Manuscript for
Air Pollution Modeling and Its Application XI, pp. 519-528. Edited by S-E. Gryning and F. Schiermeier (1996), Plenum Press, New York

TOWARD THE ESTABLISHMENT OF A COMMON FRAMEWORK FOR MODEL EVALUATION

H.R. Olesen

National Environmental Research Institute
P. O. Box 358
DK-4000 Roskilde
Denmark

1. INTRODUCTION

In this paper, a presentation will be given on current work toward the establishment of a common frame of reference for the evaluation of atmospheric dispersion models. The focus will be on issues within local-scale regulatory modelling.

First, the motivation for trying to establish a common framework will be discussed: What is unsatisfactory about the way that model evaluation is currently being practised?

Next, in Section 3, the actors on the scene will be presented. Work is being conducted internationally within the framework of several organisations in order to organise a common basis for model evaluation; some of these activities will be described.

An important activity in this respect is that of the *Initiative on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes*. Under this initiative, a series of workshops have been held on harmonisation of models. Model evaluation has been a key issue. In Section 4 some important experiences gained through this work will be discussed.

The concluding Section 5 proposes some responses to the question: What can be done better?

2. WHAT IS UNSATISFACTORY?

Model evaluation studies abound in the technical literature. Still, it is not easy for a decision-maker to determine whether a particular model is a good model, and whether it fits his particular purpose.

A fundamental problem in model evaluation practice can be illustrated by the following constructed example:

Let us assume that in the year 1990, a particular research group has received funding to evaluate models A, B and C, on the basis of data sets X and Y. The evaluation is performed. Some years later, model A is revised. However, the original model evaluation is not repeated. And when one day model D appears, the developers of model D do not have access to the data sets and software used in the 1990 evaluation. So nobody is able to make a qualified comparison of the revised model A and model D with the earlier models.

The basic problem is that, until now, a common frame of reference for model evaluation has to a large extent been lacking. Tools that would make it easier for modellers to adhere to standards have been few. Those that do exist have been unknown to most of the modelling community.

The problem is severe, because model evaluation is intrinsically difficult. In the model evaluation business, one often meets statements like the following:

"For model A, the fractional bias is 0.27 when using data from experiment X."

Such a statement by itself is totally inadequate as basis for a judgement about a model. For a serious judgement, one needs to know a lot of details. What data sets were used? What information is contained

in the data sets, and what information is missing? Were some of the original data discarded? On what criteria? Which concentration measures were considered (e.g. area-wise maxima, arc-wise maxima etc.)? What kind of averaging was performed? What processing of data took place?

Such information is usually included in the documentation from a model evaluation, but only to a limited extent. There are so many details involved that it is very difficult to get a precise picture of the abilities of a model. Only with performance evaluation results from *several models* that are formed on *exactly the same basis,* can one draw any conclusions about model performance. This is the background that there is clearly a need for "something to be done".

3. ACTORS ON THE SCENE

Over the years, several well-organised model evaluation exercises have been conducted, e.g. the work during the EPRI Plume Model Development and Validation project (e.g. Bowne et al., 1983); the DOE/AMS Air Pollution Model Evaluation Workshop in 1984 (e.g. Kurzeja and Weber, 1985); a series of activities of the US EPA (e.g. Lee, 1993); the ATMES study following the Chernobyl accident (Klug et al., 1992); the APSIS exercise concerning the Athenian photochemical smog (e.g. a series of papers in the proceedings of the previous ITM; Gryning and Millan, 1994), and the comprehensive hazardous gas model evaluation reported by Hanna et al. (1993). The list does not pretend to be complete, especially not regarding long range modelling, but the studies mentioned are among the "classics".

Limits in past computer technology have restricted the portability of software from these studies. In most cases these studies have been stand-alone exercises with no follow-ups. However, the studies do represent a valuable resource of information that should still be exploited.

Presently, there are several initiatives involved in formulating procedures for model evaluation, the activities of which shall be described in the following subsections:

- the *Initiative on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes*;
- the *European Topic Centre on Air Quality* within the framework of the European Environment Agency;
- the *Model Evaluation Group* under the EC DG XII research programme on major industrial hazards.

Conclusively, in Subsection 3.4, a few other initiatives will be briefly mentioned.

3.1 The Initiative on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes

During the past few years, a "new generation" of models has emerged with physically more justifiable parametrisations of dispersion processes than previously. A need has been felt for these new models to be developed in a well-organised manner and turned into practical, generally accepted tools fit for the various needs of decision-makers. Therefore, in 1991, a European initiative was launched for increased cooperation and standardisation of atmospheric dispersion models for regulatory purposes: the *Initiative on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes*. This initiative has organised a series of workshops to promote the use of new-generation models within atmospheric dispersion modelling, and in general improve "modelling culture". Model evaluation is a key issue in this respect.

Until now, three workshops have been held (see Olesen and Mikkelsen, 1992; Cuvelier, 1994), while a fourth will take place in May 1996. An activity closely related to the workshops is the distribution of the so-called *Model Validation Kit*. This kit was originally prepared for the second of the workshops (in Manno, Switzerland, 1993). It was revised and used again for the third workshop (in Mol, Belgium, 1994). In total it has been distributed to about 80 research groups.

The *Model Validation Kit* is a collection of three experimental data sets accompanied by software for model evaluation. It is a practical tool, meant to serve as a common frame of reference for modellers. The exercises based on the kit are limited in scope (only short-range models considered, homogeneous terrain, only three data sets, etc.). It has not been attempted to perform an in-depth investigation of models; rather, the work with the kit has served as a demonstration exercise. Many useful experiences have been gained through the use of the kit. These will be discussed in Section 4.

It should be noted that the work is continuously evolving. Up-to-date information on the workshops as well as practical tools are available through the World Wide Web (see Subsection 3.5).

3.2 The European Topic Centre on Air Quality

The European Environment Agency (EEA) is an institution under the European Union. The mandate of the EEA is to provide the member states with information on the environment. The technical tasks of the EEA are often delegated to the so-called *European topic centres*. The *Topic Centre on Air Quality* is a consortium led by the Dutch RIVM.

The 1995 work plan of the *Topic Centre on Air Quality* includes the project *MA3: Harmonisation in the use of models for ambient air quality and pollution dispersion/transport.* This project has as its objective: "To increase consistency of models already in use; To develop guidance on criteria for selecting appropriate models and their application for the assessment and management of air quality". The project includes several subtasks, one of which is to establish a documentation centre and toolkits for testing models. The work is conducted in cooperation with the above mentioned *Initiative on Harmonisation*.

3.3 Model Evaluation Group

The EC DG XII (the European Commission Directorate-General for Science, Research and Development) coordinates research programmes within the EC. Within the *Major Industrial Hazards* programme, efforts are in progress to set up a more systematic framework for model evaluation. A *Model Evaluation Group* (MEG) has devised a protocol for model evaluation (Model Evaluation Group, 1994). This protocol is generic, i.e. it is a framework to be filled in for several different kinds of technical models (not just dispersion models). An ongoing activity in this framework is an evaluation of models for dense gas dispersion (Duijm, 1995).

3.4 Others

In May 1995, the council of the British Royal Meteorological Society issued a policy statement titled *Atmospheric dispersion modelling: Guidelines on the justification of choice and use of models, and the communication and reporting of results* (Royal Meteorological Society, 1995). The aim is to promote the best practice in the use of mathematical modelling for atmospheric dispersion.

The ASTM (American Society for Testing and Materials) is considering to establish standards relating to performance evaluation of atmospheric dispersion models. The ASTM is an influential, voluntary standards development organisation which is open to members from all over the world. At an ASTM workshop held in July 1995 titled *Performance evaluation of atmospheric dispersion models* it was found that it would be both possible and useful to develop some standards for performance evaluation in cooperation with similar European initiatives.

3.5 Future developments

The activities mentioned above are *not* intended to be competing with another. They are all inspired by the same problems, and they can be seen as a an attempt to sow the same kind of seed in various contexts.

Funding is a general problem for all the activities mentioned. Much work is done on a voluntary basis, so progress must be expected to be slow.

A useful point of entry to up-to-date information on these issues is the WWW pages of the *Initiative on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes* (http://www.dmu.dk/AtmosphericEnvironment/harmoni.htm).

4. THE MODEL VALIDATION KIT: EXPERIENCES GAINED

This section sums up some experiences gained through work with the *Model Validation Kit*. Through examples, a few essential problems will be highlighted, and it will be discussed how they can be overcome.

The scenario considered by the *Model Validation Kit* is relatively simple, namely one where a single source emits a non-reactive gas in homogeneous terrain. The kit includes data sets from three atmospheric field experiments (Kincaid, USA; Copenhagen, Denmark; Lillestrøm, Norway). The data sets are accompanied by a package of model evaluation software. The software was essentially developed by Steve Hanna and his colleagues for use with hazardous gas dispersion models (Hanna et al., 1991).

In the subsequent sections, some results based on Kincaid data will be presented. During the Kincaid experiment, the network of tracer monitors was dense; it has therefore been possible to determine maximum concentrations for crosswind arcs of monitors. Such maximum arc-wise concentrations have been compared to the computed plume centre line concentration at ground level at the same distance.

More information on the *Model Validation Kit* can be found in Olesen (1995a). The results of applying the kit on five models is discussed in depth in Olesen, 1995b.

In a paper for the previous ITM, the author analyzed a number of difficulties connected to model evaluation, and listed pertinent reactions to them (Olesen, 1994). A slightly revised version of this list of difficulties looks as follows:

Why is model evaluation difficult?

- The appropriate evaluation method cannot be uniquely defined.
- Input data sets are limited - they reflect only few of the possible scenarios.
- Processing of input data for validation is far from trivial.
- The luxury of independent data sets can rarely be afforded.
- There are inherent uncertainties.

Selected items from this list will be discussed below.

4.1 The appropriate evaluation method cannot be uniquely defined

It is important to recognise that there will never be just *one* recommended method for validating models. Just as models should be fit for purpose, so should evaluation methods. When one wants to make a judgement concerning a model, one should be convinced that the tests undertaken really correspond to the questions which should be answered.

Not only do the relevant evaluation methods depend on the context of the application; they also depend on the data sets available. There is a difference between the tests one can undertake, depending on whether one has a dense set of monitors or a sparse net.

A useful reaction to the difficulty of choosing a relevant evaluation method is to *develop an array of methods to be used in various contexts.*

4.2 Input data sets are limited - they reflect only few of the possible scenarios

One fundamental difficulty with model evaluation is that experimental data sets are limited in several respects. Usually, a data set is from a campaign with only *one* source configuration, *one* set of terrain conditions, and a limited number of meteorological scenarios. Further, the number of points where the concentration has been measured is small.

In contrast, users want models that can be used for a broad range of source configurations, for a large number of meteorological scenarios, and that can compute concentrations at all points in space. A natural, and almost inevitable reaction to this dilemma is to *extrapolate model behaviour* to conditions where the model has not been validated. This implies a problem: the extrapolation will only be credible if the model has a sound description of physical processes. Thus, the validation process should be used to *develop an understanding* of model behaviour. This implies in turn that a *diagnostic approach* to model evaluation is indispensable.

In practice, a very useful approach is to stratify data according to physical parameters, and present residuals graphically. Such plots are very easy to produce with the *Model Validation Kit*. An example of a graph showing residuals is reproduced as Figure 1 (see Olesen, 1995 for full details). Some explanation is necessary.

Arc-wise maximum concentration values have been determined from tracer data, and have been predicted by models A and B. These pairs of data, (c_{obs}, c_{mod}), have been stratified according to distance from the source. Seven subsets have been formed. E.g., there is one subset corresponding to distances between 0 and 2 km. It contains 36 pairs (N=36). For each pair, the ratio (c_{mod}/c_{obs}) has been determined.

The box plots display the distribution of this ratio. The boxes indicate percentiles 5, 25, 50, 75 and 95. Taking the left plot (A) as an example, we can consider the 36 pairs in the first group. The 95 percentile is close to unity, implying that for almost 95% of the pairs model A underpredicts. The 75 percentile is approximately 0.2, implying that for 75% of the pairs model A underpredicts by more than a factor of five.

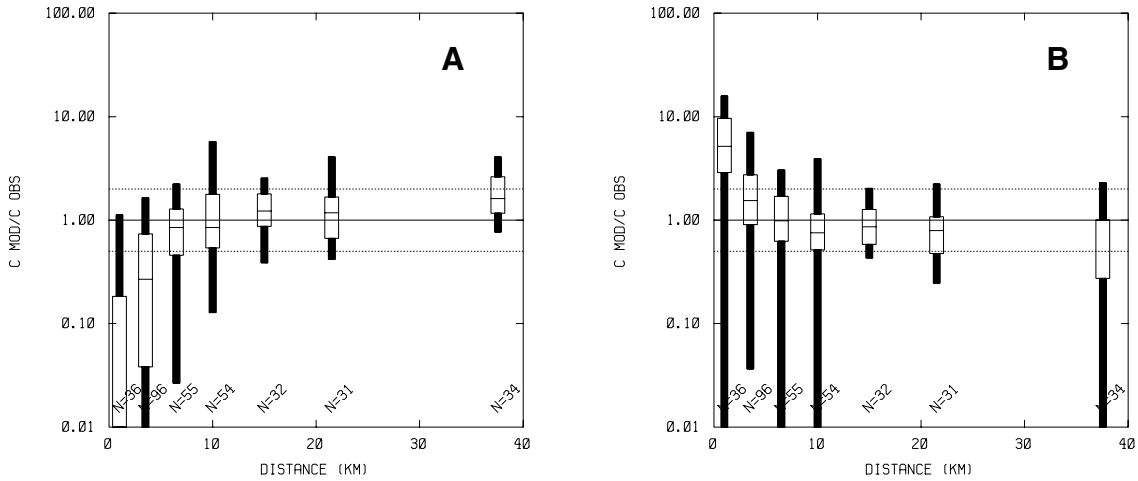

Figure 1. Model results for Kincaid data according to two models. For observations of "quality 3", the ratio of c_{mod}/c_{obs} is analyzed in terms of distance. The boxes indicate percentiles 5, 25, 50, 75 and 95. See text for further explanation.

For a physically correct model, there should be no more tendency for overprediction close to the source than far from it. Figure 1 shows that both model A and model B have problems in this respect, but behave quite differently.

Concerning data selection and processing, the graphs are based on a subset of data with well-defined arc-wise maxima (the so-called "quality 3" data). In order to reduce "noise", a filter has been imposed on the data presented in the figure: when *both* the observed and the predicted values are small (normalized concentration less than 15), the ratio is assumed to be unity.

In conclusion, one useful reaction to the problem of limited data sets is to *develop an understanding of model behaviour,* e.g. by using stratification of data as in the example.

4.3 Processing of input data for validation is far from trivial

In previous papers based on the results from the "harmonisation workshops" (Olesen, 1994 and Olesen, 1995b), it is demonstrated by many examples that processing of data is far from trivial. One important lesson to be learned is that it is extremely important to have well-defined criteria for selection of data.

The Kincaid data set lends itself as an excellent example to illustrate the problems related to quality control and rejection of data.

For the Kincaid experiment, the concentration pattern is often irregular - high and low concentrations occur intermittently along an arc. Hence, it is in many cases difficult to determine a representative maximum concentration along a cross-wind arc of monitors. Further, there are frequently gaps in the monitoring arcs. The question emerges: for cases where it is questionable whether the observed maximum is a reliable estimate of the actual maximum, should the data then be discarded from the set of arc-wise maxima?

It was found that a satisfactory solution has been to assign a *quality indicator* to each monitoring arc, indicating *how reliable* the arc-wise maximum should be considered. This quality indicator has been assigned on the basis of manual inspection of the geographical patterns of concentration distribution.

This procedure has provided a well-defined, common basis for model evaluation. If such a basis does not exist, some modellers will reject one part of the data, while others reject a different part, and any comparison will be futile.

The most important conclusion of the discussion on processing of input data is that careful work should be done. Data sets should be *prepared for ease-of-use, with their peculiarities and pitfalls well documented.* When comparing models, the models should all be run on the same data according to a common protocol.

4.4 The luxury of independent data sets can rarely be afforded

In model evaluation protocols, it is often stated that models should be independent of the data sets used for validation. At first sight this requirement may sound like common sense. However, in the author's opinion this is *not* a realistic claim in practice. A model may start its existence unbiased by existing data, but, as time passes by, no model can be expected to preserve its virginity in respect to the relatively few experimental data sets around.

A case story can serve to illustrate this, and at the same time demonstrate the capabilities of the *Model Validation Kit*.

Fig. 2 shows again results based on Kincaid data. The graphs require some explanation.

As before, arc-wise maxima are considered, and "quality 3 data" selected. The five models A-E were used to compute centerline maxima at ground level. For each model, this yields a set of modelled concentrations, which can be compared to the set of observed concentrations. The two sets have been sorted by the magnitude of concentration, so that observed and modelled values can be paired according to the rank. Finally, quantile-quantile plots as shown have been produced. Taking as an example the upper lefthand plot, it shows that for model A, the highest observed value is 319, while the highest computed value (which did not occur at the same time and arc) is 132.

Roughly speaking, a "perfect model" should be expected to produce points close to the one-to-one line*

The plots A-E were presented at the Mol workshop. There are striking differences between the models. Models A and D are not able to predict as high values as those actually observed; inversely, model B strongly overpredicts. Figure 2 is an example of a very fruitful approach to model evaluation: to run models on exactly the same basis, and then line up the results.

Actually, this graph triggered an immediate reaction when it was presented in Mol. In informal discussions it was attempted to identify the causes for the differences in model behaviour. Some potential causes were identified, and model B was revised shortly after the workshop. Therefore, now, model B should be judged not on the basis of the plot labelled B, but on the one labelled "Revised B".

This sequence of events is a perfect illustration of a mechanism, which complicates model evaluation, but must be regarded as natural and unavoidable: modellers will take advantage of the information in validation data sets, and use validation results for deducing necessary model improvements.

As a consequence, results from model validation exercises invariably become outdated. Furthermore, models will *not remain independent* of the data sets around. A suitable reaction to this problem is to validate models against many data sets.

5. WHAT CAN BE DONE BETTER?

Throughout this paper it has been emphasized that a common frame of reference is essential for model evaluation.

In the author's opinion, it would greatly enhance the productivity of the modelling community if tools for model evaluation were made generally available, so that the modelling community in practice would be able to use a common frame of reference. Such tools include *carefully prepared data sets, model evaluation software, and model evaluation protocols.* The *Model Validation Kit* is one initial step along this road.

Concerning data sets, they should be well organised, carefully checked, and with their pitfalls and peculiarities well documented. We still suffer from a lack of such data sets.

Concerning model evaluation software, we have some tools in the package of the *Model Validation Kit*. This kit is useful, but there is a need for supplementary software. An array of various model evaluation methods and corresponding software must be developed and be freely available.

Concerning model evaluation protocols, protocols for specific applications should be developed and their usability thoroughly tested.

A very real impediment to progress is that, until now, those few tools which actually did exist have been unknown to the majority of the modelling community. Now, however, recent developments in

---

* There are some reservations as to what one should expect: the prediction of a model is assumed to be an estimate of the *ensemble average* over many realisations in the centre of the plume. Even for a perfect model this value need not be identical to the one actually observed at the monitors, which is taken from a *single realisation.* Another complication is due to the fact that the observation is not necessarily performed in the plume centre. See Olesen, 1995b for a further discussion.

electronic communication have placed us in a much better position than a few years ago to share common tools - such as software and guidelines. The so-called World Wide Web (WWW) is an advanced system for accessing documents over the Internet. The most distinct feature of the WWW is its use of *hypertext* and *hyperlinks*. Some words in a document are marked, and each marked word has a link to another document or resource - which may physically be located thousands of kilometres away at another computer. Yet, the user can smoothly move between these documents without concern about their physical location.

An example of the use of the WWW in order to improve model evaluation practice has been mentioned in Subsection 3.5.

The advent of the WWW makes it possible to provide easy access to a common pool of information concerning model evaluation, such as guidelines, software and evaluation databases. A great advantage of the WWW is that it is not necessary for some central body to collect all information and continuously keep it up to date. This task would be so overwhelming for a single institution, that in practice it could never really be accomplished. On the other hand, with the use of the WWW, pooling of information can take place in a decentralized manner, so that several institutions can make contributions, each within their specific field of expertise.

REFERENCES

Bowne, N.E., Londergan, R.J., Murray, D.R. and H.S. Borenstein, 1983, *Overview, Results, and Conclusions for the EPRI Plume Model Validation and Development Project: Plains Site.* EPRI EA-3074, Palo Alto, CA.

Cuvelier, C. (editor), 1994, Proceedings of the workshop "*Intercomparison of Advanced Practical Short-Range Atmospheric Dispersion Models"*. August 30- September 3, 1993, (Manno - Switzerland), CSCS (Centro Svizzero di Calcolo Scientifico). Joint Research Centre, European Commission, EUR 15603 EN. Available from C. Cuvelier, JRC Ispra, TP 690, 21020 Ispra, Italy.

Duijm, N.J., Stork, B., Nielsen, M., Ott, S., 1995, An overview of the REDIPHEM project. Workshop on Operational Short-range Atmospheric Atmospheric Dispersion Models for Environmental Impact Assessment in Europe, Mol, Belgium, Nov. 1994, *Int. J. Environment and Pollution,* Vol. 5, Nos. 4-7, pp.-

Gryning, S-E. and Millan, M.M. (Eds.), 1994, *Air Pollution Modeling and Its Application X*. Plenum Press, New York.

Hanna, S.R., Strimaitis, D.G. and J.C. Chang, 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.

Hanna, S.R., Chang, J.C. and Strimaitis, D.G., 1993, Hazardous gas model evaluation with field observations, *Atm. Env.* 27A:2265.

Klug, W., Graziani, G., Grippa, G., Pierce, D., Tassone, C., 1992, *Evaluation of Long Range Atmospheric Transport Models Using Environmental Radioactivity Data From the Chernobyl Accident. The ATMES Report.* Elsevier. Pub. EUR 14147 of the Commission of the European Communities.

Kurzeja, R.J. and A.H. Weber, 1985, *Proceedings of the DOE/AMS Air Pollution Model Evaluation Workshop. Vol. 3. Summary, Conclusions, and Recommendations.* DP-1701-3. E.I. DuPont de Nemours & Co., Savannah River Laboratory, Aiken, SC 29808.

Lee, R.F., Overview of the U.S. Environmental Protection Agency's model evaluation activities, in: *Cuvelier, 1994*.

Model Evaluation Group, 1994, *Model Evaluation Protocol.* Can be requested from DG XII/D1, Rue de la Loi 200, B-1049 Brussels, Belgium. Fax +32 2 296 3024.

Olesen, H.R. and Mikkelsen, T. (Eds.), 1992, *Proceedings of the workshop on Objectives for Next Generation of Practical Short-Range Atmospheric Dispersion Models*, Risø, Denmark. National Environmental Research Institute, P.O. Box 358, DK-4000 Roskilde, Denmark. The volume is now out of print, but it is possible on request to obtain copies of specific papers.

Olesen, H.R., 1994, European coordinating activities concerning local-scale regulatory models, in: *Air Pollution Modeling and Its Application X.* Edited by S-E. Gryning and M.M. Millan, Plenum Press, New York.

Olesen, H.R., 1995a, Data sets and protocol for model validation. Workshop on Operational Short-range Atmospheric Atmospheric Dispersion Models for Environmental Impact Assessment in Europe, Mol, Belgium, Nov. 1994, *Int. J. Environment and Pollution,* Vol. 5, Nos. 4-7, pp.-----.

Olesen, H.R., 1995b, The model validation exercise at Mol. Overview of results. Workshop on Operational Short-range Atmospheric Atmospheric Dispersion Models for Environmental Impact Assessment in Europe, Mol, Belgium, Nov. 1994, *Int. J. Environment and Pollution,* Vol. 5, Nos. 4-7, pp.-----.

Royal Meteorological Society, 1995, *Atmospheric Dispersion Modelling: Guidelines On the Justification of Choice and Use of Models, and the Commmunication and Reporting of Results.* 104 Oxford Road, Reading RG1 7LJ, England.