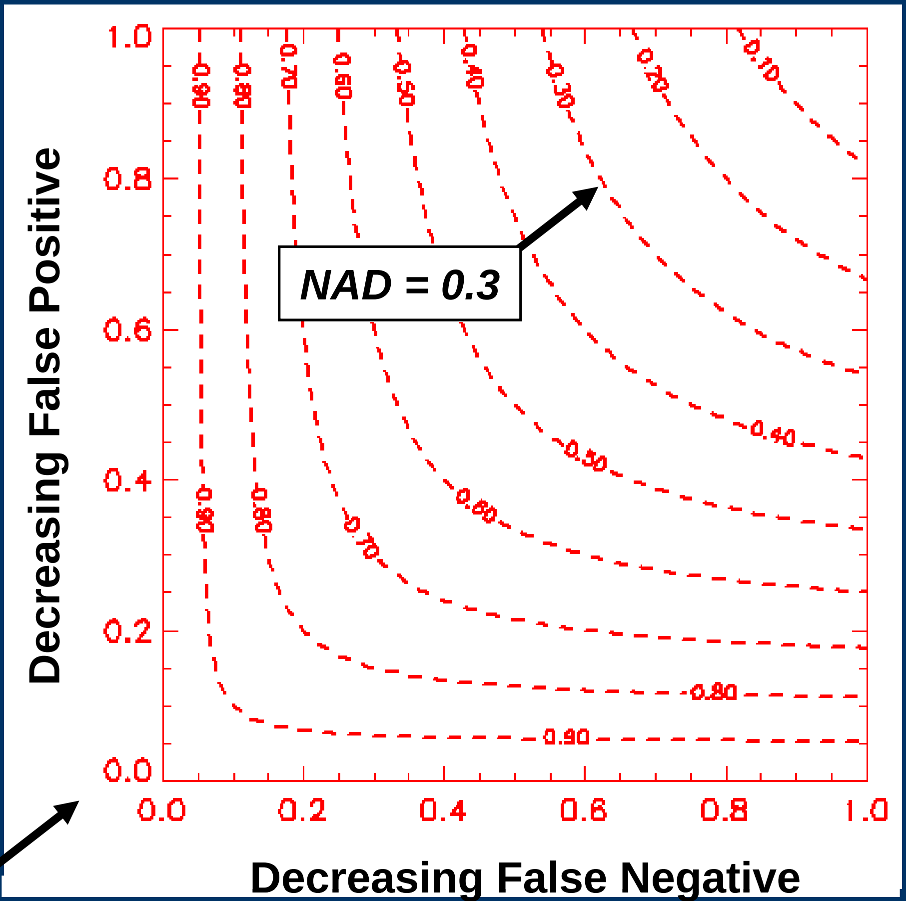
- **For average concentration, NAD has previously been developed as a scoring function in the 2D MOE space**

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

NAD = 0.0 ⇒ no scatter, perfect prediction

and, is related to the MOE(x,y) as

$$NAD = \frac{x + y - 2xy}{x + y}$$



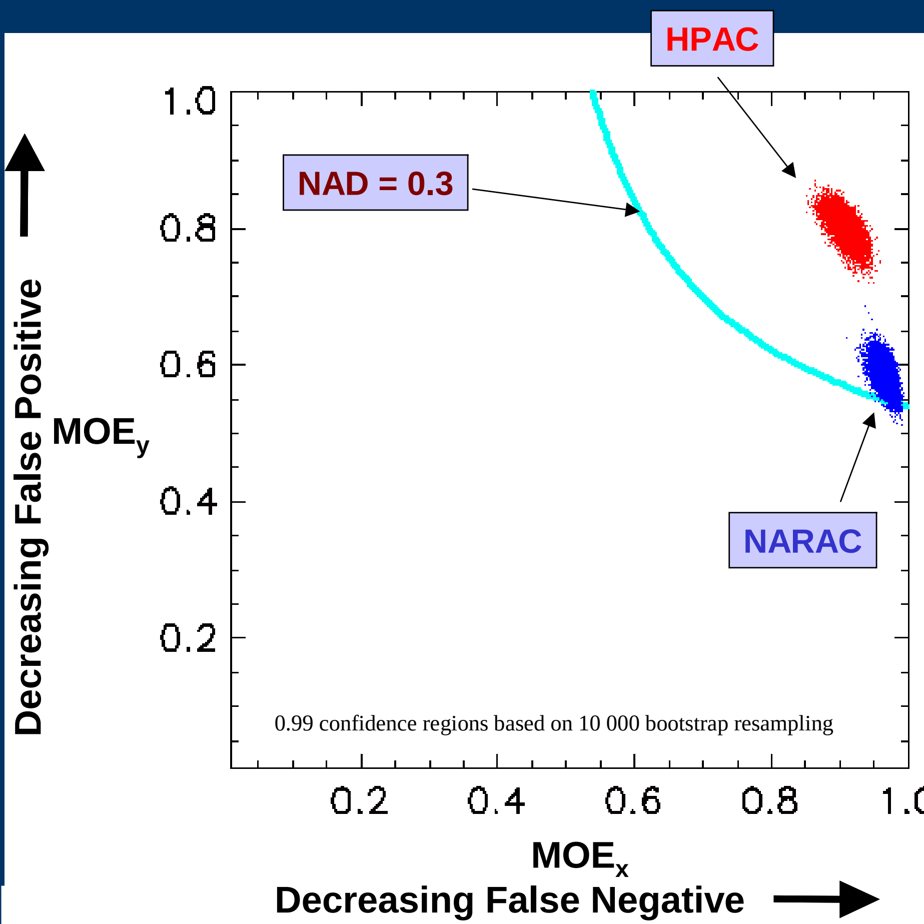
Isolines of NAD in the 2D MOE Space

But User is Interested in the Accuracy of Predicting Hazard Regions, Not Average Concentrations

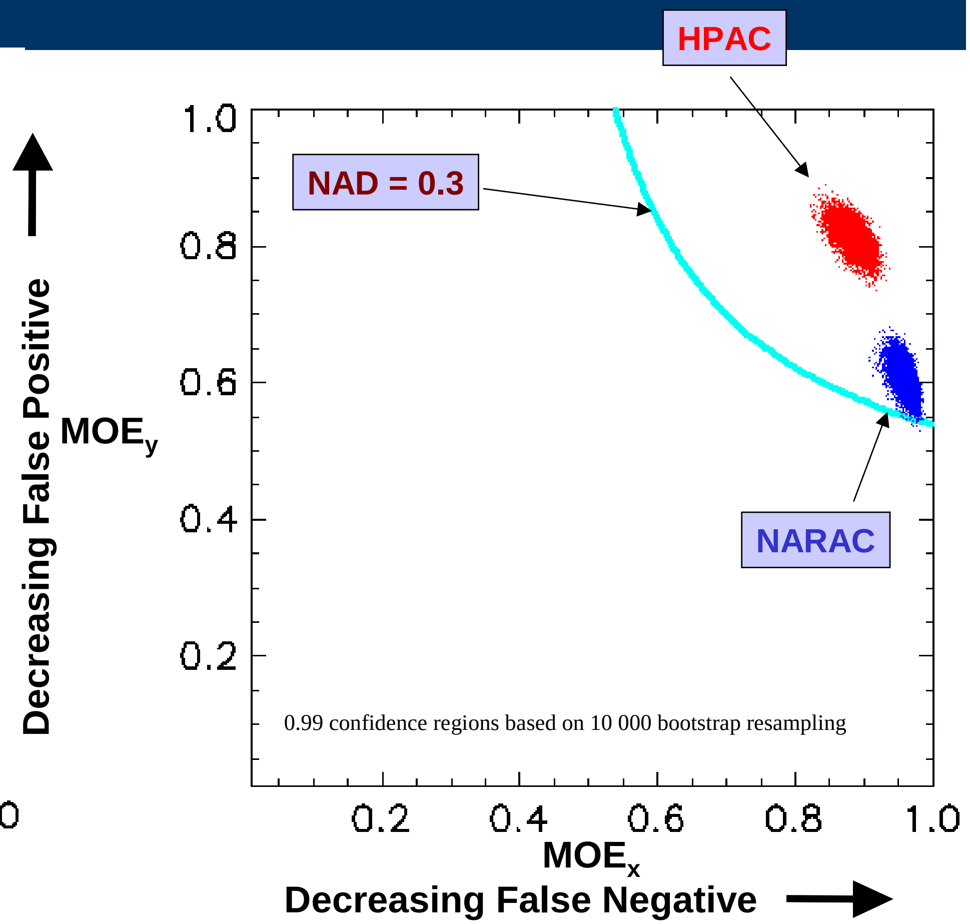
- **User's typical interest is in locations (times) where a hazard threshold is exceeded – a "hazard area"**
- **Fortunately, NAD can also be computed based on a threshold just like the MOE**
- **For the previous *Urban 2000* example, NAD = 0.46 implying an "accuracy" of predicting the downtown locations of the low-level hazard of 54%**

Example Application 1a: HPAC, NARAC, and the *Prairie Grass* Field Experiment

Threshold (T) = 15 mg sec m^{-3} ≈ 5 × sampler limit



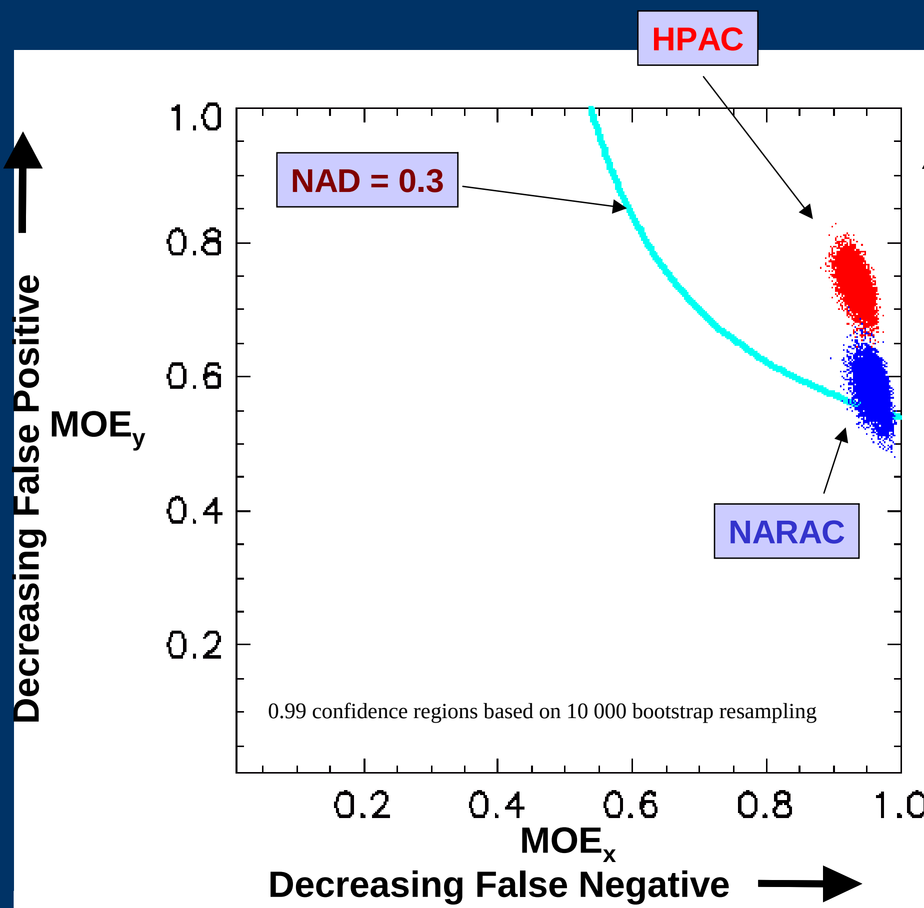
T = 60 mg sec m^{-3} ≈ 20 × sampler limit



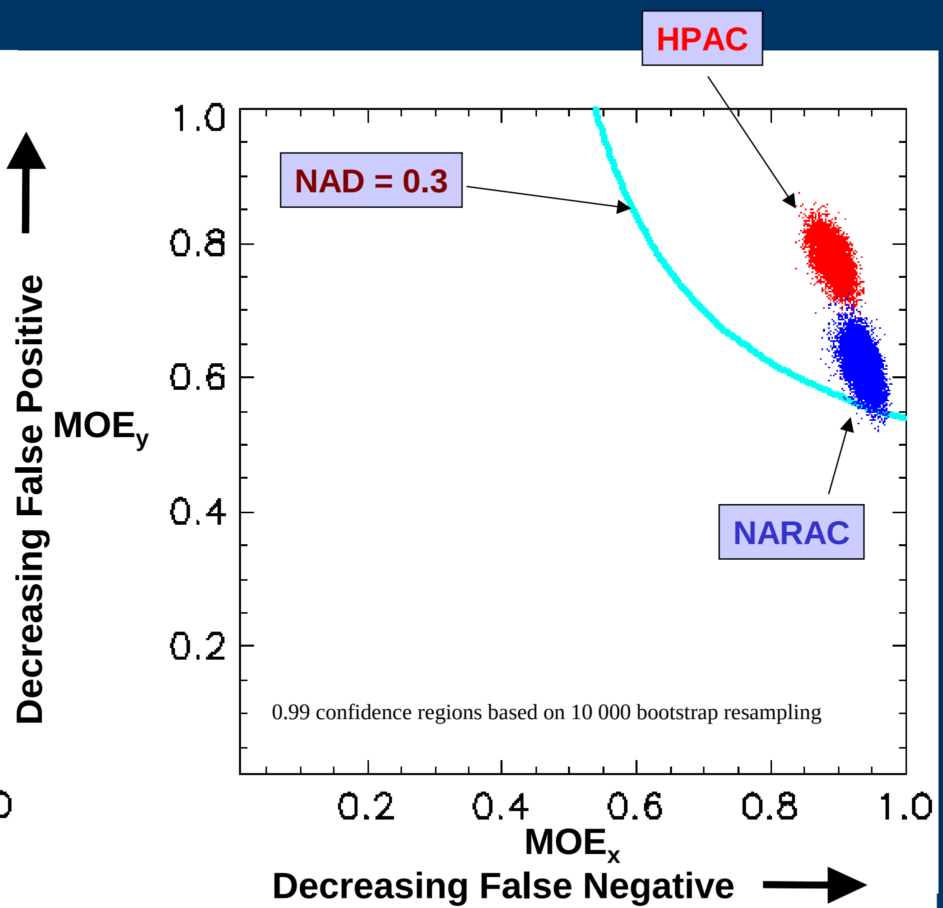
based on 51 releases

Example Application 1b: HPAC, NARAC, and the *Prairie Grass* Field Experiment After Area Interpolation – i.e., MOE components are based on real areas (km^2)

T = 15 mg sec m^{-3}
≈ 5 × sampler limit



T = 60 mg sec m^{-3}
≈ 20 × sampler limit

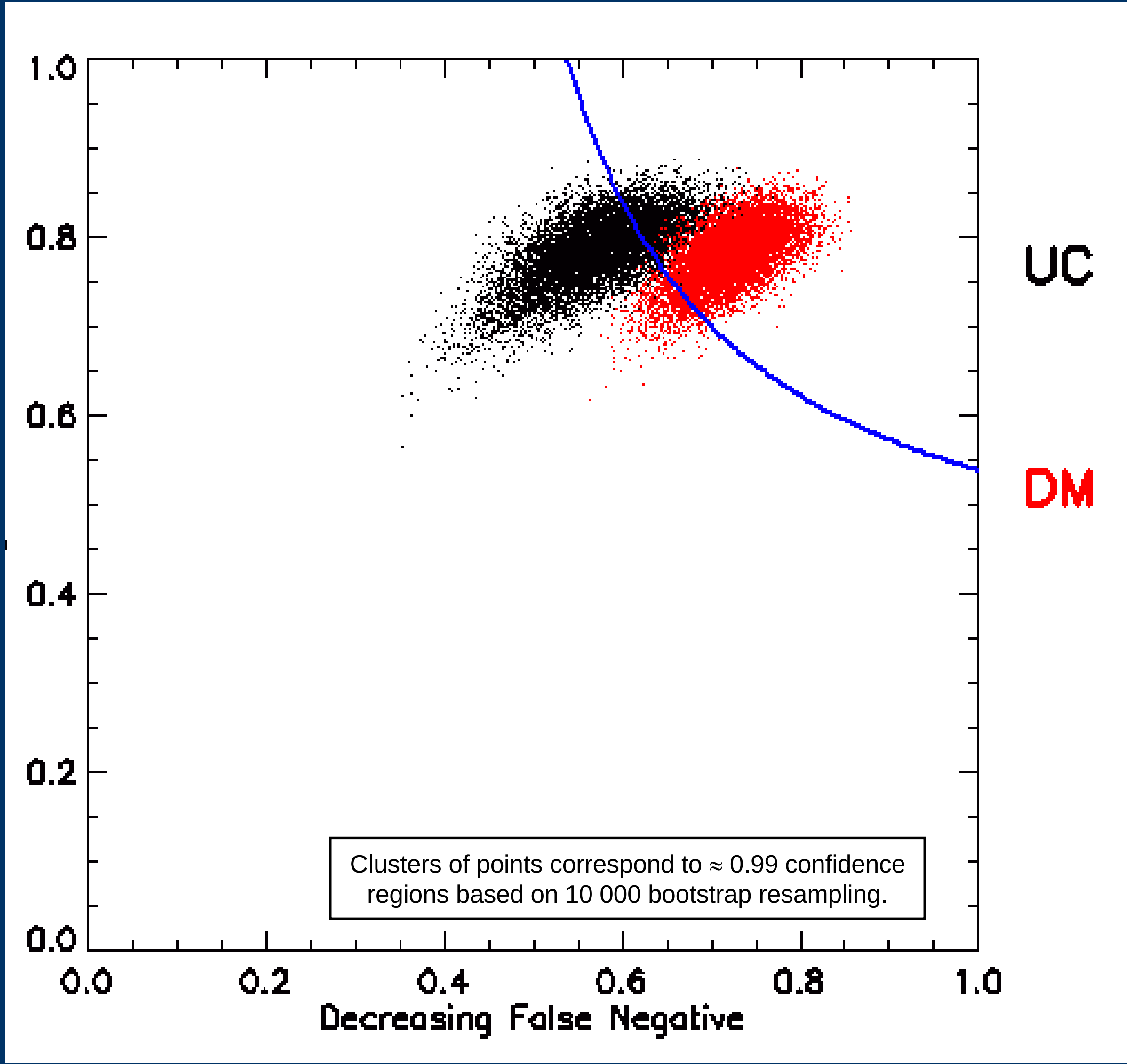


based on 51 releases

Example Application 2: Urban HPAC Baseline Mode (Urban Canopy – UC) , Urban Dispersion Model Mode (UDM), and the *Urban 2000* (Salt Lake City) Field Experiment

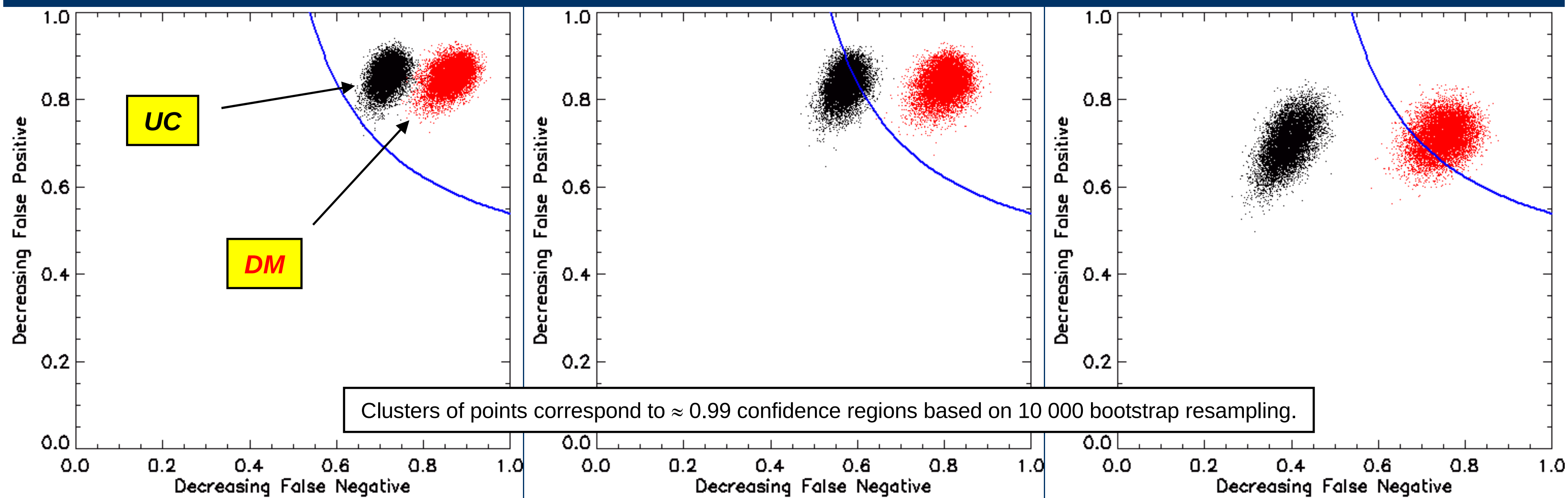
T = 30 ppt
≈10 × background
≈ 2 × MLOD

Based on HPAC (UC and DM) predictions of *Urban 200* with the "Raging Waters" upwind profile meteorological input option and 18 releases



Example Application 3: Urban HPAC Baseline Mode (Urban Canopy – UC) , Urban Dispersion Model Mode (UDM), and the *MUST* Field Experiment

T = 0.01 ppm *≈ 100 × sampler limit*	*T = 0.10 ppm* *≈ 1,000 × sampler limit*	*T = 1.00 ppm* *≈ 10,000 × sampler limit*



Based on HPAC (UC and DM) predictions of *MUST* with the "SONICs" meteorological input option and 37 releases

MUST = Mock Urban Setting Test, Dugway Proving Ground 2001

Additional Comments and Caveats

- **As would be true for any model evaluation technique, interpretation of results must be carefully considered**
  - Field trial experiments typically include high quality source term and meteorological information that will not necessarily be available for actual applications
    - » in a sense, such evaluations might be thought of as the model at its "best"
  - It is important that supporting analyses be conducted as part of any evaluation. For example, MOE values can be computed
    - » at a few different low threshold values (typically we vary by factors of about 100),
    - » as a function of downwind distance and time after the release
    - » using sampler weighting or interpolation procedures to assess results in the context of actual area sizes (where feasible), and
    - » by considering notional scenarios (that match up well with the field experiment ,where plausible) in order to consider the effects of actual agents – "effects filtering".
- **Interpretation of the JEM accuracy requirement in terms of the MOE and NAD satisfies the need for assessing this requirement for acquisition / program management decisions (e.g., when to shift resources from improving one aspect of the system to another) in the context of actual user needs**
  - Evaluations based on "armax", crosswind integrated concentration, and average concentration do not allow for this user context
  - Other scoring functions (for the 2D MOE space), have been developed that can be used for other purposes (e.g., doctrinal development)
    - » an important family of scoring functions has been developed that allows the user to weight the risks (trade-offs between false positive and false negative fractions) as appropriate to his or her application and mission

Summary

- **Issues with the application of the JEM accuracy requirement to future evaluations have been identified**
  - Not likely to be achievable or even reasonable
  - Not what the user wants, needs, or requires
    - » user's typical interest is in locations (times) where a hazard threshold is exceeded – a "hazard area"

- **Proposed solution**
  - Evaluate JEM hazard predictions user-oriented 2D MOE and actual field experiment observations
  - Assess MOE based on hazard regions, i.e., exceeding a threshold
    - » how well does the model predict the locations where a low-level threshold is exceeded
  - Use normalized absolute difference as a straightforward measure of accuracy that is directly related to the requirement

- **Example studies suggest:**
  - HPAC, and hence JEM(HPAC), can reasonably be expected to pass this accuracy threshold for simple (no complex terrain), short-range, open field experiments (e.g., *Prairie Grass*)
  - More complex situations, for example, urban environments, may require additional model features (recall baseline HPAC versus UDM mode comparisons for Urban 2000 and MUST)

User involvement in JEM evaluation remains crucial for success.