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

AN OVERVIEW OF USER-ORIENTED OPERATIONAL MODEL EVALUATION

Jeffry T. Urban and Nathan Platt

Institute for Defense Analyses, Alexandria, Virginia, USA

Abstract: This presentation provides an overview of an operational model evaluation and its role in the operational acceptance of atmospheric dispersion models and modelling protocols. We address the distinction between scientific and operational model evaluation, with an emphasis on operationally-relevant evaluation protocols, metrics, and acceptance criteria. We also discuss end-to-end vs. individual component model evaluation, the role of uncertainty in model evaluation, and model runtime, maintainability, and usability requirements. Finally, we suggest some ways in which the relationship between model developers, users, and evaluators can be structured to promote the development of realistic and testable model requirements and sound model development and operational acceptance decisions.

Key words: *operational model evaluation; verification and validation (V&V); operational V&V; scientific V&V; operational metrics*

INTRODUCTION

Atmospheric dispersion models are used to support government and commercial industry decision-making in areas such as air quality regulation, industrial health and safety planning and regulation, emergency response operations and planning, and military operations and planning. The development and use of these models is typically guided by the results of verification and validation (V&V) activities that are designed to ensure that the models correctly implement their developers' vision and produce predictions that conform to known theoretical principles and experimental observations. Often, however, V&V activities are conducted by scientists for a scientific audience using protocols that probe the models' scientific accuracy without thoroughly addressing the broader question of their utility. This is problematic because even a state-of-the-art scientific model may be unsuitable for some of its intended uses.

In this paper, we posit that a sound model evaluation regime should include not only scientific evaluation (i.e., the usual scientific V&V activities), but also operational evaluation to determine the models' utility in the context of their actual operational use. Operational evaluation requires running the models in similar, if not identical, ways that their intended users run them – perhaps using actual model users as test personnel. Crucially, operational model evaluation should also include a determination of whether the model is suitable for its intended use.

In this paper, we propose that scientific model evaluation be used to assess the technological state-of-the-art and guide model development, and that operational model evaluation be used to improve the operational utility of models and to assess whether models, in their present state of development, should be accepted for operational use. This paper focuses on hazard prediction modeling for atmospheric releases of chemical and biological weapons and toxic industrial chemicals. We cite several examples from our experiences evaluating the U.S. Department of Defense's HPAC modelling system.

OPERATIONALLY-RELEVANT MODEL INPUTS, OUTPUTS, AND EVALUATION METRICS

Operational model evaluation requires the use of model inputs and outputs that are relevant to real-world scenarios. It also requires model evaluation metrics that can be translated into model acceptance criteria that are based on the model's ability to perform its end use (e.g., to inform policy decisions).

Operationally-relevant model inputs

Since the goal of the scientific model validation is to determine how well the model represents physical reality, validation studies typically use high-quality ("pristine") measurements from atmospheric dispersion field campaigns or laboratory experiments as model inputs. Real-world atmospheric release events, however, rarely occur in such heavily-instrumented or well-controlled environments. A scientific validation study might, for example, derive model meteorological input data from SODAR and wind profiler measurements taken onsite during a dispersion field campaign: measurements that likely will not be available during a real event. Operational approaches to modelling real events include deriving model inputs from wind measurements made at the nearest airport, or from periodically-generated numerical weather prediction (NWP) data. We recently conducted an evaluation of HPAC using "operational meteorology" by comparing HPAC dispersion predictions made using NWP data available to routine HPAC users to dispersion measurements made during the Jack Rabbit II field campaign (Luong, 2016).

Another approach to "operational-like" model evaluation is to apply a data-denial protocol to experimental measurements. In a study to assess the state-of-the-art in source term estimation (STE) algorithms designed to predict the source of an airborne release from agent concentration measurements, we derived algorithm inputs by reducing FUSION Field Trial 2007 (FFT07) concentration data and meteorological data to better emulate likely operational sensor configurations and capabilities, such as using only a subset of sensors or simulating decreased sensor resolution (Platt and DeRiggi, 2012).

Operationally-relevant model outputs

The standard for scientific validation is to compare experimental measurements to model predictions of those same observed quantities: for example, atmospheric concentrations at sensor locations. For many model applications, however, it is not predicted atmospheric concentrations that are of direct interest to policy-makers' decisions, but rather derived quantities such as the size and location of hazardous areas or number of casualties. It is difficult to conduct evaluations using these outputs because of the scarcity of quality data from real-world incidents, such as industrial chemical accidents. Nonetheless, there may be value in conducting operational-like evaluations using hazard area or casualty outputs, either in direct model-to-model comparisons or comparisons with experimental data converted into hazard area or casualty estimates using human health effects models.

Operationally-relevant model evaluation metrics

The choice of model evaluation metrics – and corresponding model acceptance criteria – is determined by the modelling application. One frequently-convenient set of metrics is the maximum concentration and the cloud width on a sampler arc at a fixed distance from the release (Chang and Hanna, 2004). These simple metrics capture basic information about the shape and size of the plume, but they are less well suited for determining how well the actual locations of hazardous areas are modelled. Uncertainties in the wind direction and wind speed can result in significant errors in the prediction of the location of the plume, even if the transport and dispersion physics is modelled accurately.

Point-to-point metrics that pair each measured concentration with the model's predicted concentration at the same location and time are better suited for assessing the accuracy of hazard area modelling. Point-to-point metrics include standard statistical measures of bias and scatter (e.g., fractional bias (FB) and normalized mean square error (NMSE)) (Hanna, 1989) and user-oriented metrics like the two-dimensional measure of effectiveness (2D MOE) developed by our group (Warner et al., 2004). The 2D MOE has certain advantages for operational evaluations. It is easily interpretable in terms of quantities like false negative and false positive fractions that arguably are easier to frame in an operational context than some of the standard statistical measures. The 2D MOE also naturally lends itself to operational effects-based calculations in which concentrations are converted to, for example, lethality estimates.

Crucially, operational model evaluation also requires the selection of acceptance criteria which, if failed, indicate that the model may not be suitable for a given operational use. Hanna and Chang have proposed acceptance criteria to identify models that reside in the upper tier of historical model performance to provide an aspirational benchmark for model development (Hanna and Chang, 2012), but these criteria merely identify whether models are scientifically state-of-the-art, not whether they perform well in an

operational context. In operational evaluations the acceptance criteria, which answer questions like "how accurate is 'accurate enough'?", must be chosen according to the different applications of the model (e.g., predicting casualties or hazard areas, evaluating the performance of chemical vapor sensors, etc.).

MODEL UNCERTAINTY IN OPERATIONAL EVALUATION

Policy-makers who make decisions based on atmospheric dispersion modelling products that account for the effects of aleatory uncertainty (i.e., variability resulting from unpredictable random phenomena such as atmospheric turbulence) and epistemic uncertainty (i.e., uncertainty arising from a lack of complete knowledge of the modelled system). Scientific V&V efforts attempt to address these factors to some extent: for example, by replicating releases under similar conditions during field campaigns, or by designing experiments to close knowledge gaps. Operational evaluation, however, must be concerned with characterizing the quality of model predictions as ultimately used by policy-makers, which depends strongly on the type of model, type of modelling output, and modelling protocol.

Many atmospheric dispersion models estimate only the average concentration over a notional infinite ensemble of turbulent realizations of the plume; others try to account for turbulent variation probabilistically (e.g., HPAC's SCIPUFF model (Sykes and Gabruk, 1997; Sykes et al., 2014)). Operational evaluations must address whether policy-makers need more than a single ensemble-mean plume prediction, and if so, whether the modelling approach accounts for uncertainty well enough to support decision-making. The answers to these questions depend on the type of decision being made – i.e., on the modelling application. Real-time emergency response to a hazardous materials release, may require either a careful and robust treatment of the probabilistic ensemble and set of modelling assumptions, or a worst-case estimate that has been properly validated as actually representing something near the worst case. A planning exercise may require only a "typical case" rather than a worst case. Other planning efforts may require modelling across a probabilistic ensemble of historical weather conditions (Copeland et al. 2011; Bieringer et al., 2013). Operational evaluation also can help reveal whether simple modelling techniques are sufficient for certain applications: for example, using NATO ATP-45 chemical hazard area predictions in place of explicit dispersion modelling (Heagy et al., 2004; Platt and Jones, 2012), or using simple phenomenological urban dispersion equations in place of building-aware urban dispersion modelling (Hanna and Baja, 2009).

Much of the work in atmospheric model evaluation focuses on quantifying aleatory uncertainty (e.g., meteorological uncertainty). Comparably little attention is given to epistemic uncertainty, which is often addressed via "modelling assumptions" that need to be addressed via operational evaluation. Particular care should be made to ensure that models are not used in an improperly deterministic way. For example, the chemical and biological facilities model within HPAC employs potentially unrealistic assumptions about facility layout and weapon penetration geometry to model facility strikes (Donovan and Masiello, 2015; Dimitrov et al., 2016). As another example, the new multi-zone building interior dispersion models in HPAC include models of "generic buildings" based on common US building stock – operational evaluation should ensure that generic scenarios are used for purposes like planning exercises rather than substituting for specific real-world scenarios (Persily et al., 2010; Urban et al., 2017).

RUNTIME, RELIABILITY, MAINTAINABILITY, AND USEABILITY REQUIREMENTS

Operational evaluation includes not only a scientific component that attempts to answer questions like "is the model accurate enough for use?", but also a component that more broadly addresses the question "is the model useable?" Operational evaluations should, for example, address model verification objectives like ensuring that the model runs to completion and gives the expected outputs for runs that emulate a broad variety of operational scenarios. Evaluations should also verify that model runtimes are acceptable for operational use – again considering a broad range of scenarios since some types of runs may be much longer than others (e.g., when modelling persistent pools of evaporating agents). Evaluators should note whether the model crashes or produces errors, both to aid model development and characterize the operational availability of the model.

Ideally, evaluators should observe the model during operation by its actual users, either during everyday operations or simulated exercises, to ensure that the model is being used in the way that it is intended and

to identify potential improvements to the model or the modelling protocols. Evaluators should also make note of what types of information are typically available to model users (e.g., meteorological data, urban building databases, individual building floorplan and ventilation information, etc.) in order to design operationally-realistic evaluation protocols rather than scientific ones. Finally, evaluators may be in a position to bridge the model development and model user communities by suggesting training protocols and improvements to model documentation and the model user interface (e.g., user notifications when certain modeling capabilities are engaged during runtime).

END-TO-END MODEL EVALUATION

Modern atmospheric dispersion-based modelling software often contains many subcomponents: atmospheric dispersion models, agent source term models, toxicology models, etc. These subcomponent models should be evaluated not only individually for scientific V&V, but also integrated together to emulate how operational users run the software. This enables evaluation using operational scenarios, inputs, and outputs and helps identify user interface problems and runtime and maintainability issues.

Running the model as an integrated end-to-end product also allows evaluators to identify software errors arising from the integration of model subcomponents. Evaluators also may be able to identify conceptual problems with the model integration – these are ideally identified in the software design stage, but not always caught that early. An example is the use of the toxic load model in conjunction with ensemble-average atmospheric dispersion models to estimate human health effects. The toxic load model, unlike dosage-based toxicological models, is sensitive to the concentration fluctuations and intermittency that is averaged away in the ensemble-average approach (Czech et al., 2011).

Finally, it is the whole modelling process, not just the software itself, which should be scrutinized during operational model evaluations. For example, operational evaluation might validate not only the model physics, but also the process for creating model inputs (e.g., building layouts and ventilation representations for indoor dispersion modelling).

THE RELATIONSHIP BETWEEN MODEL DEVELOPMENT AND MODEL EVALUATION

Individual modelling capabilities that may look reasonable (even state-of-the-art) when examined by scientific V&V in isolation may nevertheless be found unsuitable for operational use either individually or in combination. Operational model evaluation can help inform model development and even model requirements to avoid or mitigate this problem.

Oversight of the development of modelling requirements can be aided by model evaluators' perspectives on what constitutes a testable requirement. A general rule of model development is that models (potentially including modelling assumptions and modelling procedures) should be testable either by experiment or well-established theory. Any requirement that is not testable either should be discarded as one that exceeds the current state of knowledge, or at least identified as a potential modelling limitation that must be advertised and mitigated if possible.

Finally, we note that a prerequisite for effective independent operational model evaluation is the availability of the model developers' detailed technical documentation of all components and algorithms within the model and the developers' verification and validation results. Another prerequisite is documentation of the model users' concepts and protocols for employing the model, representative operational scenarios that can be used to test the model, and the applications (e.g., policy decisions) that the model is intended to support. A final prerequisite is a list of model requirements to be tested along with associated operational acceptance criteria.

CONCLUSIONS

We have argued that scientific verification and validation of models is necessary but not sufficient to ensure that the models are useful in the context of their intended applications. Scientific V&V efforts need to be supplemented by operational model evaluation, which focuses running the models in operationally-representative ways using realistic scenarios, model inputs, model outputs, and evaluation metrics. Operational evaluations not only include technical evaluations of a model's accuracy in an

operational context, but also user-oriented evaluations to determine whether the models are fast enough, reliable enough, and user-friendly enough to be effective. A key feature of operational model evaluation is well-defined operational acceptance criteria, which are necessary to differentiate a prototype model from a model that is ready for operational use. An independent corps of model evaluation professionals drawn from outside the model developer and model user communities can help ensure that modelling requirements are realistic and testable and that modelling capabilities (even "state of the art" ones) are not accepted for use in operational applications for which they are unsuitable.

Acknowledgments: This effort was supported by the US Defense Threat Reduction Agency through Mr. Richard J. Fry as the project monitor. The views expressed in this paper are solely those of the authors.

REFERENCES

Bieringer, P.E., S. Longmore, G. Bieberbach, L.M. Rodriguez, J. Copeland, and J. Hannan, 2013: A method for targeting air samplers for facility monitoring in an urban environment. *Atmos. Environ.*, **80**, 1-12.

Chang, J.C. and S.R. Hanna, 2004: Air quality model performance evaluation. *Meteorol. Atmos. Phys.*, **87**, 167-196.

Copeland, J., F. Vandenberghe, and R. Babarsky, 2011: Development of typical-day scenarios for transport and dispersion consequence assessment. *George Mason University Conference on Atmospheric Transport and Dispersion Modeling*, Fairfax, Virginia, USA, July 2011.

Czech, C., Platt, N., Urban, J., Bieringer, P., Bieberbach, G., Wyszogrodzki, A., and J. Weill, 2011: A comparison of hazard area predictions based on the ensemble-mean plume versus individual plume realizations using different toxic load model. *Proceedings of the Annual American Meteorological Society Meeting, Special Symposium on Applications of Air Pollution Meteorology, Dense Gas Experiments & Modeling II*, paper 2.5, January 2011.

Dimitrov, I.K., J.T. Urban, and N. Platt, 2016: Review of the CBFAC Technical Documentation Versions 5.3.226 and 6.2. Memorandum to the Defense Threat Reduction Agency, 12 December 2016.

Donovan, M.T. and P.J. Masiello, *Chemical and Biological Facilities Model CBFAC Version 6.2*. Leidos technical report prepared for DTRA under contract HDTRA1-12-C-0054 (September 2015).

Hanna, S.R., 1989: Confidence limits for air quality model evaluations, as estimated by bootstrap and jackknife resampling methods. *Atmos. Environ.*, **23**, 1385-1398.

Hanna, S.R. and E. Baja, 2009: A simple urban dispersion model tested with tracer data from Oklahoma City and Manhattan. *Atmos. Environ.*, **43**, 778-786.

Hanna, S. and J. Chang, 2012: Acceptance criteria for urban dispersion model evaluation, *Meteorol. Atmos. Phys.*, **116**, 133-146.

Heagy, J.F., Platt, N., and Warner, S., 2004: A quantitative comparison of HPAC predictions and ATP45(B) chemical templates, *Proc. of the 8th International Symposium on Protection Against CBW Agents*, Gothenburg, Sweden.

Luong, K., 2016: User-oriented assessment of a chemical hazard prediction model, IDA Summer Associate Program Final Brief.

Persily, A., A. Musser, and S.J. Emmerich, 2010: Modeled infiltration rate distributions for U.S. housing. *Indoor Air*, **20**, 473-485.

Platt, N. and D. DeRiggi, 2012: Comparative investigation of source term estimation algorithms using FUSION Field Trial 2007 data: Linear regression analysis. *Int. J. Environ. Pollution*, **48**, 13-21.

Platt, N. and L. Jones, 2012: Potential use of transport and dispersion model output to supplement Allied Tactical Publication-45 hazard area prediction templates, *Int. J. Environ. Pollution*, **48**, 30-38.

Sykes, R.I. and R.S. Gabruk, 1997: A second-order closure model for the effect of averaging time on turbulent plume dispersion. *Journal of Applied Meteorology*, **36**, 1038-1045.

Sykes, R.I., Parker, S.F., Henn, D.S., Chowdhury, B., 2014: *SCIPUFF Version 2.8 Technical Documentation*. Sage Management technical report prepared for DTRA under contract No. DTRA01-03-D-0013 (June 2014).

Urban, J.T., J. Henrikson, and N. Platt, Summary of indoor modeling assessment in HPAC 6.3 (Build 6.3.176). Memorandum to the U.S. Defense Threat Reduction Agency, 16 May 2017.

Warner, S., N. Platt, and J.F. Heagy, 2004: User-oriented two-dimensional measure of effectiveness for the evaluation of transport and dispersion models. *Journal of Applied Meteorology*, **43**, 58-73.