A Platform for Model Evaluation

H.R. Olesen
National Environmental Research Institute (NERI)
P.O. Box 358, DK-4000 Roskilde, Denmark

Keywords: model evaluation, Model Validation Kit, atmospheric dispersion models, model evaluation, ASTM, inherent uncertainty.

1 Introduction

During the past decade we have had a series of Harmonisation workshops and conferences where a central topic has been model evaluation. The present paper is an attempt to synthesise some of the views which have been expressed and the work that has been done. I will here recapitulate some important issues: *Where is there a consensus concerning the approaches to take? Which problems are we confronted with? Which solutions are in sight?*

In the oral presentation I will extensively quote many papers, but space does not allow the same number of quotes in the present extended abstract (see URL 1 for additional references).

The problem of model evaluation is intrinsically difficult. P. Chatwin (1992) made a thought-provoking and very concise statement in an abstract for the first Harmonisation workshop:

In view of the fact that atmospheric dispersion is a stochastic phenomenon in all cases, it is undoubtedly scientifically wrong to describe it by mathematical models that are deterministic. Scientific correctness requires the use of statistical models.

In contrast to Chatwin's statement stands the fact that we apply a large number of deterministic models. We do it with relative success, and we are frequently required to assess model reliability. But there is much truth in Chatwin's statement, and as a result we are faced with serious problems when dealing with deterministic models.

Regulators would prefer models they can blindly trust. If they cannot trust model results completely – and they cannot – they need instead information concerning the uncertainty of model results. This represents a severe challenge.

2 Components of model evaluation

There seems to be reasonable consensus in the modelling community concerning the components of model evaluation. As a minimum an adequate model evaluation should comprise two components (according to a recent ASTM guide, 2000):

1. Science peer reviews
2. Statistical evaluations with field data

There are other tasks supportive to model evaluation, such as verification of correct coding and sensitivity analysis regarding model response to input data. These other tasks may be considered part of model evaluation – such is the view expressed in a "Model Evaluation Protocol" published under European Commission's Major Industrial Hazards Programme (Model Evaluation Group, 1994).

It is important to note that a model evaluation is not complete without diagnostic studies, attempting to ensure that the model gives the right result for the right reason. The ASTM guide sees this as part of the scientific review:

A key part of the scientific peer review will include the review of residual plots where modelled and observed evaluation objectives are compared over a range of model inputs, for example, maximum concentrations as a function of estimated plume rise or as a function of distance downwind.

3 Why is model evaluation difficult?

During the past decade, I have used various versions of Table I in order to list in a structured fashion the many problems involved in model evaluation. Discussions of the table can be found in Olesen (1994 and 1996). In the present extended abstract, the table will not be fully explained – only a few comments will be made.

3.1   The appropriate evaluation method depends on the context...

It is my impression that in the modelling community there is largely consensus on how to deal with the first difficulty mentioned in Table I: that *the appropriate evaluation method depends on the context of the application and the data sets available.* I think that there is general support for the view of Schatzmann and Leitl (1999), who write:

Which particular tests and which particular model/dataset comparison (one should make) *for a given model type can ultimately be based only on a consensus. Such a consensus needs to be built up for individual groups of models (e.g. obstacle resolving prognostic models) within and by the scientific and operational community which develops and uses those models.*

Similarly, the previously mentioned ASTM guide defines a framework, which can be filled in with specific procedures as experiences are gained for each class of models.

3.2   Processing of input data is far from trivial

Behind the innocently looking statement *"Processing of input data is far from trivial"* a host of problems are concealed. Many papers at this series of harmonisation conferences discuss different ways to process input data. For instance, there have been detailed analyses of how various concentration parameters – arcwise maxima, near-centerline concentrations and cross-wind integrated concentrations –can best be determined. Many authors have described pitfalls in datasets and processing methods, so it is obviously necessary to be careful in processing of data.

One experience from the past work – an experience that has been repeatedly confirmed – is the usefulness of assigning a quality indicator to experimental data, indicating how reliable a particular set of observations is. Such a quality indicator can be assigned by subjective methods (inspection of graphs), or it can be assigned by a computer code according to certain criteria. The use of a quality indicator is a great advantage, because subsets of data can be selected in a structured manner. Therefore, it is possible in a structured manner to discard data that would have been misleading if they were blindly included in an analysis.

Table I: Why is model evaluation difficult?

Difficulty	Reaction	Implied problems
The appropriate evaluation method depends on the context of the application and the data sets available.	An array of various evaluation methods must be developed.	What weight should be ascribed to various performance measures?
Input data sets are limited – they reflect only few of the possible scenarios.	a) Extrapolate model behaviour outside of validation domain.	a) Does the model give the right result for the right reason? We must understand model behaviour!
	b) Use many data sets.	b) Hard work! Ambiguous results.
Processing of input data for validation is far from trivial.	Take care! Identify pitfalls. Use quality indicators.	Numerous problems!
The luxury of independent data sets can rarely be afforded.	Use many data sets.	Hard work! Ambiguous results.
There are inherent uncertainties.	Use Venkatram's conceptual framework with emphasis on ensembles, not individual realisations.	Ensembles are difficult to establish.

3.3 There are inherent uncertainties

The last entry in Table I concerns the question of model uncertainty. A good starting point for discussing deviations between model estimates and observations is a conceptual framework, which has been described in several papers by A. Venkatram (e.g. 1982; 1999). The framework is used also in the recent ASTM guide (ASTM, 2000); here we adhere to the ASTM notation. The key idea is to decompose, respectively, *observed* concentrations and *modelled* concentrations.

It is a basic assumption that we have a model formulated in terms of a set of input parameters called α (this could be mixing height, Monin-Obukhov length and a number of other parameters). There is also a set of parameters for which the model does *not* account explicitly (such as the number of large convective cells passing during the sampling period), and this set is called β.

Let us consider the situation when the parameter set α has a certain set of values. Let us assume that our model is deterministic and thus purports to predict one number for each α: the *ensemble average* $\overline{C}_o(\alpha)$ for a large number of realisations. If we consider just one realisation, then the observed concentration (C_o) can be decomposed as follows

$$C_o(\alpha, \beta) \quad = \quad \overline{C}_o(\alpha) \quad + \quad c(\Delta c) \quad + \quad c(\alpha, \beta)$$

| Observed concentration | Ensemble average | Measurement error | Inherent uncertainty |

$$(1)$$

Note that the decomposition depends on the definition of α and thus is model-dependent. This is an annoying fact, which we must accept.

The *"measurement error"* term accounts not only for trivial instrument inaccuracy, but also for the fact that we may not measure the parameter that we assume. As an example, if we attempt to measure a cross-wind integrated concentration for a plume, we may be mistaken: the observed concentration may be different from what a more extensive set of measurements would tell us. The measurement error can in principle be reduced by increasing the number and the accuracy of measurements. As noted earlier, a way to avoid unnecessarily severe effects of measurement error is through the use of quality indicators assigned to data, so that misleading observations can be discarded.

Equation (1) further accounts for the fact that because of unresolved processes there is an *"inherent uncertainty"* $c(\alpha, \beta)$ which cannot be eliminated; these unresolved processes are represented by the β parameter set. Note that the inherent uncertainty is model dependent: if we include more variables in our set of input parameters α, the inherent uncertainty will presumably decrease.

Next, let us consider a *modelled* concentration for the α in question. It attempts to represent the ensemble average of observations, $\overline{C}_o(\alpha)$, but does so only approximately:

$$C_m(\alpha) \quad = \quad \overline{C}_o(\alpha) \quad + \quad d(\Delta\alpha) \quad + \quad f(\alpha)$$

| Modelled concentration | Ensemble average | Input uncertainty | Model formulation error |

$$(2)$$

This equation accounts for the possibility that the input to the model is incorrect (e.g., mixing height is another than assumed) and further recognises that the model may be incorrectly formulated.

A main inference from the fact that (1) and (2) are composed of totally different components is that observations and model results come from different statistical populations, *so their distributions cannot be expected to be identical.*

When comparing observations and model predictions we have a residual. As a consequence of (1) and (2) the residual can be seen to consist of the following terms:

$$C_o(\alpha, \beta) - C_m(\alpha) = \quad c(\Delta c) \quad + \quad c(\alpha, \beta) \quad - \quad d(\Delta \alpha) \quad - \quad f(\alpha)$$

| Residual | | Measurement error | | Inherent uncertainty | | Input uncertainty | | Model formulation error | (3) |

Eq. (3) is *the basic equation for interpretation of model evaluation results*.

The framework here can be used to see more clearly the difficulties we are confronted with when we want to compare our models with actual observations. It is clear from (3) that the residual may be severely affected not only by the skill of the model, but also by a mixture of other effects. This is the challenge that has frequently led to confusion and has provoked a never-ending sequence of scientific papers during the past decades.

If we consider a large number of measurements, and we have a perfect model, and we do not have systematic measurement errors, then on average the residual should be zero. But typical residuals for individual realisations are large. Of course they depend on precisely which physical reality we let our equations describe. But as an example, A. Venkatram (1999) stated (with reference to evaluation studies by Hanna):

The geometric means of the ratios of the model estimated to the observed concentrations, derived from several field studies, for two of the most recently developed models are close to 2.5. This means that there is over 30% chance that an observation will be 2.5 times greater or less than any model prediction.

It is clear that such large deviations between model estimates and observations are unavoidable, and thus needs to be accounted for in both judging model performance and in applying them. This requires a framework that defines the relationship between model prediction and observation.

I believe that it will be useful if we in the model evaluation community make more extensive use of the conceptual framework underlying Eqs. (1)-(3). It will make it easier for us to separate and analyse problems.

4 Which solutions are in sight?

In the modelling community there are groups of modellers working with various classes of models, each group trying to establish databases and evaluation methodologies for their respective class of model. Such groups are for instance the activity concerning street canyon modelling, and the activity related to the classic problem, where an inert gas is emitted from an isolated stack – typified by the Kincaid experiment.

Such activities are useful, as they may lead to the consensus referred to by Schatzmann and Leitl.

When looking for solutions – "toolboxes" – for model evaluation, I will note the two, which I am best acquainted with. One is the so-called Model Validation Kit. This kit is a practical tool meant to serve as a common frame of reference, but it does have well recognised limitations. It considers the classic single-stack problem. The kit is a collection of four data sets as well as tools for statistical and graphical analyses (Olesen, 1995 and 1997; URL 2). An important limitation of the procedure used in the kit is that it does not explicitly address the stochastic nature of atmospheric dispersion. This means that especially the quantile-quantile (Q-Q) plots typically produced by the kit should be used with great care. As noted earlier, even a "perfect" model cannot be expected to provide the same frequency distribution of concentrations as the one observed, and thus Q-Q plots should not result in a one-to-one correspondence. Residual plots (residuals versus physical parameters) can be a very useful supplement to Q-Q plots, because they provide an insight into model behaviour. The various issues related to the Model Validation Kit have been discussed in detail at previous conferences (e.g. Olesen 1999 and 2000, Cooper 1999, McHugh et al. 1999).

An alternative "toolbox" which covers the classic single-stack problem, but also has the potential to be extended to other dispersion problems, is the recently published "ASTM Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance" (ASTM 2000). This guide is primarily the result of John Irwin's work: in an attempt to address the problem with the stochastic

nature of observations in a consistent manner, he has produced a series of tools and papers concerning a methodology, which has evolved over the previous harmonisation conferences (e.g. Irwin 1999; 2000). Eventually, it has resulted in the ASTM Guide.

In my opinion this Guide represents a foundation on which we can base much of our future work. It should be emphasised that the Guide is of an open nature and allows developments to occur. Thus, the introduction states:

This guide provides techniques that are useful for the comparison of modeled air concentrations with observed field data... Methodologies for such comparisons are yet evolving; hence, modifications will occur in the statistical tests and procedures and data analysis as work progresses in this area.

The ASTM standard guide contains detailed discussions on the framework and procedures for model evaluation. The framework is general in the sense that it does not assume that we deal with a certain type of model or with a certain concentration variable. However, there is an appendix to the guide, which specifies an example where the framework is used. This example deals with the classic problem of a plume being emitted from an isolated point source.

The software and data used for this example are available, but presently not in a very well-organised and user-friendly form. There are plans to improve upon this. As this work progresses, there will be links to the currently available tools through the web site of the Harmonisation initiative (URL 3).

An important aspect of the ASTM guide is the emphasis on bringing implicit assumptions into the light of day. Some of its concluding words are:

There are evaluations in literature that implicitly assume that what is observed is what is being modeled, when this is not the case. This guide is attempting to heighten awareness and thereby promote statistical model evaluations...

We have a framework, but with lots of work ahead. As stated earlier, processing of input data is far from trivial and if this aspect is neglected, we can draw erroneous conclusions concerning model performance. I would like to conclude by repeating what I said after having worked extensively with preparation of data for the Model Validation Kit (Olesen, 1997):

In this paper, besides some facts on the pilot study data sets, many experiences on preparation and use of performance evaluation data sets have been conveyed. The moral of the story is that as a producer of data you have to work your way through the data and test things out; you should <u>not</u> just take a data set from the shelf and distribute it, assuming that your job is over. When working through the data, you will encounter numerous problems on your way, both tiny problems and larger. All of these problems should be eliminated one by one, laying the road open for future users of data...

Experience has shown that the process of creating useful data sets takes time; it takes time to prepare the data, it takes time for modellers to use them, and it takes time to revise the data set in response to the feedback received. All parties involved in evaluation activities must be aware of this nature of things. We should build on the experiences of others, and this is a long, continuing process.

5 Acknowledgements

The author wishes to acknowledge numerous contributors at the harmonisation conferences, but especially John Irwin, whose work has formed basis for much of the present discussion.

6 References

Due to space limitations in the present extended abstract, only very few of the references cited are indicated below. The full reference list is available through URL 1.

ASTM 2000, Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance, D6589. American Society for Testing and Materials (available through www.astm.org).

URL 1: www.dmu.dk/AtmosphericEnvironment/harmoni/Belgirate.htm. 'Papers for Belgirate'. Visited March 2001.

URL 2: http://www.dmu.dk/AtmosphericEnvironment/ m_v_kit.htm. 'Model Validation Kit'. Visited March 2001.

URL 3, http://www.dmu.dk/AtmosphericEnvironment/harmoni.htm. 'Initiative on Harmonisation...". Visited March 2001.