MODEL VALIDATION KIT - RECENT DEVELOPMENTS

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 ground is presently being prepared for further development of the kit. Based on data from the Kincaid experiment, investigations have been carried out in order to illuminate features and problems with a methodology for model evaluation proposed in the context of the ASTM. Certain problems with the determination and use of so-called near-centreline concentrations (NCC's) are identified and discussed. A study of crosswind integrated concentrations from the Kincaid experiment has been conducted; as a by-product, the study has revealed data problems with a certain version of the Kincaid data set.

KEYWORDS
Model Validation Kit, Kincaid, atmospheric dispersion models, model evaluation, ASTM, near-centreline concentrations, crosswind integrated concentrations.

INTRODUCTION

One of the key elements of the work of the initiative on *Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes* is model evaluation.

The so-called *Model Validation Kit* was created for the second "Harmonisation..." workshop in 1993, and since then it has undergone gradual improvements. The Model Validation Kit has formed the basis for much work on model evaluation, and it has been requested by 160 modelling groups over the past few years.

The Model Validation Kit provides a simple way of evaluating model results against experimental data, where model results are compared *directly* against observations; various statistical and graphical analyses are easy to undertake with the tools of the kit. These tools include a package of model evaluation software developed by Hanna et al. (1991) which is of a flexible and general nature.

The kit comprises data sets where a single source emits a tracer gas while arcs of monitors are positioned downwind of the source. The data to be compared are arc-wise maximum concentrations and (for some of the data sets) crosswind integrated concentrations along arcs. The kit has been described by Olesen (1995) and is available from the author. More information is available through the Internet
(http://www.dmu.dk/AtmosphericEnvironment/harmoni.htm).

There is a basic conceptual problem with the procedure of directly comparing arcwise maxima to modelled maxima, and as a consequence the results should be interpreted with care. The problem is that even a "perfect" model cannot be expected to give the same frequency distribution of arcwise maxima as the one observed. If the monitoring network is sufficiently dense and the data represent a sufficient number of scenarios, *it must be expected that a "perfect model" will underpredict the highest observed concentrations*. This is because atmospheric dispersion is a

stochastic process, and 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. However, it has the advantage of being straightforward and practically oriented. When the Model Validation Kit is used, one must interpret the results from it with care, keeping in mind that for a good model underprediction of the highest concentrations is to be expected. In particular, the so-called quantile-quantile plots from an entire experimental database should not stand alone as the result from a model evaluation. A very useful supplement are the so-called residual plots which provide more insight into model behaviour.

Work is in progress, which addresses the challenge of the stochastic nature of the atmosphere in a more explicit fashion than the current kit. A methodology is under development where *ensembles* of observed values are considered rather than individual values. This work may eventually result in tools which can be widely used by the modelling community and which can be incorporated in a future version of the Model Validation Kit.

However, we are not yet at a stage where a well-established set of tools accompanied by "authoritative" data sets can be distributed.

The focus of the ongoing work is on a draft for an ASTM standard guide on statistical evaluation of atmospheric dispersion models (ASTM Designation Z6849Z. Draft, April 1999[1]). The contents of this procedure have been described at earlier Harmonisation conferences (e.g. Irwin, 1998; Olesen, 1998) and will be briefly recapitulated in the next section.

In the remainder of the present paper, various aspects of the ASTM methodology will be explored. There are certain problems with the methodology and its current implementation, implying, i.a., that the procedure will favour models, which give too low concentration values.

THE ASTM PROCEDURE

The ASTM package is based on a philosophy where model results are compared not against individual observations, but against some form of ensemble average. This is in accordance with the assumption that a model normally is not capable of reproducing stochastic variations, only ensemble averages.

Conceptually, the philosophy is appealing, but implementing it in the form of an operational procedure entails a number of problems.

The present paper discusses such problems, illustrated by examples. Users of the methodology should be aware of these problems, because otherwise they may become misled by the results of the procedure.

When using the ASTM procedure one can choose between several concentration parameters as the subject for study. For instance, the so-called “near-centerline concentrations” (NCC’s) can be considered. Optionally, one can use other concentration parameters, such as crosswind integrated concentrations.

Until now, use of the ASTM procedure in conjunction with NCC’s has received the most thorough treatment. Software and procedures, which deal with this parameter have been made generally available.

After a general introduction to the ASTM methodology here, we will focus on NCC’s and on crosswind integrated concentrations in following two sections.

The steps in the ASTM procedure can be outlined as follows. We assume for the time being that the concentration variable of interest is NCC’s.

- 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.

[1] A review copy of the draft Guide is available from J. Irwin on request. Related software and datasets can be found at http://www.dmu.dk/atmosphericenvironment/Harmoni/ASTM_key.htm

- 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 for the arc. Based on all observations in the regime, compute the lateral dispersion, σ_y, for the regime.
- Go back to the individual arcs and select "near-centre-line" concentrations. "Near-centreline" is defined in terms of the *regime-averaged* σ_y. Thereby, create a data set with observed concentrations. This will be the basis for all further work.

The above steps concern observed data. Model results corresponding to observations can now be computed and grouped regime by regime. It is a subject open to discussion exactly how model results "corresponding" to observations should be determined (see later for details).

When the groups of observed and modelled data have been prepared, one can go on and consider the absolute fractional bias for each regime

$$AFB = \frac{2\left|\overline{c}_m - \overline{c}_{obs}\right|}{\overline{c}_m + \overline{c}_{obs}} \qquad \textbf{(1)}$$

where c_m represent modelled concentrations, c_{obs} observed concentrations, and an overbar means averaging within a regime.

Actually, according to the ASTM methodology, a resampling procedure is used so that the concentrations in Eq. (1) are selected by conducting a bootstrap sampling among the observed and modelled values in the current regime. Thus, there may be, e.g., 500 AFB values for each regime. Subsequently, the AFB's for the various regimes are processed in order to compute a composite index, a "Model Comparison Measure". Here, we will not be concerned with the statistical processing of data, but focus on the underlying data that enter Eq. (1). These data are crucial as to whether the entire procedure makes sense.

A number of points of concern with the ASTM methodology have emerged since it was first proposed and some have been discussed at previous harmonisation conferences and elsewhere (e.g. Irwin and Rosu, 1998; Olesen, 1998). Points of concern include the following:

- A fundamental difficulty concerns the definition of regimes. Ideally, ensembles are groups of observations collected under identical external conditions, so that any variations are due to stochastic fluctuations. But in practice when the ASTM methodology is used, ensembles are composed of all observations in a "regime", which represents a *range of* external conditions, not just one set of conditions. Thus, scenarios may be grouped together although they have quite distinct characteristics.
  This fundamental difficulty may have undesirable implications. Regimes as defined from real-life experiments must necessarily be rather "broad". As explained in Olesen (1998), a judgement based on averages only (this is equivalent to considering fractional bias only) does not always provide a fair distinction between models.
  The question of sensitivity to regimes deserves exploration. In some cases it would be alarming if a change of regime definitions leads to changes in evaluation results. On the other hand, sometimes it is perfectly understandable and desirable if we get a better distinction between a poorer and a better model when we improve the resolution of regimes.
- Problems with use of the NCC methodology:
  a) The near-centreline concentrations selected for evaluation will inevitably be "contaminated" by off-centerline values. If NCC's are taken at face value, models that give too low results will be favoured. This will be discussed in a separate section.
  b) How are model results "corresponding" to observations determined? In the current implementation of the methodology, a crude method is used, whereby model results for the *centerline itself* are compared to observed *near-centerline* values. Also this will favour models which predict too low values. The problem will be discussed later.

Some additional problems are in principle trivial, but nevertheless act as obstacles to a more widespread use of the ASTM methodology.

- There are problems with the particular version of the Kincaid data set that is presently distributed with the ASTM procedure.
- The currently available software tools are not user-friendly.

The following sections discuss problems related to the NCC methodology; furthermore, some results are reported from studies of crosswind integrated concentration, which are of relevance to the ASTM methodology.

NEAR-CENTRELINE CONCENTRATIONS

Contamination with off-centerline values

Fig. 1 show two examples of crosswind profiles for the Kincaid experiment, which will be used to illustrate some problems, related to determination and use of NCC's.

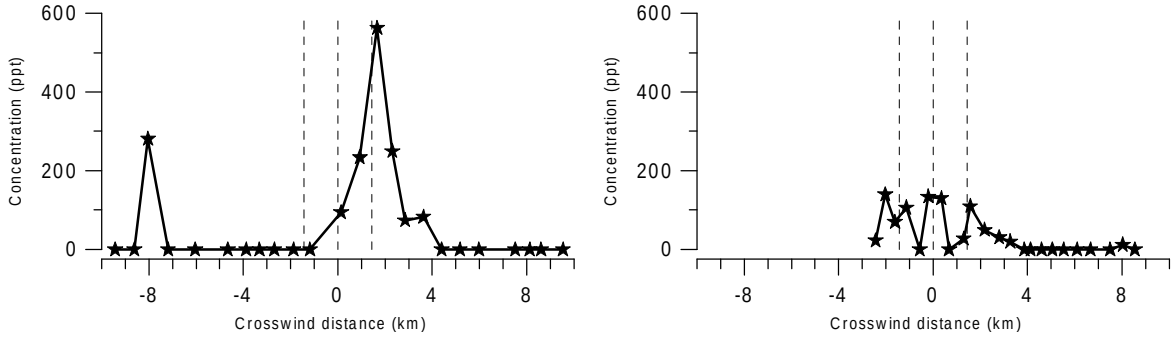

Fig 1. Crosswind profiles at a distance of 7 km observed during the Kincaid experiment in 1980.

Fig. 1a (left) is a case from July 13, 13 hours; Fig 1 b (right) from July 25, 15 hours.

Both examples represent arcs at a distance of 7 km. The ASTM procedure has been used to process the data, and the arcs have been classified according to a proposed set of regime definitions; they belong to the same regime (z_i/L between -1000 and -150).

According to the ASTM methodology, observations which are no further than $0.67\sigma_y$ (regime-wise σ_y) from the centre-of-mass of the arc, are classified as NCC's. The limits for acceptance are indicated on the figures ($\pm$1432 m).

Given a real scenario, it is not always obvious what one should expect from a "perfect model" and what NCC's are. But in the case represented in Fig 1a, from a common-sense viewpoint, it would be reasonable to classify the three largest concentrations (560, 250 and 230 ppt) as being near-centerline. Thus, the "common-sense" average NCC for this case alone is 350 ppt. However, following the proposed procedure strictly, three rather low values are selected as NCC's, resulting in an average NCC for the regime at approximately 110 ppt.

Similarly, in the case depicted in Fig. 1b, a common-sense determination of NCC's would result in two values of approximately 130 ppt, whereas the proposed procedure selects 6 values as NCC's, resulting in an average of approximately 65 ppt.

If the two cases in Fig.1a and 1b were the only ones in the regime, then according to the ASTM methodology a "perfect model" would be a model with a centerline concentration of 80 ppt - far lower than common sense would deem reasonable.

The point I wish to make by these two examples is that there are cases where the near-centreline concentrations are "contaminated" by observations which are off-centerline. Any "contamination" which takes place will lead to inclusion of lower values, so that our NCC's will have a systematic bias. When we use the NCC's for comparison with model results, the comparison will favour models that give too low results.

In my view this is a serious problem with the NCC methodology. By refining the criteria for selecting NCC's we may alleviate the problem, but to some extent it seems inevitable with the NCC methodology

Pairing of model results with "corresponding" NCC's

As currently implemented the ASTM draft applies a rather crude approximation when determining which model results correspond to NCC's. Modelled *centerline* concentrations are compared to observed *near*-centerline concentrations, thereby introducing systematic bias. This bias can be significant. Olesen (1998) presented some investigations related to this question.

The principles can be explained as follows: Let us consider two versions of the methodology with the use of near-centreline concentrations: an accurate method where observations are compared to modelled values at the corresponding, off-centerline points, and a "crude" method where observations are simply compared to modelled values at the plume centre line.

At first sight the expected average ratio between an accurate "NCC methodology" and the "crude approximation NCC methodology" would be seem to be around 0.9 (if we have a Gaussian plume with a centerline concentration of 1, the average within $\pm 0.67\sigma_y$ is 0.9).

However, looking closer into things it becomes clear that one should expect a lower ratio. *How much* lower is an open question. The results from Olesen (1998) indicate a value of 0.69, but this value is model-dependent.

The key to the details of the issue resides in the fact that there is a difference between the σ_y *for a regime* and σ_y for the individual cases.

Assume that we have a very simple regime with only two cases, one with a very wide plume (case A), and the other with a very narrow plume (case B). The ASTM methodology does not preclude that two such cases end up in the same regime.

It is the *average* σ_y for the regime, which is used for selection of "near-centreline concentrations". This is fair enough for case A (the wide plume), but not for case B (the narrow plume).

Assume now that we have a model, which perfectly reproduces both cases. For the model results corresponding to case A there will only be a small difference between modelled values at the centerline and modelled values at off-centerline points. Hence, as long as we consider only case A observations, it doesn't matter whether we use the crude or the accurate method. However, for model results corresponding to case B there may be a large difference between the modelled centerline value and the computed concentrations at the off-centerline (so-called "near-centreline") points, because the points are actually *not* near the centerline. So for case B observations, the ratio between the accurate and the crude method may be *far* lower than 1. Therefore it is *not* correct to expect an average ratio of 0.9 between the "accurate NCC methodology" and the "crude approximation NCC methodology".

CROSS-WIND INTEGRATED CONCENTRATIONS

The ASTM procedure is not restricted to considering NCC's, but can also be used for other concentration variables, such as crosswind integrated concentrations.

A study has been conducted where crosswind integrated concentrations were derived from the Kincaid data set. The original intention with the study was to provide a set of quality-controlled values, which could be used in conjunction with the ASTM package or be used independently. Although the current ASTM procedure does provide values for crosswind integrated concentrations, these are determined by an automatic (blind) procedure, and a check of these values was considered appropriate.

The study, however, revealed that there are inconsistencies between various versions of the Kincaid data set. The particular version of the data set that is being distributed as part of the ASTM-related package suffers from a problem with the definition of arcs which will be described in more detail in the subsequent section. The implication is that exercises with the ASTM package on its accompanying Kincaid data set should be regarded merely as *exercises,* and their results not considered trustworthy. The data problems do not affect the Kincaid data in the current Model Validation Kit.

Due to the data problems, an authoritative data set for Kincaid with quality-controlled crosswind integrated concentrations has not yet been prepared and the results indicated below are only preliminary. One lesson learned from the work with quality control of the crosswind integrated concentration is that a manual quality control of crosswind integrated concentrations is highly desirable.

In the present Model Validation Kit, arcwise maxima have been determined by manual inspection, and a quality indicator assigned to each value. The quality indicator is a number between 0 and 3,

and it indicates how reliable the arcwise maximum is. 338 arcs were assigned the highest quality of 3.

A similar approach was attempted with the crosswind integrated concentrations. The arcs were graded, so that a complete arc (with zero values at the edges and with a reasonably dense coverage of monitors along the arc) was assigned a quality indicator of 3.

Based on preliminary results 661 arcs fulfilled the minimum requirements for the ASTM procedure to attempt an arcwise integration. After a manual inspection 387 (60%) of these were classified as being of highest quality in respect to crosswind integrated concentration.

There is not a one-to-one correspondence between high-quality arcwise maxima and high-quality crosswind integrated concentrations. 65% of the observations with well-defined arc-wise maxima were also considered complete enough to warrant a reliable crosswind integrated concentration. Conversely, 35% were not considered pertinent for determination of crosswind integrated concentration - this reflects the fact that it does not necessarily require a complete arc to determine a maximum value.

Details on the work of determining crosswind integrated concentrations can be found on the Internet (http://www.dmu.dk/atmosphericenvironment/Harmoni/ASTM_key.htm).

KINCAID: DATA PROBLEMS

Potential users of the version of the Kincaid data set, which accompanies the ASTM software, should be aware of a problem with the definition of arcs.

A number of ideal arcs have been defined as circles with centre at the source and radii 1 km, 2 km, 3 km, 5 km etc. During work with the data it became clear that this version of the Kincaid data set is subject to an erroneous assignment of monitors to arcs. For instance, many monitors have been included in the 3-km arc, although they are at a distance of 5 km. This means that misleading information is present concerning the 3-km arc, and that the information concerning the 5-km arc is incomplete. The problem is most noticeable at the 3 and 5-km arcs, but definition of arcs in the entire data set ought to be re-examined before results based on it are used for serious evaluation of models.

CONCLUSION

The existing Model Validation Kit provides a simple way of evaluating model results against experimental data, but its results should be used with caution because it does not explicitly account for the fact that atmospheric processes are stochastic. One should therefore expect that even with a “perfect model”, the model would underpredict the highest observed concentrations. The proposed ASTM procedure is intended to address the problem of ensemble averaging. Optionally, one can consider one of several concentration variables as the basis for the procedure. The so-called NCC’s (near-centreline concentrations) have received particular attention.

Studies of the properties of NCC’s as determined and used by the present implementation of the ASTM procedure point to the problem that the procedure will favour models, which give too low results.

Part of the problem is due to the fact that NCC’s are “contaminated” by low values and part of it due to the fact that - as currently implemented - a crude approach is used where modelled *centerline* values are compared against observed *near-centreline* concentrations that are typically *off-centerline* values.

Studies have been conducted of crosswind integrated concentrations for the Kincaid data set. The crosswind integrated concentration is one of the concentration variables that can be used in conjunction with the ASTM procedure. These studies revealed problems with data in one particular version of the Kincaid data set (the one distributed together with the ASTM package, not the one in the Model Validation Kit). The implication is that exercises with the ASTM package on its accompanying Kincaid data set should be regarded merely as *exercises,* and their results not considered trustworthy. So far, no authoritative dataset with crosswind integrated concentrations has been produced. The study confirmed that a manual quality control of crosswind integrated concentrations is highly desirable.

Development of the ASTM methodology should be continued, but with due attention to the problems identified here.

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.

Irwin, J. & 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.

Irwin, J.S. (1998), 'Statistical Evaluation of Atmospheric Dispersion Models'. Presented at the 5th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purpose in Rhodes, May 1998. To be published in IJEP.

Olesen, H.R. (1995), '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., 1998, 'Model Validation Kit - Status and Outlook.' Presented at the 5th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purpose in Rhodes, May 1998. To be published in the IJEP.