Model Validation Kit – status and outlook

H.R. Olesen

National Environmental Research Institute (NERI)
P.O. Box 358, DK-4000 Roskilde, Denmark
E-mail hro@dmu.dk

Abstract: Over the past few years, the so-called Model Validation Kit has been the basis for much work on model evaluation. The kit has recently been enhanced with a supplement containing experimental data from Indianapolis. A change of methodology is under consideration, based on the concept of near-centreline concentrations. The paper examines some consequences of such a change in methodology.

Keywords: atmospheric dispersion models, model evaluation, Model Validation Kit, near-centreline concentrations.

1 Introduction

The present conference is the fifth in a series of meetings which have been organised by the initiative on "Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes" since 1992. At these meetings, model evaluation has been a key issue. Work has been conducted in order to start establishing a toolbox of recommended methods for model evaluation. The basis for much of this work has been a so-called *Model Validation Kit* (Olesen, 1995a). Over the past few years, the kit has been distributed to approximately 140 research groups.

Since the previous harmonisation workshop in Oostende in 1996, a supplement to the Model Validation Kit has become available. The supplement comprises data from an experiment in Indianapolis, USA, and some related software tools. One section in the present paper introduces this supplement.

The Model Validation Kit has been used for a large number of studies reported at the previous workshops (see IJEP, 1995; 1997). It will also play a role in connection with a new *Model Documentation System* which has become publicly available under the auspices of the European Topic Centre on Air Quality of the European Environment Agency (Moussiopoulos, 1998). This Model Documentation System is a catalogue of models available through the Internet (http://www.etcaq.rivm.nl). The Model Documentation System as such does not contain detailed information on model performance, but model quality is an issue of obvious interest to the users of the system. Therefore, modellers are encouraged to conduct model evaluation exercises according to standard methodologies — to the extent that standard methods exist.

The Model Validation Kit provides one such *de facto* standard methodology and can be recommended for use for the time being. But the kit has limitations, and model evaluation results should be used with prudence as pointed out in a series of previous papers (Olesen, 1994; 1995b; 1996).

If we look into the future, there may be changes to the methodology of the Model Validation Kit. Such potential changes are a topic for discussion in the present paper. The consequences of changes are presently being explored in the model evaluation community, but until consensus has been reached on a new set of well-tested standard tools, the existing Model Validation Kit provides the best available common frame of reference for evaluation of short-range models.

The bulk of this paper is devoted to a discussion of one particular issue which eventually can have great implications. According to the methodology used in the present version of the kit, observed arcwise maxima are compared directly to modelled centreline values. In a draft for an ASTM standard (American Society for Testing of Materials) an alternative methodology has been proposed which uses the concept of "near-centreline concentrations". The implications of using either of the two methodologies will be discussed.

2 Supplement to the Model Validation Kit

The original Model Validation Kit comprises three experimental datasets (Kincaid, Copenhagen and Lillestrøm) and software. The software includes a program for statistical model evaluation, a program for analysis of

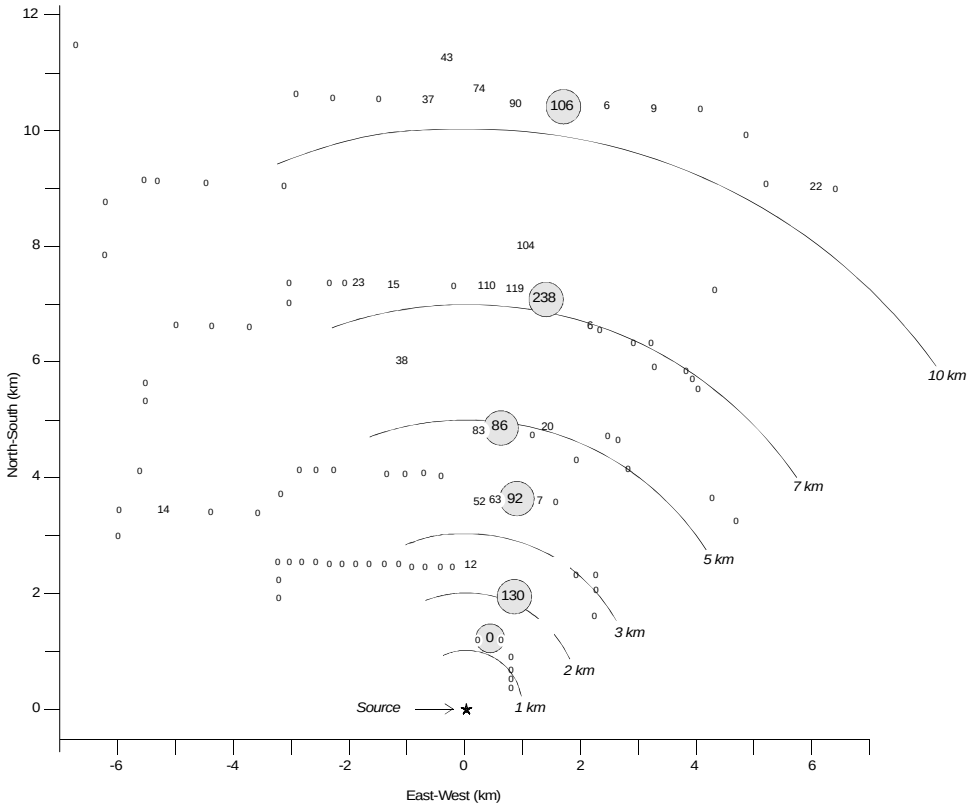

Figure 1 Geographical distribution of measured concentrations at Kincaid, 22 May 1981, 10-11 hours. Values are in ppt, and the arcwise maxima are enclosed in circles.

residuals, and a simple plotting package well suited for presenting the results from these programs. The software was originally developed by Hanna et al. (1991).

The supplement to the Model Validation Kit which became generally available in 1997 includes a data set from Indianapolis (USA) as well as some software tools.

Among these tools are utilities specifically for handling Indianapolis data. They are intended to make it relatively easy for a modeller to combine his modelled results with the observed tracer concentration data from Indianapolis. Additionally, the software tools include an enhanced version of the SIGPLOT software – which can also be used in conjunction with the original Model Validation Kit.

During the Indianapolis experiment, SF_6 tracer was released from an 84 m power plant stack in the town of Indianapolis, USA. 170 hours of tracer data are available from monitoring arcs at distances ranging from 0.25 to 12 km from the source.

The data set will not be described here in any detail. We refer to the descriptions by Murray and Bowne (1988), by Olesen (1997a), and on the Internet. The Indianapolis data set is an interesting complement to the other data sets of the Model Validation Kit because it represents urban conditions.

The Model Validation Kit and its supplement are available free of charge from the author. Information on the kit can be found on the Internet through the home page of the initiative on *Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes* (http://www.dmu.dk/AtmosphericEnvironment/harmoni.htm).

3 The near-centreline methodology

3.1 Introduction

When evaluating dispersion models, various sets of concentration variables can be considered as being of interest. For instance, such variables are cross-wind integrated concentrations and arcwise maximum concentrations. Both of these variables are considered in the Model Validation Kit (however, in the current version of the kit cross-wind integrated concentrations are not included for the case of Kincaid). There is a possible alternative to the use of observed arcwise maxima, which will be discussed here at some length. It uses the concept of *near-centreline concentrations.*

An evaluation methodology using this concept has been proposed by J. Irwin in the context of the American Society for Testing and Materials (ASTM, 1997). This methodology is also being considered in the context of an

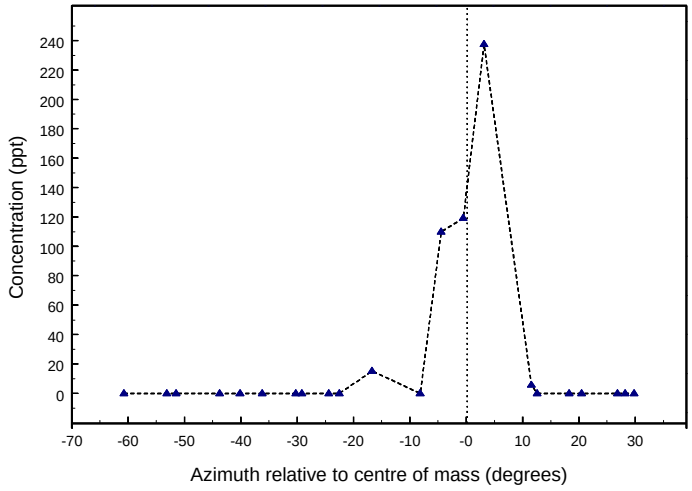

Fig. 2. Kincaid data. Concentrations along the arc 7 km downwind, 10-11 hours May 22, 1981.

ISO ad hoc work group on model evaluation[1]. There is an overlap between the persons involved in the ASTM and the ISO work as well as in the "Harmonisation..." workshops, so results obtained in one of these frameworks are likely to propagate to the others.

The essence of the ASTM methodology is that we should consider not only *one* observed value (the maximum) per arc, but consider several "near-centreline" concentrations. A model's ability to reproduce the near-centreline concentrations should be evaluated by first classifying the experimental data into regimes with similar physical properties, and then assessing statistical performance measures within each regime. In particular, the fractional bias (FB) is considered. Finally, a composite performance measure over many regimes can be constructed.

The remaining part of this paper is devoted to the question:

What are the consequences if we change methodology – from a methodology focusing on maximum arcwise concentrations (MAC's) to a methodology focusing on near-centreline concentrations (NCC's)?

The complete methodology involves a wide range of issues, but we will here restrict ourselves to the most basic questions. The present paper can be regarded as one among many building blocks in a foundation for a decision on recommended model evaluation methodologies.

3.2 Why consider a change?

The background for considering a change in methodology compared to the Model Validation Kit lies in the fact that atmospheric dispersion processes are stochastic.

Models can be expected only to predict ensemble averages – not the results of specific realisations. The Model Validation Kit in its present form does not explicitly address this issue. Anyhow, it has the advantage of being straightforward and practically oriented.

On the other hand, a methodology building on the concept of near-centreline concentrations is better suited to handle the question of ensemble averages. This has a price in terms of increased complexity. Also, the definition of a "perfect model" will change with a change of methodology.

In the following subsections, the procedure for defining near-centreline concentrations and the implications of using them will be explored and contrasted with the procedures of the existing Model Validation Kit.

3.3 Basic concepts – defining an ensemble average

An example of the layout of a dispersion experiment is shown as Figure 1. It displays the geographical distribution of measured ground-level tracer concentrations during one particular hour of the Kincaid experiment. According to the methodology used in the Model Validation Kit, arcs of monitors are considered where all monitors in an arc have approximately the same distance from the source. For each arc, the observed maximum is compared directly to the modelled value for the plume centreline at ground level.

Take as an example the 7 km arc as illustrated in Fig. 2. The maximum value (238) is compared to a modelled value for the plume centre line.

[1]ISO is the International Standards Organisation, and the workgroup is under TC 146/SC5: the subcommittee on Meteorology under the Technical Committee on Air Quality.

In contrast, according to the methodology involving NCC's we should instead take several, near-centreline concentrations (in this case for instance the three highest values) and compare them to modelled values. The details in this will be discussed later, but let us first consider a basic question: *How should we ideally form an ensemble average of arcwise centreline concentrations?* (Under the assumption that we have many realisations of one "event", i.e. a meteorological scenario.)

This is not *a priori* clear. Let us assume that we know all details of the concentration distribution along an arc. We may choose between at least two methods for forming an ensemble average of arcwise centreline concentrations:

(i) The maximum based method: Take the maximum concentration for each arc and form the average

$$\overline{c}_{\max} = \frac{1}{n}\sum_{i=1}^{n} c_{\max,i}$$

(where $c_{max,i}$ is the maximum concentration along the arc for the i'th realisation of the event).

(ii) The centre-of-mass based method: Find the concentration at the centre-of-mass for each arc, and form the average

$$\overline{c}_{c\_o\_m} = \frac{1}{n}\sum_{i=1}^{n} c_{c\_o\_m,i}$$

(where $c_{c\_o\_m,i}$ is the arcwise centre-of-mass concentration of the arc for the i'th realisation of the event).

Method *(i)* fits with an evaluation methodology where we compare observed arcwise maxima to modelled centreline values as we do in the case of the Model Validation Kit. A "perfect model" would be a model which correctly predicts ensemble averages of arcwise maximum values. In practice, we cannot determine the *true* arcwise maximum because our network of monitors has gaps. Strictly speaking, this “receptor spacing effect” makes it impossible for us to assess whether a given model is perfect according to definition *(i)*.

Method *(ii)* corresponds to an evaluation methodology where we identify the centre of mass for each arc and compare the concentration there with the modelled centreline concentration. A "perfect model" is one which correctly predicts ensemble averages of *arcwise centre-of-mass* values.

Some properties of the two types of "perfect models" are trivial from a mathematical standpoint, but very interesting from a practical point of view.

In the first place, if we have a perfect model (according to either method), and if we have a reasonable number of realisations of the same event, then there will exist observed values greater than our model prediction. Thus, *it is a characteristic feature of a perfect model that it underpredicts the highest concentrations*. There is a slight modification: in case of a perfect method-*(i)*-model, the underprediction can to some extent be concealed by the “receptor spacing effect”.

Secondly, a model which is perfect according to the maximum based definition *(i)* will predict larger concentrations than a model which is "perfect" according to the centre-of-mass based definition *(ii)*.

The practical consequences of this will be treated in the subsequent sections.

A discussion along similar lines as above, but with a few additional details can be found in another paper by the author (Olesen, 1997b).

3.4 Basic concepts – preparation of a data set with near-centreline concentrations

The basis of the ASTM methodology, with changes as suggested by Irwin (1998) will be outlined here.

- First, take an experimental data set with a good coverage of samplers along the monitoring arcs.
- Next, classify the observations into regimes. Each regime should represent uniform physical conditions (arcs at the same distance from the source, with the same stability etc.). The definition of a “regime” is important because ensemble averages will be determined regime by regime.
- Consider an arc, and accept it or reject it for further processing. For instance, it will be rejected if there are unacceptably few monitors.
- Determine the centre-of-mass and the lateral dispersion, σ_y, for the arc. Based on all observations in the regime, compute the average σ_y for the regime.
- Go back to the individual arcs and select “near-centre-line” concentrations. On a given arc, there may be none, one, or several selected values. “Near-centreline” is defined in terms of the *regime-averaged* σ_y. The resulting data set with observed concentrations will be the basis for all further work. It can be used by modellers for computing concentrations at corresponding points.

The above steps can be performed once and for all for a given experimental data set, once the set of regimes is defined.

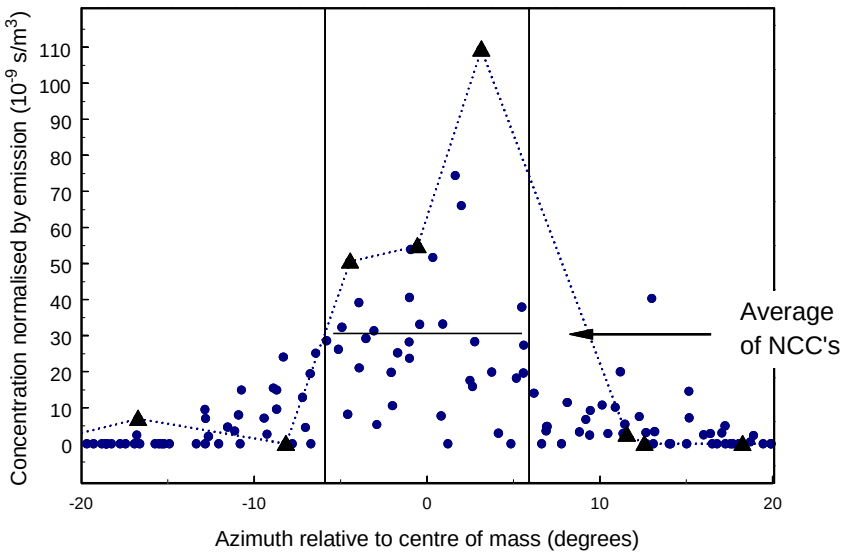

Fig. 3. Kincaid data. Concentration from 14 arcs belonging to a regime with arcs 7 km downwind and -25<z_i/L<0.

Fig. 3 shows an example of all observations in a certain regime at Kincaid as defined by Irwin (personal communication; near-neutral but unstable conditions, -25<zi/L<0; 7 km arc). Observed concentrations from 14 arcs have been grouped together in one regime. The observations from one arc – the one discussed previously (7 km, 11 hour, May 22, 1980) are identified with a line. The concentration units are different from those of Figs 1 and 2; here, concentration is normalised by the emission rate. Observations with azimuth between -0.67 σ_y and 0.67 σ_y (regime-averaged σ_y) are considered near-centreline concentrations; these NCC's lie between the two vertical lines on the figure.

The ASTM methodology goes on to define a set of procedures for statistical treatment of the NCC's. The procedure involves resampling and estimation of the *median* concentration within each regime (original ASTM draft), or resampling and estimation of the *average* concentration within each regime (revised proposal, Irwin and Rosu 1998).

It will be argued here that averages are not a sufficient base for assessing model performance, but that more information should be retained for analyses of model behaviour.

Let us consider the example in Fig. 3. If we retain only the average, a model which predicts a single value of 30 for all 14 arcs would be deemed perfect. This is unsatisfactory from a regulatory point of view. It is frequently a regulatory requirement that models predict high percentiles well. Therefore, if a model shows skill in predicting the high end of the frequency distribution of observations in each regime, this capacity should be acknowledged.

Here, we are faced with a fundamental difficulty. The various realisations in a regime are different, and this *might* be due solely to stochastic variations. But, more likely, these variations are the consequence of an imprecise definition of our regimes. And in the latter case, it would be unfair to rate a model only on its ability to predict *averages* over our imperfectly defined regimes. Therefore, an analysis of averages is necessary, but not sufficient.

3.5 Definition of regimes

Ideally, a regime should be an ensemble of observations which represent several realisations of one "event" – one dispersion scenario as defined by a combination of meteorological conditions, source terms etc. This is hardly attained in practice where a regime must fulfil the following opposing requirements:

1) The regime represents uniform physical conditions:
2) The regime contains enough observations to allow use of various statistical techniques.

The task of defining regimes should in general not be left to individual modellers because that would deprive us of the ability to compare model evaluation results on a common basis. This job – or art – should be trusted to the providers of model evaluation data. As a help in defining regimes, one should inspect frequency distributions of various properties within each regimes and ensure uniformity.

3.6 Consequences of a change of methodology

Model evaluation entails a comparison of observations with model results. If we change methodology – from a MAC-based (maximum arcwise concentrations) to a NCC-based (near-centreline concentrations) – we must change *both* sets of data involved in the comparison:

1) The set of observations. Instead of a set of observed maximum concentrations we must consider a much larger set of near-centreline concentrations, containing many lower values.
2) The set of model results. Instead of a set of computed centreline concentrations we must use a larger set of modelled values corresponding to the near-centreline concentrations. One important note is pertinent here: The ASTM methodology uses a very crude approximation as it is presently implemented. Modelled *centreline* concentrations are used directly in the analyses instead of *near-centreline* concentrations . As it will be shown, this approximation is not reasonable and should be abandoned in future.

If we consider the entire Kincaid data set, and for illustration purposes consider one particular model, namely the Danish OML model (Olesen et al., 1992), we can in gross terms demonstrate the consequences of a change of methodology.

The key question to be considered is the following: *Will a model which behaves reasonably in terms of the MAC methodology also be acceptable according the NCC methodology?*

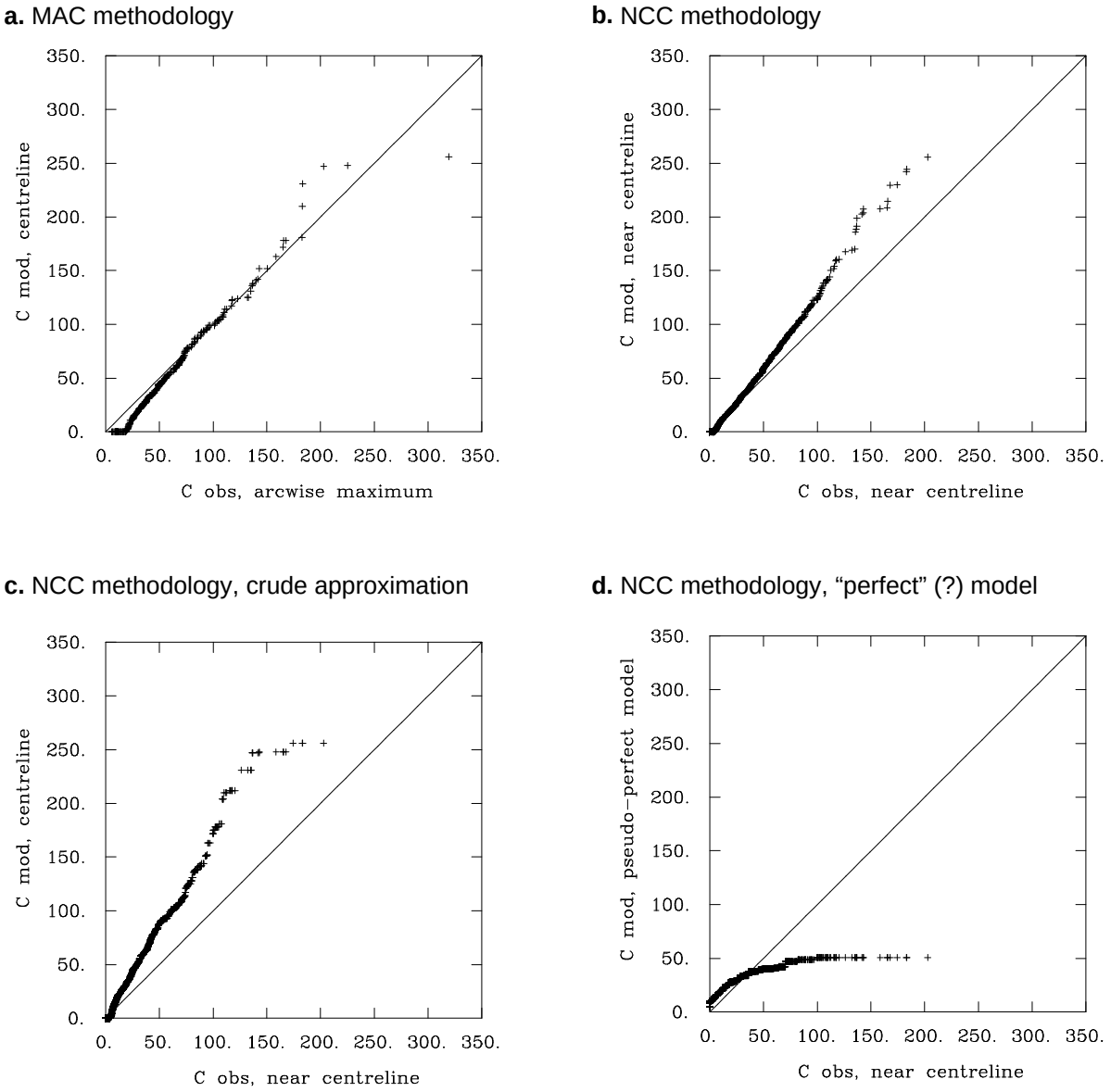

Fig. 4. Kincaid data. Model performance in terms of quantile-quantile plots. Fig. 4a-4c illustrate the performance of the OML model according to various methodologies (see text), while Fig. 4d illustrate the performance of a constructed model which is perfect for predicting regime-wise averages.

Table I Effect of a change in methodology as evidenced by overall averages.

	MAC methodology (Fig. 4a; 338 obs)	NCC methodology (Fig 4b; 1866 obs)	NCC methodology, crude approximation (Fig 4c; 1866 obs)	NCC methodology, pseudo-perfect model (Fig 4d, 1866 obs)
Average of modelled values	47.5	28.5	41.5	25.1
Average of observed values	54.3	25.1	25.1	25.1
Overprediction in percent	-13	14	65	0

The answer can be found in Fig. 4 which is based on Kincaid data. Each panel in Fig. 4 shows a quantile-quantile plot where the distribution of observed and modelled values are compared. The data are ordered by rank, so for instance the highest observed concentration is paired with the highest modelled concentration.

Fig. 4a shows model performance according to the MAC methodology (as in the Model Validation Kit). Only observations of the best quality (quality indicator 3, see Olesen 1995b for details) are considered. The model behaviour is acceptable, although a greater tendency for underpredicting the highest values might be expected from a "perfect model". On the other hand, the “receptor spacing effect” implies that some overprediction should be expected in general.

Fig 4b shows all observations according to the NCC methodology. Here, 1866 concentration values are represented (whereas there were only 338 concentration values in Fig 4a). The modelled values have been computed for points corresponding to the observed concentrations – i.e., typically at off-centreline points.

A comparison of 4a and 4b reveals the consequences of a change of methodology. With changed methodology, a model will have an increased tendency towards overprediction. The same point can be seen by taking the simple average of all the numbers underlying Fig 4, as it is done in Table I. The overall model bias changes from underprediction to overprediction when going from a MAC methodology to a NCC methodology.

Fig 4c illustrates the consequences of using the crude approximation which is applied in the present ASTM draft code. Instead of using modelled values computed at the correct off-centreline positions, modelled centreline values are used. The effect of this approximation is certainly too severe to be neglected.

Fig 4d serves to illustrate how much information we loose if we focus solely on average values within each regime. If we pursue this line of thought, we can easily construct an artificial "perfect model": We define a "perfect" model which predicts only one value for each regime, namely the average value. Although "perfect", such a model would be characterised by the frequency distribution displayed in Fig 4d. From a common-sense point of view, such a model would definitely not be perfect. Therefore, we should not let averages be the only success criterion for models.

3.7 Unsettled problems

It is a characteristic feature of a “perfect model” that it underpredicts the highest concentrations. The most fundamental problem concerning a change of methodology is whether the regulatory community will accept a new notion of a perfect model, such that a perfect model underestimates the highest concentrations more severely than previously. This is the price to be paid for a more consistent framework which can better cope with the problems of inherent uncertainty.

Besides this main issue, there are many other questions to be resolved before the NCC methodology can find widespread use as a well-defined, common standard. Such problems include:

- If averages should not be the only success criterion, then what should be added?
- How should regimes be defined for the various datasets? Regimes must be defined with due respect for the characteristics of each dataset.
- Do the proposed methods for selecting arcs and determining centre-of-mass work satisfactorily?
- There are various options concerning resampling and statistical treatment of data. These must be settled.
- At various points in the methodology, some more or less arbitrary choices have been made (such as defining “near-centreline” as being within $\pm 0.67\ \sigma_y$). Is the outcome of the methodology sensitive to any of these choices?
- The presently available data sets are not thoroughly checked (they contain duplicate values and outliers) and are not in a very convenient format.
- The available software tools are not yet well-documented nor easy to apply.

Anybody interested can contribute in exploring this set of issues. The relevant software and data sets are available in draft form through the Internet (see http://www.dmu.dk/AtmosphericEnvironment/harmoni.htm for

further details). Note, however, that these tools have not yet reached maturity. Over the coming years, the tools will probably develop, so they eventually provide a well-defined common frame of reference. Information on their current status can be found on the Internet.

4 Conclusions

A change of methodology of the Model Validation Kit is under consideration. The proposed new methodology has attractions, but also poses many problems. As yet, many questions are not settled, so modellers who wish to use a "common currency" when evaluating their models should use the established Model Validation Kit.

On the other hand, in order for the new methodology to reach maturity, it should be tested by researchers willing to put an effort into some pioneering work.

This paper reports some results concerning the "exchange rate" between the two methodologies which can provide a background for decisions on standard methodologies.

Acknowledgments

The author wishes to acknowledge John Irwin, who has been a valuable source of inspiration. Thanks are also due to Steve Hanna who has contributed in numerous ways to the work with the Model Validation Kit.

References

Hanna, S.R., Strimaitis, D.G. and Chang, J.C. (1991), 'Hazard Response Modeling Uncertainty (a Quantitative Method). Vol. I: User's Guide for Software for Evaluating Hazardous Gas Dispersion Models'. Sigma Research Corporation, Westford, Ma.

IJEP (1995), 'The Workshop on Operational Short-range Atmospheric Dispersion Models for Environmental Impact Assessment in Europe. 21-24 November 1994, Mol. Belgium', *Int. J. Environment and Pollution*, Vol. 5, Nos. 4-6.

IJEP (1997), '4th Workshop on Harmonisation within Dispersion Modelling for Regulatory Purposes. 6-9 May, 1996, Oostende, Belgium'. *Int. J. Environment and Pollution*, Vol. 8, Nos. 3-6.

Irwin, J.S. (1997), 'Standard Practice for Statistical Evaluation of Atmospheric Dispersion Models'. Draft. ASTM Designation Z6849Z. ASTM, 100 Barr Harbor Drive, West Conshohocken, PA 19428-2959, USA.

Irwin, J. and Rosu, M-R. (1998), 'Comments on a Draft Practice for Statistical Evaluation of Atmospheric Dispersion Models', Proceedings of the 10th Joint Conference on the Applications of Air Pollution Meteorology. American Meteorological Society, Boston, pp. 6-10.

Moussiopoulos, N., de Leeuw, F., Karatzas, K. and Bassoukos, A. (1998), 'The Air Quality Model Documentation System of the European Environment Agency'. This set of pre-prints.

Murray, D.R. and Bowne, N.E. (1988), 'Urban Power Plant Plume Studies'. EPRI report EA-5468, available from EPRI, 3412 Hillview Ave., Palo Alto, CA 94304, USA.

Olesen, H.R., Løfstrøm, P., Berkowicz, R. and Jensen, A.B. (1992), 'An Improved Dispersion Model for Regulatory Use - the OML Model'. *In:* Air Pollution Modeling and its Application IX, edited by H. van Dop and G. Kallos, Plenum Press, New York, 1992

Olesen, H.R. (1994), 'European Coordinating Activities Concerning Local-Scale Regulatory Models'. *In:* "Air Pollution Modeling and Its Application X", Plenum Press, New York.

Olesen, H.R. (1995a), 'Data Sets and Protocol for Model Validation'. Workshop on Operational Short-range Atmospheric Dispersion Models for Environmental Impact Assessment in Europe, Mol, Belgium, Nov. 1994, *Int. J. Environment and Pollution*, Vol. 5, Nos. 4-6, 693-701.

Olesen, H.R. (1995b), 'The Model Validation Exercise at Mol. Overview of Results'. Workshop on Operational Short-range Atmospheric Dispersion Models for Environmental Impact Assessment in Europe, Mol, Belgium, Nov. 1994, *Int. J. Environment and Pollution*, Vol. 5, Nos. 4-6, pp. 761-784.

Olesen, H.R. (1996), 'Toward the Establishment of a Common Framework for Model Evaluation', *in:* Air Pollution Modeling and Its Application XI, pp. 519-528. Eds. S-E. Gryning and F. Schiermeier, Plenum Press, New York.

Olesen, H.R. (1997a), 'Pilot study: Extension of the Model Validation Kit'. 4th workshop on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Oostende, Belgium, 6-9 May, 1996. *Int. J. Environment and Pollution*, Vol. 8, Nos. 3-6, pp. 378-387.

Olesen, H.R. (1997b), 'Tools for Model Evaluation'. Paper presented at the 22 NATO/CCMS International Technical Meeting on Air Pollution and Its Application, June 2-6, 1997, Clermont-Ferrand, France. To appear in: Air Pollution Modeling and Its Application XII. Edited by S-E. Gryning and N. Chaumerliac, Plenum Press, New York.