H13-101
METHODOLOGY FOR STATISTICAL EVALUATION OF ATMOSPHERIC DISPERSION MODELS IN A RISK ASSESSMENT CONTEXT

Bertrand Sapolin[1], Gilles Bergametti[2], Philippe Bouteilloux[1] and Alain Dutot[2]

[1]DGA Maîtrise NRBC, Vert-le-Petit, France
[2]Laboratoire interuniversitaire des systèmes atmosphériques (LISA), Créteil, France

Abstract: Atmospheric dispersion models are used in chemical risk assessment studies to predict the potential consequences of toxic releases into the atmosphere. Owing to the large public health and economic impacts at stake, the models need to be evaluated extensively. In this study, emphasis is put on statistical evaluation of models compared to experimental field data. It is shown that risk assessment oriented evaluations cannot exclusively rely on methodologies traditionally used in air quality model evaluations. Some features of a methodology better suited to risk assessment specificities are outlined.

Key words: *risk assessment, statistical evaluation, acute inhalation toxicity.*

INTRODUCTION

The threat of chemical, biological and radiological (CBR) terrorist attacks on civilian or military populations has been given much attention in recent years. Such attacks involve releases of potentially highly toxic substances into the atmosphere, which may produce adverse effects on the population. CBR risk assessment activities aim at predicting potential consequences of such scenarios, using transport and dispersion models. Risk assessment must emphasize its operational purpose in order that the model outputs can be used for decision making by the military community or homeland security services. If possible, risk assessment shall also produce an estimate of the uncertainty associated with the model's results.

The models have to be evaluated to assess the confidence level one can put into their predictions. In this study, emphasis is put on statistical evaluation of models compared to experimental field data. The Kit Fox experiment (WRI, 1998) was chosen because it includes short-scale instantaneous or finite duration releases implying non-stationary transport and diffusion in an environment where some complex effects are expected. Hence, Kit Fox is representative of risk assessment scenarios.

The study presented here focuses on chemical risk assessment. The model which has been tested is HPAC (DTRA, 2004). The aim of the study is not to assess absolute model performance but rather use the evaluation results to investigate new methodologies for model evaluation, which would consider risk assessment features better than usual methodologies do.

KIT FOX FIELD EXPERIMENT

The Kit Fox field experiment is a series of short-duration dense gas CO_2 releases conducted in 1995 at the US Department of Energy Nevada Test Site. The flat desert surface of the area was artificially roughened by means of two arrays of obstacles: the ERP (Equivalent Roughness Pattern) near the source as shown by the dashed rectangle in 0, and the URA (Uniform Roughness Array) represented by the outer solid rectangle. The roughness length was about 0.12-0.24 m for the ERP and 0.01-0.02 m for the URA. The roughness length of the surrounding flat desert was about 0.0002 m (Hanna, S. R. and J. C. Chang, 2001).

52 trials were conducted with both nearly instantaneous 20 s releases ("puff releases" below) and 2-8 min continuous releases. The source was a ground level 1.5 m × 1.5 m square near the middle of the roughness arrays. The first 19 trials were carried out with both ERP and URA roughness arrays set up (trials referred to as "ERP" in the following), the remaining trials took place with the ERP removed (trials referred to as "URA" in the following).

77 concentration monitors with a 1 second time resolution were placed along four downwind lines at 25, 50, 100 and 225 m as shown in 0.

The releases were conducted during neutral and stable atmospheric conditions. Several meteorological stations were placed within and outside the roughness arrays, with a time resolution ranging from 1 to 10 s.

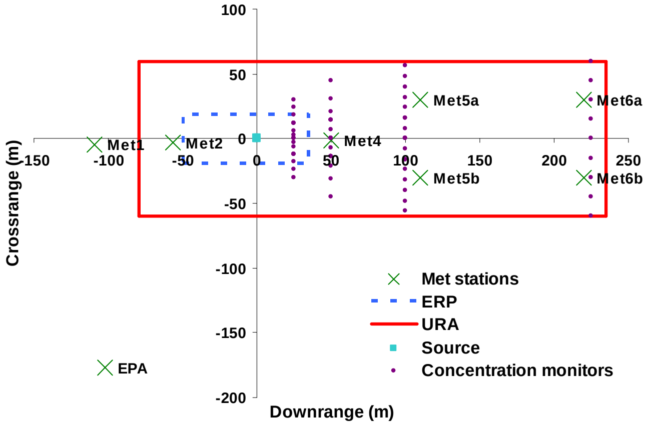

Figure 1. Plot plan of the Kit Fox experiment

HPAC

HPAC is a software from the US Defense Threat Reduction Agency used for assessing consequences of hazardous releases into the atmosphere. The dispersion capability of HPAC is provided by SCIPUFF (Sykes, R. I., S. F. Parker *et al.*, 2006), a Lagrangian model in which Gaussian puffs are used to represent the concentration field. The gridded meteorological data required by SCIPUFF is calculated by one of the two optional diagnostic wind-field models: MC-SCIPUFF or the more complex SWIFT model. SCIPUFF uses a second-order turbulence closure scheme to derive the predicted dispersion rates from velocity fluctuation statistics. The closure scheme is also used to estimate the variance of the statistical distribution of concentration. Based on a theoretical form for the distribution, probabilistic results can be provided. Due to its ease of use and low computational times SCIPUFF is one of the first operational dispersion models with such probabilistic capabilities.

HPAC (version 4.04 SP4) was run against the 52 Kit Fox trials. The modelling domain was a 420 m × 420 m square covering the URA roughness pattern and all the meteorological stations. A 42 × 42 cells grid was used to represent the varying roughness in the dispersion area. The model was run under numerous configurations (various source term models, input data, configuration parameters). The results presented in this study were all obtained with the following parameters:

- Source term: stack release (stack height set to zero to mimic a ground surface source)
- Meteorological data: all stations and vertical levels used, 20 s averaged data, SWIFT wind-field model
- Conditional averaging time and meteorological observation time bin size left to default values
- Sensible heat was calculated from Hanna, S. R. and J. C. Chang (2001).

In order to compare HPAC to the high frequency observations, the model was run with an output frequency of 1 s.

PERFORMANCE MEASURES FOR KIT FOX

Arc max concentration (MVK protocol)

The Model Validation Kit (MVK) is a package of datasets and software for evaluation of atmospheric dispersion models supported by the European Initiative on “Harmonization within Atmospheric Dispersion Modelling for Regulatory purposes” (Olesen, H. R. and J. C. Chang, 2005). The MVK includes a statistical tool (BOOT) based on the work of (Hanna, S. R., J. C. Chang *et al.*, 1993). The MVK emphasizes evaluation based on arc max or crosswind-integrated concentrations even if the BOOT software has a more general scope. For the sake of consistency with the MVK, comparisons of arc max concentrations were made first and are presented in Table 1. Instantaneous and 20 seconds moving averaged concentrations were considered. Following Hanna, S. R. and J. C. Chang (2001), the trials were split into four blocks.

Table 1. Typical FAC2 values (%) and their 95% confidence intervals, based on comparisons of predicted and observed arc max concentrations for Kit Fox - FAC2 is the fraction of predictions within a factor of two of measures.

		Instantaneous concentration	20s moving average concentration
Block results	ERP puff	63.5[49-76.4]	50[35.8-64.2]
	ERP continuous	54.2[32.8-74.4]	45.8[22.1-63.4]
	URA puff	65.5[54.3-75.5]	66.7[55.5-76.6]
	URA continuous	45.8[29.5-58.8]	41.7[27.6-56.8]
Overall results		**59.2[52.1-65.9]**	**54.3[46.8-60.8]**

This methodology is widely used in air quality modelling where the peak concentration is of interest for regulatory applications. However it might prove inappropriate in risk assessment. Firstly the arc max value is not the most relevant variable of concern. It is indeed more important to know what happens on the borders of the cloud rather than in the centre. In a first attempt to satisfy this condition one can consider doing point-to-point comparisons. Secondly, concentration cannot be directly related to a toxic effect so variables in closer connection to toxicity should be used instead, as explained hereafter.

Effect-related variables of interest for model evaluation

Acute inhalation toxicity of chemicals is a nonlinear function of concentration and duration. The usual relation suggests that a given level of effect is reached by a fixed value of dosage Ct (concentration integrated with time) Owing to frequent departures from this relationship, a more general description may be adopted (ten Berge, W. F., A. Zwart *et al.*, 1986):

$$TL(t) = k \quad \text{where} \quad TL(t) = \int_0^t [C(\xi)]^n d\xi \tag{1}$$

TL is the toxic load (also written $C^n t$). The exponent *n* depends on the substance and the effect. Typical values for *n* are in the range 0.5-4 (ten Berge, W. F., A. Zwart *et al.*, 1986).

Statistical distribution of the population response to toxic load is usually lognormal (Sommerville, D. R., K. H. Park *et al.*, 2006), which means that equation (1) can be extended to a cumulative distribution function (cdf) of the population function and then linearized using a probit function (Finney, D. J., 1971) to give:

$$Y = a.Ln(C^n t) + b \tag{2}$$

Where *Y* is the probit value associated to the impact on the population;
a, *b* and *n* are constants depending on the effect and the toxic agent, determined by analysing experimental data.

For a toxic agent where *a*, *b* and *n* are known, the fraction Φ of the population suffering effect for a given toxic load (*TL*) can then be inferred from the cumulative distribution function of a standard normal distribution:

$$\Phi(TL) = \frac{1}{2}\left[1 + erf\left(\frac{a.\ln(TL) + b}{\sqrt{2}}\right)\right] \tag{3}$$

Chemical risk assessment studies consider agents of variable toxicity, from toxic industrial chemicals to highly toxic warfare agents. In order to investigate the validity of our approach over a large toxicity range, a hierarchy of substances was established by firstly ranking them into four classes, from the least to the most toxic, and then choosing a representative agent in each class. The criterion for establishing the classes is based on AEGL-3 values for a 10-min exposure time (Acute Exposure Guideline Levels). The AEGL-3 values were retained as they represent a widely accepted threshold. The 10-min exposure time is typical of accidental releases. Table 2 summarizes our choice.

Table 2. Description of the four classes spanning the toxicity range of chemical agents, and properties of a representative agent in each class. Probit parameters are taken from INERIS (2003; 2008)

Classes			Benchmark agents			
Rank	Toxicity	AEGL-3 10 min range (mg/m^3)	Agent name	Probit parameters (C in ppm, t in min)		
				a	*b*	*n*
I	Low	AEGL-3>500	Ammonia NH_3	2.17	-47.4	1.83
II	Moderate	50<AEGL-3<500	Hydrogen fluoride HF	2.63	-29.9	1
III	High	5<AEGL-3<50	Phosphine PH_3	16.81	-120.89	0.5
IV	Very high	AEGL-3<5	Arsine AsH_3	2.65	-26.08	1.18

Point-to-point comparisons based on effect-related variables

Dosage or toxic load can be inferred from the concentration time series using equation (1). Paired-in-space predicted and measured dosage and toxic load for the four benchmark agents were performed and the results are presented in Table 3.

Table 3. FAC2 values (%) and their 95% confidence intervals for the same model configuration as in Table 1. In order to remove the smallest values from evaluation a cut-off dosage value of 1 ppm.s was used.

		Ct	C^nt			
			NH_3	HF	PH_3	AsH_3
Block results	ERP puff	21.1[18.2-24]	13.8[11.4-16.4]	21.6[18.6-24.6]	33.9[30.4-37.4]	19.2[16.4-22.1]
	ERP cont.	22.9[18.7-27.1]	13.1[9.7-16.6]	23.3[19.2-27.8]	34.9[30.1-39.8]	22[17.8-26.2]
	URA puff	29.5[26.6-32.5]	18.3[15.8-21]	30.4[27.4-33.5]	55[51.4-58.3]	26.1[23.2-29]
	URA cont.	35.5[32-39.1	20.4[17.4-23.4]	36[32.5-39.6]	61.2[57.3-64.7]	29.9[26.5-33.3]
Overall results		**27.8[26.1-29.5]**	**16.9[15.5-18.3]**	**28.4[26.7-30.1]**	**47.8[45.9-49.7]**	**24.6[23-26.2]**

As already explained above, the main purpose of the study is to investigate how model performance may be influenced by the evaluation objective. Table 3 shows overall poor performance compared to Table 1. Pairing in space is indeed more stringent than just comparing arc max values. Only FAC2 was presented here for illustration but the same conclusions are obtained with the other basic performance measures of the BOOT software. It is to be noted that the FAC2 decreases as the *n* exponent increases, as more weight is given to the uncertain concentration when *n* is increased.

Suggested use of effect-related variables in model evaluation

One may wonder whether dosage or toxic load are the most appropriate variables, and whether usual criteria such as FAC2 are applicable for a risk related evaluation framework.

The fraction of fatalities can be directly inferred from toxic load using equation (3). 0 illustrates this for the four benchmark agents. Note that this representation may be deceptively giving the impression that arsine (AsH_3) is not the most toxic agent, and yet it is.

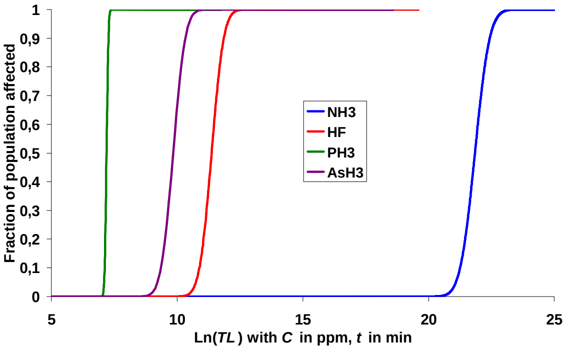

Figure 2. Population response as a function of toxic load

The curves in 0 all exhibit the same pattern which can be broken down into three distinct parts: two plateaus (“no effect” and “full effect”) connected by a very sloping part. The slope is seemingly not very different from one substance to the other, the main difference being the threshold at which it starts. Let C5 and C95 be the concentrations for respectively 5% and 95% of population response. The ratios r=C95/C5 for NH_3, HF, PH_3 and AsH_3 are respectively 2.29, 3.48, 1.47, and 2.85. It is to be noticed that even though C95 and C5 are both functions of exposure duration, C95/C5 is a constant value for each agent.

These ratio values show that the population response only changes on a very narrow range of concentration. A more extensive investigation should be made to confirm this statement over a broader number of agents but our experience shows that C95/C5 hardly ever exceeds 5. A small C95/C5 ratio implies that a slight error in model’s predictions might have important consequences if the measured data is in the range C5-C95. Conversely large errors in the steady parts may not impair the accuracy of model prediction.

Owing to the relatively small values of r, statistical criteria such as FAC2 may prove inappropriate. More generally, usual performance measures such as FAC2, FB, NMSE (Chang, J. C. and S. R. Hanna, 2005), which emphasize the amplitude of the differences measure / prediction, fail to elicit the non linear influence of a given difference measure / prediction on the overall model performance. For this reason it is suggested in the following to introduce toxicological laws into the evaluation methodology and to compare predicted and measured fractions of affected population. Recognizing that the most important in risk assessment is to know whether a model is able to predict the contours of effect at various population response levels, we suggest comparing predicted and measured fractions of population based on a given incidence level.

The results can be presented in a contingency table as shown on the left part of 0. To get the contingency table, one has to calculate for each monitor whether the beforehand fixed incidence level is exceeded or not, by the measures and the model. Several performance criteria may be defined as shown on the right part of 0.

Event observed? Event predicted?	Yes	No	Total
Yes	*A*	*D*	*A*+*D*
No	*C*	*B*	*C*+*B*
Total	*A*+*C*	*D*+*B*	*N* = *A*+*B*+*C*+*D*

$$R_{fp} = \frac{D}{D+B}, \quad R_{fn} = \frac{C}{A+C}, \quad R_d = \frac{A}{A+C}$$

$$R_{ga} = \frac{A+B}{N}, \quad R_{ba} = \frac{C+D}{N}$$

Figure 3. Left: contingency table for a fixed incidence level. Right: several measures based on contingency tables (Fienberg, S. E., 1980): R_{fp} = false positive rate, R_{fn} = false negative rate, R_d = detection rate, R_{ga} = good analysis rate, R_{ba} = bad analysis rate.

Table 4 shows the results obtained for the four benchmark substances with exactly the same model configuration as that used to obtain Table 1 and Table 3. The incidence level was set to 1%.

Table 4. Lethal effect for comparisons based on a variable representing the statistical population response and a 1% incidence level. n.s.: not significant (NH_3 is not toxic enough for inducing significant overlap, false positive and false negative values.

Agent	R_d	R_{fn}	R_{fp}	R_{ga}	R_{ba}
NH_3	n.s.	n.s.	n.s.	n.s.	n.s.
HF	82%	18%	6%	93%	7%
PH_3	80%	20%	17%	98%	2%
AsH_3	72%	28%	19%	79%	21%

The overall results show false negative rates under 30%, detection rates over 70%, false positive rates under 20%, good analysis rates over 75% and bad analysis rates under 25%, which seems encouraging. But what is most important is that these criteria are more suited to the expectations of operational users than usual statistical criteria are.

There is a similarity between the proposed criteria and the Measures of Effectiveness MOE (Warner, S., N. Platt *et al.*, 2001) where contour areas are compared. Because of the limited spatial resolution of concentration monitors in most field experiments, point-to-point data summation is suggested as a substitute for the area estimates (Chang, J. C. and S. R. Hanna, 2005), thus allowing to calculate the variables of the contingency table in a relatively straightforward manner. However, to build Table 4, we have not used point-to-point data summation because the latter approach emphasizes the amplitude of the difference between measure and prediction. Data summation is an unnecessary stringent condition in risk assessment evaluation, and it does not consider the non linearity of the population response to toxic load. For these reasons, it is suggested that data summation be replaced by a simple counting of individuals exceeding or not a beforehand fixed incidence level (fraction of population). Acting so permits to give the same weight to a big or a small overestimate (or underestimate). The results in Table 4 were obtained with this method and the results appear fairly good.

INCLUDING CONCENTRATION FLUCTUATIONS IN RISK-ORIENTED MODEL EVALUATIONS

The work presented before focuses on model's mean results compared to experimental data. It could be extended to include inherent uncertainties, which is an important concern of operational decision makers. A risk-oriented model evaluation should indeed take this into account. Elements for consideration are suggested hereafter.

The atmospheric boundary layer is random by nature so a fully deterministic prediction of pollutant dispersion cannot be achieved, even with a "perfect" model. Most models provide a mean result, representing the average of the ensemble of all possible random realizations for a given set of input parameters. Conversely, observations are individual realizations of the ensemble. Thus, comparing a mean prediction to an experiment amounts to introducing an arbitrary bias into the evaluation. It should be investigated how SCIPUFF probabilistic capabilities can be used in risk-oriented evaluations to remove this bias. SCIPUFF calculates the first and second moment of the instantaneous concentration distribution. Then the probabilistic prediction for concentration is achieved by assuming a clipped normal distribution (Lewellen, W. S. and R. I. Sykes, 1986). Other authors suggest that a left-shifted and clipped gamma distribution be used (Yee, E., 2008). The concentration time series shown in 0 below illustrate an example of probabilistic predictions obtained with HPAC against Kit Fox.

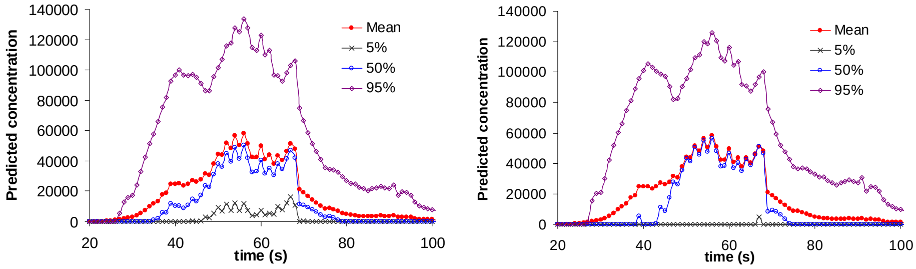

Figure 4. Modelled time series of concentration statistical distributions using a left-shifted clipped gamma distribution (left picture) or a clipped normal distribution (right picture) based on HPAC predictions of mean and variance.

If HPAC were a perfect representation of reality, 90% of the observations would lie within the uncertainty intervals shown on 0. For risk assessment purposes, one would need to build the dosage or toxic load distribution from the known concentration distribution but this is not easily achievable. Indeed the concentration pattern shown in 0 corresponds to a time series of correlated random variables. For each time step, SCIPUFF calculates the mean and variance as well as the integral timescale for the concentration fluctuations from which a covariance matrix can be constructed (two-time point statistics). Hence, dosage is a random variable resulting from the sum of correlated random variables, and there is no way to calculate the theoretical distribution of it. SCIPUFF only provides mean dosage and (under some assumptions) dosage variance but knowledge of a theoretical dosage probability density function is lacking before going further into probabilistic predictions. It is even worse for toxic load because variance is unknown. A possible answer would be to derive many realistic concentration time series from the uncertainty intervals, calculate toxic load and population fraction for each, then build empirical distributions. This would imply developing a sampling method of correlated variables. This work has not been investigated yet.

CONCLUSION

Risk assessment activities for military or homeland security require that the dispersion models be able to provide features such as the contours reached by a given toxic effect. Statistical evaluations of models used in this context have to consider this specificity. The Model Validation Kit protocol focuses on arc max or crosswind integrated concentrations, which makes it a tool better suited to air-quality applications than risk assessment. In this study, we try to outline the main features for the development of a risk assessment oriented evaluation methodology.

We suggest dealing with variables closely related to the acute inhalation toxic effects suffered in a population. The most useful variable of interest is the fraction of population affected. To calculate it, a toxicological model for the distribution of the population response is used and tested against several toxic substances covering a wide toxicity range. We found that the population response exhibits a similar pattern for all products. Using this feature combined to contingency tables and detection based criteria, it is shown that models can exhibit poor performance in point to point comparisons with usual criteria, and though be fairly rated when using the suggested approach. The latter focus on what is really important for risk assessment. In particular, unnecessary stringent conditions are removed, such as the amplitude of measure / prediction discrepancies.

Dispersion in the atmospheric boundary layer is a random process which involves inherent uncertainties. Owing to this, a part of the measure / prediction discrepancies may not be ascribed to the model. Hence, the work presented here should be extended to include uncertainties due to concentration fluctuations. Some dispersion models like SCIPUFF are designed to provide probabilistic concentration fields. This capability could be used to investigate the statistical distributions of risk-related variables such as the fraction of population suffering adverse effect. Due to great complexity, this work has not been done yet.

REFERENCES

Chang, J. C. and S. R. Hanna (2005). Technical Descriptions and User's Guide for the BOOT Statistical Model Evaluation Software Package, Version 2.0.

DTRA, 2004: Hazard Prediction and Assessment Capability, version 4.04, DVD containing model and data.

Fienberg, S. E., 1980: The analysis of cross-classified categorical data, 2nd edn, Cambridge, Massachusetts, MIT Press.

Finney, D. J., 1971: Probit Analysis, 3rd edn, Cambridge Univ. Press.

Hanna, S. R. and J. C. Chang, 2001: Use of the Kit Fox data to analyse dense gas dispersion modeling issues. *Atmospheric Environment*, **35**, 2231-2242.

Hanna, S. R., J. C. Chang and D. G. Strimaitis, 1993: Hazardous gas model evaluation with field observations. *Atmospheric Environment*, **27A**(15), 2265-2285.

INERIS, 2003: Seuils de toxicité aiguë - Acide fluorhydrique (HF).INERIS 03DR072, 42 pages.

INERIS, 2003: Seuils de toxicité aiguë - Ammoniac (NH_3).INERIS 03DR035, 40 pages.

INERIS, 2008: Seuils de toxicité aiguë - Arsine.INERIS 05DR115, 48 pages.

INERIS, 2008: Seuils de toxicité aiguë - Phosphine.INERIS 06DR071, 46 pages.

Lewellen, W. S. and R. I. Sykes, 1986: Analysis of concentration fluctuations from Lidar observations of atmospheric plumes. *Journal of Climate and Applied Meteorology*, **25**, 1145-1154.

Olesen, H. R. and J. C. Chang, 2005: Consolidating tools for model evaluation, *10th International Conference on Harmonization within Atmospheric Dispersion Modelling for Regulatory Purposes*, Sissi, Crete.

Sommerville, D. R., K. H. Park, M. O. Kierzewski, M. D. Dunkel, M. I. Hutton and N. A. Pinto, 2006: Toxic Load Modeling, pages 137-158 in H. Salem and S. A. Katz, eds, Inhalation toxicology, Taylor & Francis Group.

Sykes, R. I., S. F. Parker, D. S. Henn and B. Chowdhury, 2006: PC-SCIPUFF Version 2.2 - Technical Documentation. Titan Corporation, Princeton, New Jersey pages.

ten Berge, W. F., A. Zwart and L. M. Appelman, 1986: Concentration-time mortality response relationship of irritant and systemically acting vapours and gases. *Journal of Hazardous Materials*, **13**, 301-309.

Warner, S., N. Platt, J. F. Heagy, S. Bradley, G. Bieberbach, G. Sugiyama, J. S. Nasstrom, K. T. Foster and D. Larson, 2001: User-Oriented Measures of Effectiveness for the Evaluation of Transport and Dispersion Models. *IDA Paper P-3554*, 815 pages.

WRI, 1998: Final Report on the 1995 Kit Fox Project, Vol. I - Experiment Description and Data Processing, and Vol. II - Data Analysis for Enhanced Roughness Tests. Western Research Institute, Laramie, Wyoming, USA pages.

Yee, E., 2008: The concentration Probability Density Function with implications for probabilistic modeling of chemical warfare agent detector responses for source reconstruction.TR 2008-077, Defence Research and Development Canada (DRDC), 44 pages.