Developing a Common Framework for Model Evaluation

Panel Discussion
7th Transport and Diffusion Conference
George Mason University
19 June 2003

Dr. Priscilla A. Glasow
The MITRE Corporation

MITRE

Topics

- Joint Effects Model evaluation approach
- Model evaluation in the Department of Defense
  - Methods
  - Evaluation method deficiencies
  - Lessons learned in building a common evaluation framework

MITRE

JEM Model and Evaluation Approach

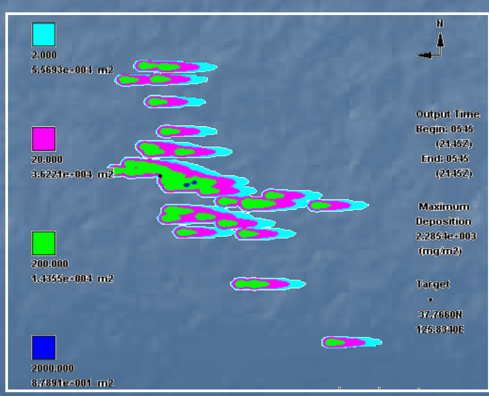
CBRN Hazard Model Components

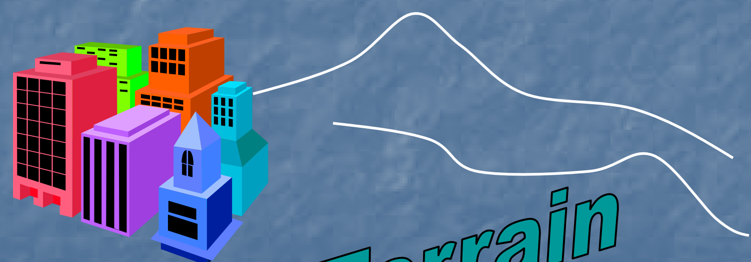
Source

Meteorology

Transport & Dispersion of Matter & Energy

Decay, Deposition Processes





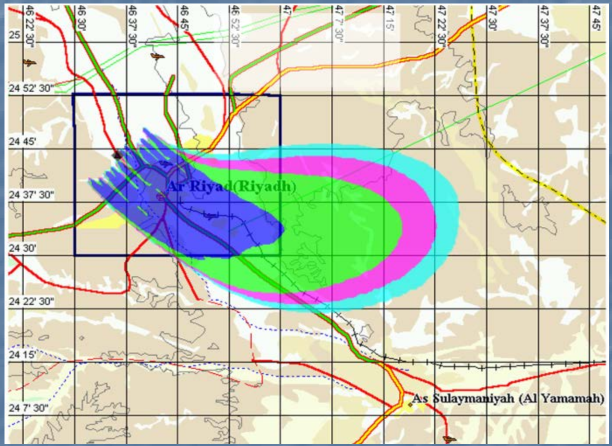
Terrain

Human Response

JEM

System Description

- Models the transport and diffusion of hazardous materials in near real-time
  - CBRN Hazards
  - Toxic Industrial Chemicals / Materials
- Modeling approach includes
  - Weather
  - Terrain / vegetation / marine environment
  - Interaction with materials



- Depicts the hazard area, concentrations, lethality, etc.
  - Overlays on command and control system maps
- Will deploy on C4I systems and standalone
  - Where JWARN is not needed, but hazard prediction is needed

System Description (continued)

- Legacy systems being replaced
  - HPAC - a DTRA S&T effort
  - VLSTRACK - a NAVSEA S&T effort
  - D2PUFF - a USA S&T effort
- Improved Capabilities
  - One model, accredited for all uses currently supported by the 3 Interim Accredited DoD Hazard Prediction Models
    - Warfighter validated Graphical User Interface
    - Uses only Authoritative Data Sources
  - Integrated into Service Command and Control systems
  - Flexible & Extensible Open Architecture
    - Seamless weather data transfer from provider of choice
  - Planned program sustainment and Warfighter support
    - School-house and embedded training
    - Reach Back

JEM

Calculating Hazard Predictions

- **How might a hazard probability be calculated?**

$$P_{HAZ} = P_{HAZ/VULN} * P_{VULN/OCC} * P_{OCC}$$

P_{HAZ} = probability of the hazard
$P_{HAZ/VULN}$ = prob. of the hazard for a given vulnerability
$P_{VULN/OCC}$ = prob. of vulnerability for a given occurrence
P_{OCC} = prob. of the occurrence

Example: Calculating Casualties

- The probability of casualties can be defined as:

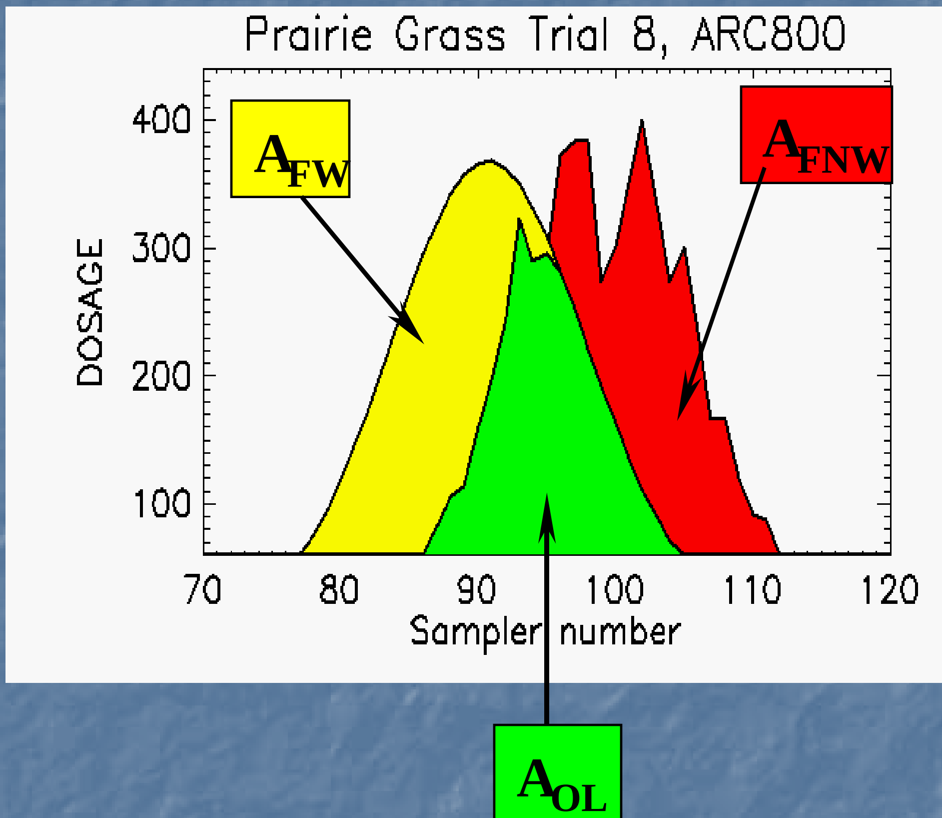
$$P_{CAS} = P_{CAS/CONT} * P_{CONT/EXP} * P_{EXP}$$

P_{CAS} = prob. of casualties
$P_{CAS/CONT}$ = prob. of casualties for a given contamination
$P_{CONT/EXP}$ = prob. of contamination for a given exposure
P_{EXP} = prob. of exposure

Probability of Casualties

$P_{CAS/CONT}$	$P_{CONT/EXP}$	P_{EXP}
• Probability of hazard for a given vulnerability • Toxicity is already a probability • Data for Inhalation or contact	• Probability of vulnerability for a given occurrence • Can be empirical or computed by JOEF • Can account for Protection Levels (by inverting protection factors – ratios of concentrations outside a system to inside a system)	• Probability of occurrence • Performed by JEM's Atmospheric Transport and Dispersion algorithms • MET DATA • SOURCE TERMS • T&D PHYSICS

Falsely and Falsely Not Warned



Adapted from IDA work by Dr. Steve Warner

- Can compute fractions of areas (under curves) of FW and FNW to observed and/or predicted areas.
- Given enough field trial data, means of FW and FNW can be used as probabilities.
- $\text{Accuracy}_{\pm}$ then $1 - \overline{FW}$, $1 - \overline{FNW}$, (false pos. acc. and false neg. acc.)
- *Propose a tentative ~ 0.2 FW accuracy standard and a tentative ~ 0.4 FNW accuracy standard which 1) HPAC has demonstrated and 2) is apparently accepted by HPAC users.*
- *Comparisons to ATP-45 (as current JORD language reads) may be problematic for FNW.*

Transport and Diffusion Statistical Comparisons


Designation: D 6589 – 00

Standard Guide for
Statistical Evaluation of Atmospheric Dispersion Model Performance[1]

1. Scope

1.1 This guide provides techniques that are useful for the comparison of modeled air concentrations with observed field data.

2. Referenced Documents

2.1 *ASTM Standards:*

D 1356 Terminology Relating to Sampling and Analysis of Atmospheres[2]

3. Terminology

- ASTM D 6589-00, Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance
- Geometric mean, geometric variance, bias (absolute, fractional), normalized mean square error, correlation coefficient, robust highest concentration
- Use all available Validation Database data
- Allows comparison to other T&D models and legacy codes

Anticipated Approach

- Use both the IDA methodology and the T&D statistics using the validation database
- The IDA metrics are easy for the warfighter to understand and visualize the threat
- T&D statistics provide a basis for comparison to legacy codes (comparison of observed and predicted means)

Model Evaluation in the Department of Defense

MITRE

Model Evaluation Methods

- The Department of Defense (DoD) assesses the credibility of models and simulations through a process of Verification, Validation and Accreditation (VV&A)
- The JEM model evaluation approach will follow DoD VV&A policy and procedures
  - DoD Instruction 5000.61, May 2003
  - DoD VV&A Recommended Practices Guide, November 1996

Current Methods Deficiencies
Science View

- Inadequate use of statistical methods for assessing the credibility of models and simulations
- Over-reliance on subjective methods, such as subject matter experts, whose credentials are not documented and who may not be current or unbiased

Current Method Deficiencies
User View

- Insufficient guidance on what techniques are available and how to select suitable techniques for a given application
- Overemphasis on using models for their own sake rather than as tools to support analysis
- Failure to carefully and clearly define the problem for which models are being used to provide insights or solutions
- Not understanding the inherent limitations of models

JEM

Is Developing a Common Framework an Achievable Goal?

- The Military Operations Research Society (MORS) sponsored a workshop in October 2002 to address the status and health of VV&A practices in DoD
- An Ad Hoc committee was also formed by Mr. Walt Hollis, Deputy Undersecretary of the Army (Operations Research) to address general concerns regarding capability and expectations for using M&S in acquisition

Best Practices

An Initial Set of Goals for Developing a Common Framework for Model Evaluation

- Understand the problem
- Determine use of M&S
- Focus accreditation criteria
- Scope the V&V
- Contract to get what you need

MITRE

Best Practices
Understand the Problem

- Understand the problem/questions so that you can develop sound criteria against which to evaluate possible solutions.
- This first step is frequently disregarded
  - Too hard to do
  - Assumption that the problem is “clearly evident”
  - No real understanding of what the problem is
- Failure to understand the problem results in:
  - Unfocused use of M&S
  - Unbounded VV&A
  - Unnecessary expenditure of time and money without meaningful results

Best Practices:
Determine Use of M&S

- M&S should be used to:
  - help critical thinking
  - generate hypotheses
  - provide insights
  - perform sensitivity analyses to identify logical consequences of assumptions
  - generate imaginable scenarios
  - visualize results
  - crystallize thinking
  - suggest experiments the importance of which can only be demonstrated by the use of M&S
- Know up front what you will use M&S for
  - What part of the problem can be answered by M&S?
  - What requirements do you have that M&S can address?
  - Use M&S appropriately: If the model isn't sufficiently accurate vis-à-vis the test issues, it may lead to flawed information for decisions, such as test planning and execution.

Best Practices:
Focus Accreditation Criteria

- Focus the accreditation on the criteria; focus the VV&A effort on reducing program risk.
- Accreditation criteria are best developed through collaboration among all stakeholders:
  - Get System Program Office buy-in on the M&S effort
  - Foster collaboration with testers, analysts, and model developers...and key decisionmakers.
  - Build the training simulation first to elicit the user's needs and ideas about "key criteria" before building the system itself. If this is not feasible, think "training" and ultimate use to define key criteria.

Best Practices:
Scope the V&V

- Focus the accreditation on the criteria as that will properly focus the verification and validation (V&V) effort.
- Write the accreditation plan so that it scopes and bounds the V&V effort.
- Plan the V&V effort in advance to use test data to validate the model
  - Need to ensure that the data are used correctly
  - Need to document the conditions under which the data were collected

Best Practices:
Contract to Get What You Need

- Include M&S and VV&A requirements in Requests for Proposals to obtain bids and estimated costs
  - Require that bidders provide a conceptual model that can be used to assess the bidder's understanding of the problem
  - Make documentation of M&S and VV&A deliverables under the contract
  - Competition among bidders will generate alternative models and a broader range of ideas
  - Require that the Government announce what M&S it will use as part of their source selection
  - Ensure that the VV&A team is multidisciplinary and includes experts who understand the mathematical and physical underpinnings of the model
  - Ensure that Subject Matter Experts (SMEs) have relevant and current expertise, that these qualifications are documented, and that the rational of each SME for his or her judgements and recommendations are recorded

Other Goals for Developing a Common Approach for Model Evaluation

- Consensus and commitment to developing and using credible models
- Adequate resources to conduct credible model evaluation efforts
  - Qualified evaluators
  - Funding
- Provide program incentives to those who are expected to perform model evaluations

Other Goals for Developing a Common Approach for Model Evaluation

- Use independent evaluators, not the model developers, to obtain unbiased evaluation results
- Institutionalize the use and reuse of models that have documented credibility
- Establish standards for model evaluation performance to ensure quality and integrity of evaluation efforts