CONSOLIDATING TOOLS FOR MODEL EVALUATION

Helge R. Olesen[1] and Joseph C. Chang[2]
[1]National Environmental Research Institute (NERI), Roskilde, Denmark
[2]School of Computational Sciences, George Mason University, Fairfax, Virginia, USA

INTRODUCTION

During the series of Harmonisation conferences, many papers have used the so-called Model Validation Kit, which was introduced in 1993. A related subject of interest has been the evaluation methodology of the American Society for Testing and Materials (ASTM) standard guide D6589 on statistical evaluation of dispersion models.

The purpose of the present paper is to serve as a guide to the tools - data sets and software packages - that are currently available for general use. There have been recent updates to some of the tools, and only few people are aware of the status of the available material. The present paper serves as a key to the Model Validation Kit, while it also outlines some main features of the ASTM-related material that is now available on the Web. There is a need to consolidate existing tools, and the present paper represents one step towards that goal. However, consolidation is far from complete, as there are still many issues to be resolved and many tools deserving further improvements.

SOME BASIC RECOMMENDATIONS

It is recommended that any model evaluation exercise start with clear definitions of the evaluation goal and the variables to be considered, followed by exploratory data analysis, and then statistical performance evaluation. The implications of this are discussed in numerous papers, e.g. by Chang and Hanna (2004).

Therefore, statistical model performance evaluation should not be a stand-alone exercise. It is highly recommended to be coupled with exploratory data analysis, which can reveal model errors, and errors and inconsistencies in data.

THE MODEL VALIDATION KIT

The Model Validation Kit is intended to be used for evaluation of atmospheric dispersion models. It is a collection of four field data sets as well as software for model evaluation. The Kit is a practical tool intended to serve as a common frame of reference for model performance evaluation. It is, however, limited in scope, as described in subsequent discussions.

The Kit has been used for the series of Harmonisation workshops and conferences. A preliminary version of the Kit was used for the workshop in 1993, while a subsequent version was used essentially unchanged throughout the period 1994 - 2005 (in 1997, a supplement was added). It has been distributed to more than 250 research groups during that period.

The package was recently updated to Version 2.0 (in September 2005). The new version allows the same studies to be carried out as the previous version, but has been revised in several respects. New software and computing environments have made it necessary to update the package. Furthermore, the documentation is significantly improved and brought up to date. The package can be downloaded from the Internet at www.harmo.org/kit.

The package contains the following elements:

Field data sets from Kincaid, Indianapolis, Copenhagen and Lillestrom;
The BOOT statistical model evaluation software package;
Tools for exploratory data analysis, useful for diagnostic model evaluation;
A recommended procedure (protocol) for model performance evaluation. This procedure is relatively simple and thus has some limitations.
Note that although the emphasis of the Model Validation Kit is on the protocol, some tools included in the Kit - in particular the BOOT software - are general and can be applied for problems beyond the scope of the protocol.

1.14. The BOOT software

The main tool for statistical performance evaluation is the BOOT software package. The BOOT program has been improved and is now available in version 2.0 with a comprehensive, rewritten User's Guide (Chang and Hanna, 2005). Besides detailed technical description of performance measures and the use of the software, the User's Guide also provides a discussion of model evaluation objectives and exploratory data analysis. The BOOT package is flexible and general in nature. Although it has been primarily used to evaluate the performance of air dispersion models, the same procedures and approaches implemented in BOOT also apply to other types of models.

Compared to the previous version of BOOT, the package now includes some additional performance measures, and an implementation of the ASTM statistical model evaluation procedure (see later). The BOOT package is capable of computing performance measures such as the Fractional Bias (FB), the Normalised Mean Square Error (NMSE), the Geometric Mean Bias (MG), the Geometric Variance (VG), the fraction within a factor of 2 (FAC2), the Measure of Effectiveness (MOE), as well as several others. (FB and MOE are in fact closely related.) With the new software version, FB and MG can be separated into overpredicting and underpredicting components. Bootstrap resampling is used to estimate the confidence limits of a performance measure - hence the name BOOT of the package.

1.15. Tools for exploratory data analysis

When performing model evaluation, it is not sufficient to consider just statistical evaluation that produces some performance metrics. Rather, it is recommended that exploratory data analysis also be performed using graphical techniques.

The Model Validation Kit includes some tools for such graphical analyses in the form of the SIGPLOT graphical package. The SIGPLOT package is offered as an option that is specifically tailored for model performance evaluation. It must be mentioned that the SIGPLOT program, as well as a number of associated utility programs included in the Model Validation Kit only function in a DOS environment. The package can produce residual plots, where model residuals are depicted as a function of independent variables such as the downwind distance and time of day.

It is recognised that the somewhat archaic SIGPLOT package is only one of the many ways of performing exploratory data analysis. Its purpose is primarily to demonstrate the types of analyses that should be done. More modern and interactive tools than the SIGPLOT package can certainly be used to achieve the same goals. For example, a potential alternative is to use Microsoft Excel for data handling and graphical analyses. Excel offers some very powerful tools for interactive data analysis. Some useful hints and utilities related to Excel can be found in the Kit. Nevertheless, Excel does not offer the specialised plots that SIGPLOT produces. The advantages of using SIGPLOT are that you will be able to produce residual and other types of specialised plots with data in a relatively standardised format, that the required utilities are already prepared, and that the procedures for using the software are described in detail in a Compendium, which is part of the Model Validation Kit.

1.16. Data sets

The Model Validation Kit addresses the classic problem of a single stack emitting a non-reactive gas. The Kit comprises data from the following four field experiments:

The Kincaid experiment (1980-81) with tracer releases from a 187-m stack. There are 171 hours of tracer data from monitoring arcs at distances from 0.5 to 50 km. In the Model Validation Kit, the emphasis is on arc-wise maximum concentrations.

The Indianapolis experiment (1985) with tracer releases from an 84-m power plant stack in the city of Indianapolis, USA. There are 170 hours of tracer data from monitoring arcs at distances from 0.25 to 12 km. The emphasis is on arc-wise maxima.

Data from an experiment in Copenhagen, Denmark in 1978-79 with releases from a non-buoyant elevated source (115 m) in neutral and unstable conditions. Nine hours of tracer data are available on arcs from 2 to 6 km. Both arc-wise maxima and crosswind-integrated concentrations are considered reliable.

Data from an experiment in Lillestrøm, Norway (1987) with tracer releases from a non-buoyant source at 36 m in stable (winter) conditions. Sampling took place during 8 separate 15-minute periods.

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 (e.g., inspection of graphs), or assigned by a computer code according to certain objective criteria. The use of a quality indicator is valuable, because subsets of data can be selected in a well-defined manner. This can be utilised to discard data that would have been misleading if they were blindly included in an analysis. For two of the experiments, Kincaid and Indianapolis, the tracer data have been flagged by a manually assigned quality indicator assessing the quality of arc-wise maximum concentrations.

1.17. Limitations

It must be recognised that model evaluation studies performed on the basis of the Model Validation Kit are limited in scope. These limitations can be summarised as follows:

Only four experimental data sets are considered.

The emphasis is on operational short-range models.

The problem of interest is relatively simple, namely a point source emitting a non-reactive gas over flat terrain, due to the fact that this is the scenario represented by the four field experiments. On the other hand, much of the software included in the Kit is general and applicable to many different release scenarios.

Further, the emphasis is primarily on a) arc-wise maximum concentrations, and to some extent b) cross-wind integrated concentrations.

The Kit does not explicitly account for the stochastic nature of dispersion problems.

The tools in the Kit can be used to diagnose strengths and weaknesses of the models, but as a consequence of the above limitations, you should be careful in interpreting the results.

To further elaborate the last bullet in the above list, atmospheric dispersion processes are stochastic, whereas models in general predict only ensemble averages - not individual realisations. This means that there is a basic conceptual problem with the procedure of directly comparing model predictions to observations, as they cannot be expected to have the same statistical distribution. One consequence is that if the monitoring network is sufficiently dense and if the data represent a sufficient number of scenarios, then a "perfect model" is likely to underpredict the highest observed concentrations.

Despite its limitations the Model Validation Kit has the advantage of being straightforward to apply and practically oriented. It also provides a common framework where the results of different studies can be intercompared.

AN ALTERNATIVE: THE ASTM METHODOLOGY

As noted, there is a concern that direct comparison of model predictions against observations could cause misleading results. Therefore, an alternative approach has been proposed by John Irwin, and has resulted in ASTM Standard Guide D6589. This procedure has also been incorporated in the latest version of the BOOT software as an option. However, there exists also a separate package (software and data sets), specifically devised as an implementation of the ASTM procedure - here referred to as the *ASTM package*. It was prepared by John Irwin and is available on the Internet (www.harmo.org/astm). It can be used as a supplement or an alternative to the Model Validation Kit. We will here outline the main principles of the ASTM methodology, and further explain some features that distinguish the two packages.

The ASTM guide is a guide on "statistical evaluation of dispersion model performance", and is general in the sense that it is not confined to the problem of dispersion from an isolated point source - which is the focus of the Model Validation Kit. However, as an example of application of the principles in the Guide, the particular problem of point source dispersion is addressed in an Appendix to the Guide, as well as in the ASTM package.

The fundamental premise of the ASTM procedure is that observations and model predictions should *not* be compared directly, and that observations should be properly averaged before comparison. The comparison takes place within *regimes*, which for example can be defined according to atmospheric stability and distance to the source. The ASTM procedure then calculates performance measures based on *regime averages* (i.e., averaging over all experiments within a regime), rather than the values for individual experiments.

In the specific implementation of the ASTM methodology found in the ASTM package, the observations of interest are near-centreline concentrations (NCCs). NCCs bear some relation to arc-wise maxima, but for a given experiment and arc, there may be several NCCs as opposed to only one arc-wise maximum. NCCs are selected among those observations that lie within a distance of 0.67 σ_y from the cloud centre, where σ_y is the cloud width. The software of the ASTM package is capable of retrieving NCCs from observations and process them in accordance with the prescribed methodology. Note, however, that selecting NCCs is not straightforward, as it depends on regime definitions, and there are questions as to which arcs can be considered having enough data for NCCs to be defined.

The ASTM package includes software, documentation and three datasets (Kincaid, Indianapolis and Prairie Grass). The data in the package have not been quality flagged, but the software performs certain checks when retrieving NCCs.
The ASTM procedure implemented in the BOOT software in the Model Validation Kit assumes that NCCs have been retrieved separately.

The ASTM procedure represents a framework and is not a fixed protocol. For example, regimes can be defined in many different ways, and this may lead to differing results of a performance evaluation, depending on regime definitions.

Altogether, the ASTM procedure represents a promising approach, but still with some issues that are not fully resolved. Some issues deserving attention are:

- There is a need to study the sensitivity of the evaluation results to the definition of regimes (i.e., how data are stratified).
- There is always only a limited number of regimes (e.g., ~20 to 40) that can be defined. The performance measures are always determined by this limited number of regime averages. It is necessary to carefully examine the implication of accounting for only the variance in regime averages, rather than the full variance in the complete dataset.
- In the current implementation of the procedure with near-centreline concentrations (NCCs), it is problematic that NCCs are compared compared to model predictions in the *exact* centerline, which by definition are higher than near-centerline values.
- The basic assumption that (averaged) model results should fit (averaged) observations may be unwarranted if the quality of observed data is not properly assured. This is especially a concern when experimental data are fed into a statistical "blackbox" where these data are processed and averaged before a result is inspected. Problems with the observed data or the way they are interpreted may easily pass unnoticed. Data quality should be assured. Use of a quality indicator could alleviate such problems.

CONCLUSION

In order to ensure model quality, it is desirable to have access to carefully prepared data sets, model evaluation software, and model evaluation protocols.

This paper gives an overview of some available tools. These tools include some of the desired features just mentioned. However, there are also some limitations, such as not enough *carefully prepared* data sets, and lack of robust evaluation protocols that can be applied without reservations and lead to definitive conclusions. A model evaluation exercise will usually result in somewhat "inconclusive conclusions" due to the limitations listed in the paper. In general, modellers would be hesitant to designate a "best performing model" without reservations, although it is relatively easier to designate a group of "good performing models."

Nevertheless, model evaluation based on available tools and data sets is extremely useful to promote the quality of models. The evaluation process provides an opportunity for modellers to recognise severe model weaknesses, and subsequently correct them.

ACKNOWLEDGEMENTS

Special thanks go to John Irwin, whose persistent work on the ASTM methodology has led it to its current status, and to Steven Hanna, who has contributed in numerous ways to the work with the Model Validation Kit.

REFERENCES

Chang, J.C. and Hanna, S.R., 2004: Air quality model performance evaluation. *Met. Atmos Phys.*, **87**, 167-196.

Chang, J.C and Hanna, S.R., 2005: Technical Descriptions and User's Guide for the BOOT Statistical Model Evaluation Software Package. Available through www.harmo.org/kit

There are numerous papers by Irwin and Olesen concerning the issues of the present paper. References can be found through two Web sites:

http://www.harmo.org/kit : Web site of the Model Validation Kit

http://www.harmo.org/astm : Web site providing access to the ASTM package.