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

18th International Conference on Harmonisation within
Atmospheric Dispersion Modelling for Regulatory Purposes
Bologna, Italy
9–12 October, 2017

An overview of operational model evaluation

Jeffry T. Urban and Nathan Platt

Institute for Defense Analyses

Sponsor: US Defense Threat Reduction Agency – Joint Science and Technology Office for Chemical and Biological Defense (DTRA-JSTO/CBD)

IDA

Overview

- This presentation draws on IDA’s experience in independent model evaluation to discuss the operational evaluation of models
  - What is operational evaluation and why is it important?
  - What are its elements and how is it performed?

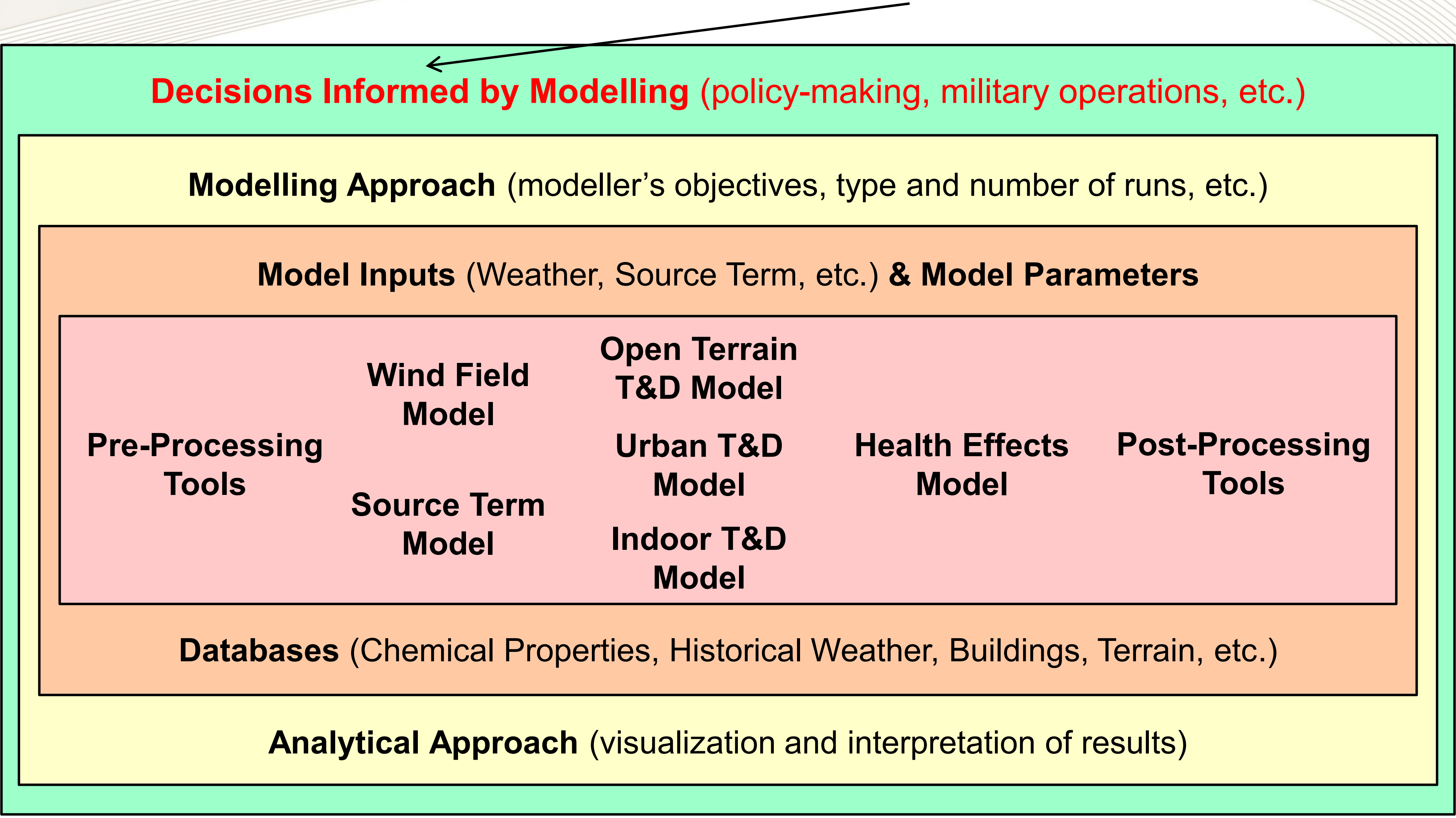
- Purpose is to initiate discussion within the community on how to approach operational evaluation
  - Operational evaluation is not often discussed in the scientific community
  - There is no universal procedure for performing operational evaluation
  - Our group’s focus is modelling for chemical and biological defense applications, but many evaluation principles can be generalized
  - The presentation covers many topics – follow-up dialogue is welcome!

Operational Evaluation vs. Scientific Evaluation

- **<u>Scientific Evaluation:</u>** Does the model meet its technical requirements, and does it represent physical phenomena accurately?
  - Does the model contains errors? How close to the state-of-the-art is it?
  - Scientific evaluation usually focuses on individual models or their subcomponents
  - "Gold standard": Validate the model using the best-quality experimental data or by comparing to other validated, high-fidelity models
  - **Bottom line: Is the model scientifically accurate?**

- **<u>Operational Evaluation:</u>** Is the model acceptable for its intended uses?
  - Evaluates the "modeling enterprise" (the model in its operational context)
    - Requires end-to-end evaluation of all models in the modelling system
    - Also includes evaluation of data limitations, modelling protocols, etc.
  - "Intended use" = end user's intent (maybe different from developer's intent)
  - Operational evaluation can help determine whether a prototype model has become mature enough for operational use
  - Can help determine the uses (if any) for which a model should be applied
  - **Bottom line: Is the model good enough for specific applications?**
    - "State-of-the-art" ≠ "Good enough"
    - "Good enough" = policy-makers make better decisions with model than without it?
      - Maybe not . . . if model is inaccurate, or misleading, or misapplied, or is subject to large uncertainties
    - Policy-makers care about the real-world effects of releases, not their scientific characteristics

The Modelling Enterprise



This is what modelling – and model evaluation – ultimately supports

Decisions Informed by Modelling (policy-making, military operations, etc.)

Modelling Approach (modeller's objectives, type and number of runs, etc.)

Model Inputs (Weather, Source Term, etc.) **& Model Parameters**

Pre-Processing Tools

Wind Field Model

Source Term Model

Open Terrain T&D Model

Urban T&D Model

Indoor T&D Model

Health Effects Model

Post-Processing Tools

Databases (Chemical Properties, Historical Weather, Buildings, Terrain, etc.)

Analytical Approach (visualization and interpretation of results)

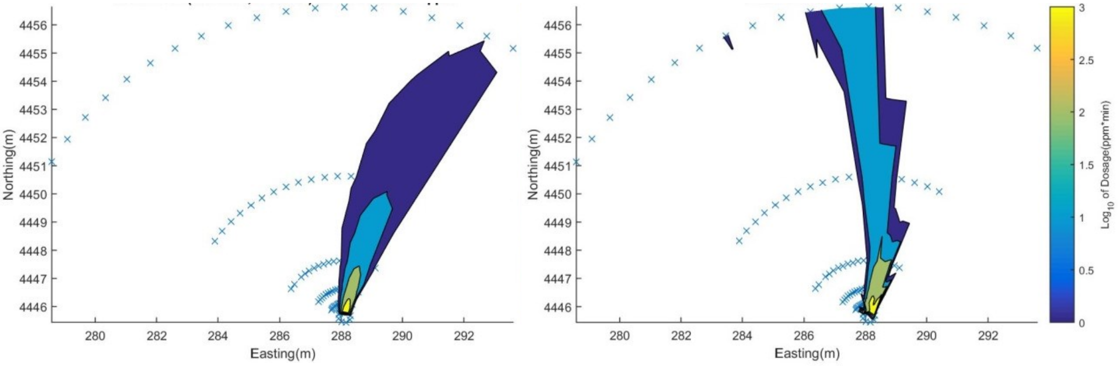
IDA | Model Inputs for Operational Evaluation

- Inputs for Scientific Evaluation:
  - High quality measurements of meteorological parameters, chemical source term parameters, etc. from field campaigns, wind tunnel experiments, etc.

- Inputs for Operational Evaluation:
  - Whatever the modeller would have available during real operations
    - Airport weather observations, numerical weather predictions (NWP), WeatherBug?
    - Rough estimates of emission sources

Emulating operational inputs in field campaign-based evaluations

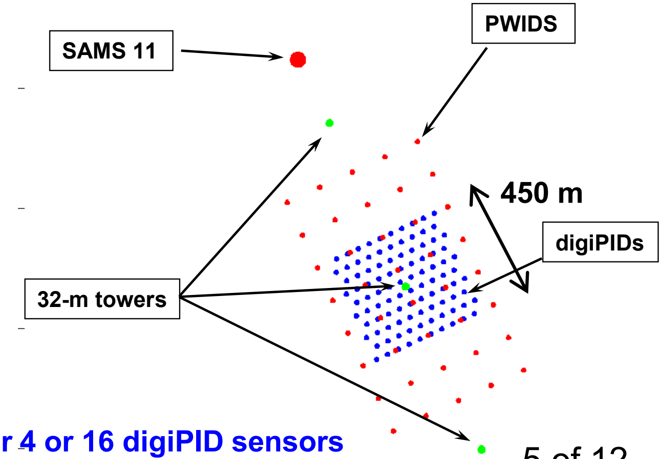
HPAC Evaluation:
JRII with "Operational" NWP Inputs



HPAC Prediction with NAM218 NWP input

Observation

Source Term Estimation Algorithm Evaluation:
FFT07 with Data Denial Protocol



Either 4 or 16 digiPID sensors (of 100) used in evaluation

Model Outputs for Operational Evaluation

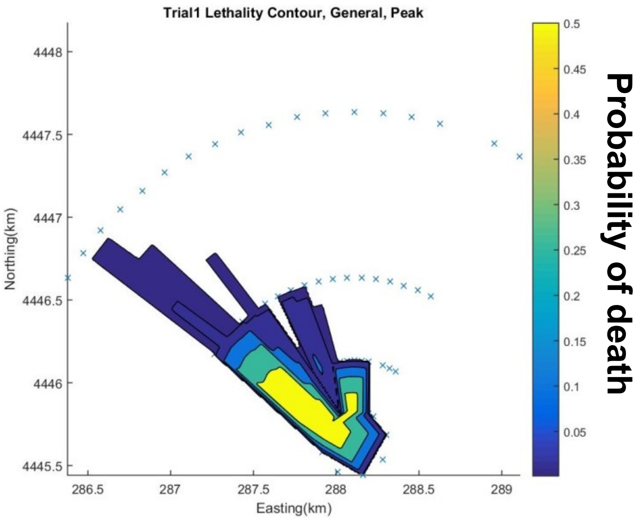
- <u>Outputs for Scientific Evaluation:</u>
  - Usually arc-maximum concentrations and arc-wise plume widths, or sometimes “point-to-point” average concentrations at sampler locations

- <u>Outputs for Operational Evaluation:</u>
  - Whatever the operational modeller provides to customers (e.g., policy-makers)
    - Probably something beyond just concentrations or dosages without further context
  - For hazard predictions, could be number of fatalities, or the locations over which an operationally-relevant average concentration (e.g., 1 hr.) is exceeded

JRII Chlorine Hazard Areas

<u>Observed</u> chlorine concentrations converted to probability of death using toxicological modeling

Basis for evaluating model predictions of hazard areas?

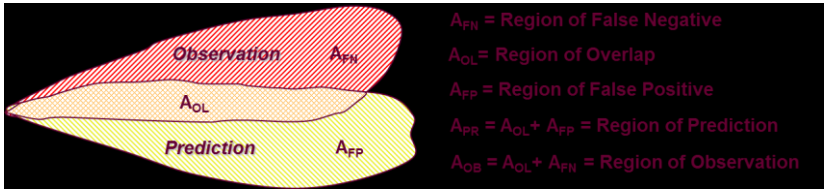


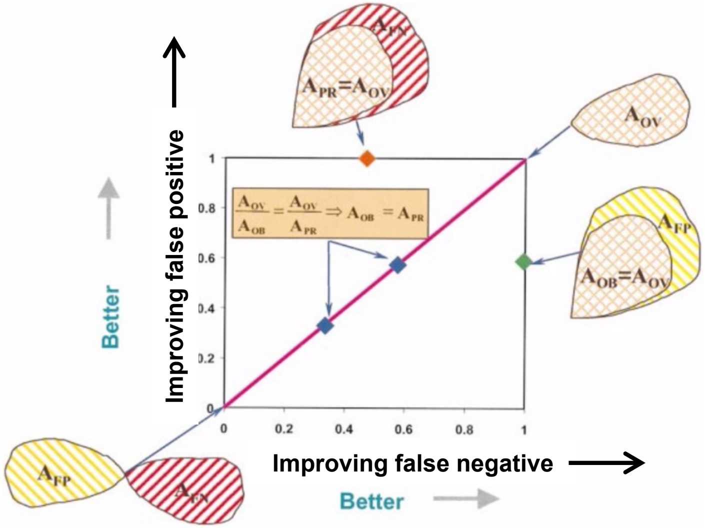
Metrics for Operational Evaluation

- Metrics for Scientific Evaluation:
  - Usually statistical comparisons of observed concentrations to predictions
  - Acceptance criteria are designed to identify state-of-the-art (e.g., |FB| < 0.67, NMSE < 6, FAC2 > 0.3 for urban models), not to assess operational utility

- Metrics for Operational Evaluation:
  - Depends on the application (casualty estimation, hazard area prediction, etc.)!
  - **Critical question: What are the acceptance criteria?** Depends on end user.
    - A state-of-the-art model might not be "good enough" for certain uses (or is overkill)
    - Note: Urban modelling is harder (lower standard for state-of-the-art?), but could be more important because of large civilian populations (higher standard for operations?)

A two-dimensional user-oriented measure of effectiveness

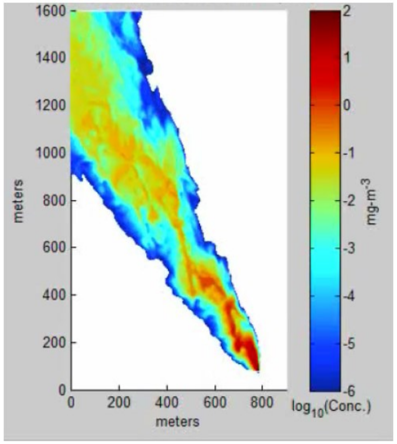

Observation
A_{FN}
A_{OL}
Prediction
A_{FP}
A_{FN} = Region of False Negative
A_{OL}= Region of Overlap
A_{FP} = Region of False Positive
A_{PR} = A_{OL}+ A_{FP} = Region of Prediction
A_{OB} = A_{OL}+ A_{FN} = Region of Observation



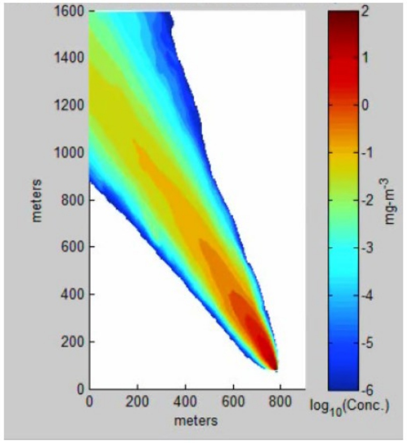
Addressing Uncertainty [1 of 2]

- Policy makers need to manage risk – how so depends on type of application
  - Real-time response: worst case (validated as such!)?; or probabilistic treatment of plume meander, parametric variation of release size, etc.?
  - Training exercise: typical case?
  - Policy planning: probabilistic treatment of historic weather ensembles?

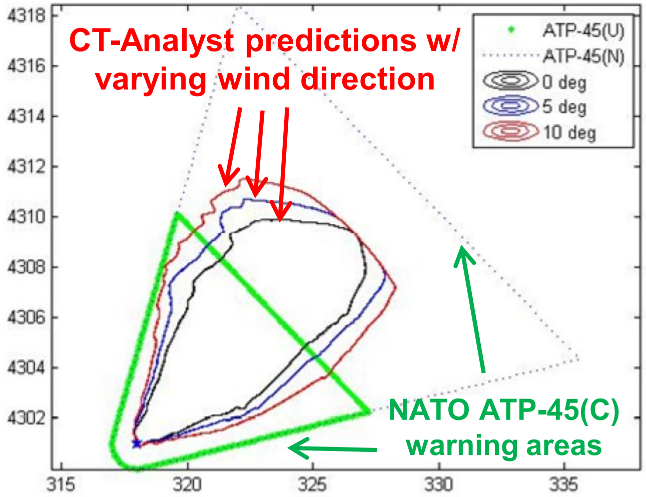
- How uncertainty is addressed depends on the type of model(s)
  - Ensemble average plume
  - Ensemble average + variance (e.g., SCIPUFF, meandering plume model, etc.)
  - “Single-realization” (CFD-like)
  - “Ensembles of models” (like tropical cyclone “spaghetti model” forecasting)?



VTHREAT single realization



Average of 20 VTHREAT realizations



CT-Analyst predictions w/ varying wind direction

NATO ATP-45(C) warning areas

Addressing Uncertainty [2 of 2]

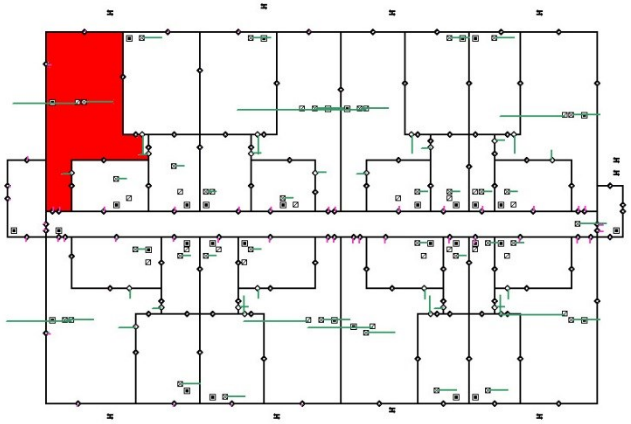
- Epistemic uncertainty (i.e., knowledge gaps) can be as important as – or more important than! – aleatory uncertainty (e.g., arising from stochastic turbulence)
  - Epistemic uncertainty is usually addressed via "modelling assumptions"
  - Modelling assumptions are not always transparent or well-vetted
  - Beware "generic scenarios" with overly-specific inputs to deterministic high-fidelity models – low-fidelity modelling, or probabilistic modelling, might be better
  - Sometimes a "complex" model can give worse results than a simple one because epistemic uncertainty – yet be trusted more because it "has more physics"!

- **Saying "I don't know" is sometimes OK!**
  - "'Can-do' is not 'must-do' or 'should-do'"
  - Addressing uncertainty openly allows policy-makers to manage risk better
  - Operational evaluation can assess how risk is managed in the modelling enterprise

CONTAM building zone and ventilation representation

Maybe simple "box model" would be better if building layout is not known?

Could parameterize box model using ensembles of CONTAM runs?



Runtime, Reliability, and Usability Requirements

- Operational evaluation can help ensure model meets operational requirements:
  - Model runs without errors <u>under operational conditions across relevant cases</u>
  - Model meets runtime requirements <u>under operational conditions across relevant cases</u>

- Ideally, some operational evaluation will involve observing actual users running the model under typical conditions
  - Identify differences between developers' and users' expectations for the model
    - Is the model being used correctly?
    - How well do users' trust the model in different circumstances?
  - Identify deficiencies in model documentation and training
  - Identify user interface problems
  - Refine modelling protocols
  - Understand real-world data limitations, time constraints, policy-maker decisions, etc.

Recommendations [1 of 2]

- Consider the operational context of the model at all stages of development (including the conceptual design of models and integrated modelling systems)
    - Models, and the modelling approach, must differ according to the operational use (consequence planning, real-time response, assessing protective equipment, etc.)
    - The approach to operational evaluation also depends on the models and their uses
        - So no “standard approach” to operational evaluation – although there are general principles!

- Operational evaluations should include the following elements:
    - Modelling protocols that emulate operational use
        - Protocols for developing input databases (e.g., building databases) should be evaluated too!
    - Operational-like model inputs (e.g., not just meteorological data from field campaigns)
    - Operationally-relevant model outputs (e.g., not just concentrations)
    - Evaluation metrics that link model performance to mission effectiveness

- Develop criteria for distinguishing between research tools and operational tools
    - Models should be mature and fit for purpose
        - Models may be fit for some purposes but not others
    - Operational evaluations should use operationally-relevant model inputs and modelling protocols, and evaluation results should
    - Acceptance criteria should be well-defined in terms of the model’s impact on decision-making (preferably before operational evaluation occurs)

Recommendations [2 of 2]

- Explicitly consider the impact of knowledge gaps and other forms of uncertainty
  - Affects everything from conceptual design to development to operating procedures
  - Also can help policy-makers manage risk when consuming modelling products
  - Don’t try to model everything! It’s OK to admit limitations of models and knowledge.
- Operational evaluation is informed by, and can inform, good documentation
  - Technical documentation and operating concept should be in place before evaluation
  - Modelling assumptions and model logical flow should transparent (and users should be notified when deterministic sub-models are engaged)
  - Users should develop operating procedures during development and document them
    - Requires coordination with model developers
  - Operational evaluation can inform the development of “capabilities and limitations” documents for users
- Consider the role of <u>independent</u> model evaluation
  - Professional evaluators with specialized expertise
  - Can bridge the scientific and operational communities
  - Not burdened by operational tempo (users) or product deadlines (developers)
  - No stake in the outcome: helps ensure models are not pushed into inappropriate uses
  - Can also help define modelling requirements (e.g., by ensuring that they are testable)