

How do we demonstrate that a consequence model is fit-for-purpose?

Simon Coldrick, Health and Safety Executive, Harpur Hill, Buxton, Derbyshire, SK17 9JN, UK

simon.coldrick@hsl.gsi.gov.uk

In consequence assessment for major hazards, the ability to predict things, rather than measure them or wait to learn from an accident, is almost essential. However, since predictions rely upon models of reality, rather than reality itself, there is a need to examine the correspondence between the model predictions and the real world. Model evaluation is a means for assessing the credibility or fitness-for-purpose of a model and for providing assurance to a decision maker that model results are credible. Approaches for evaluating models have existed since the early years of computer modelling where the appearance of the phrase "garbage in – garbage out" signals an early mistrust of results of simulations. As computers have become more powerful, software has become more accessible, and modelling techniques have advanced. It is now possible to create predictions that appear at first sight to be highly realistic, but the need to question the accuracy of models remains as relevant as ever. The model evaluation process typically comprises a number of activities: scientific assessment, verification, validation, user-oriented assessment and sensitivity analysis. There is a vast array of material on how each of these stages may be carried out and recommendations on procedures for evaluating different types of consequence model. Despite this, there are many aspects of model evaluation which remain uncertain and the whole process is far from trivial. Recent experiences in developing and applying model evaluation procedures have shown that the lack of relevant experimental data for validating consequence models remains a problem. Another critical issue is the lack of consistency in the way models are compared to the data and deemed to be acceptable. This emphasises the need to ensure that the evaluation process is consistent, transparent, straightforward and comprehensive, and not simply a data-comparison exercise. The time and cost involved in model evaluation also lends weight to the idea that model evaluation should be considered as an ongoing process of model improvement, rather than as a one-off pass/fail test.

Keywords: Model evaluation, protocol, validation, verification, scientific assessment

Modelling and model evaluation – why do it?

Mathematical models of physical processes and their embodiment in computer simulations are widely used in many areas, because they allow us to learn something about a particular situation that would otherwise be too difficult, expensive or impractical to achieve by other means. Modelling is an integral part of the risk assessment process for this reason - it allows us to make an estimate of what might happen under a particular set of circumstances. This modelled estimate may be based on test data for a slightly different scenario, or at another scale or it may not be based on test data at all, just the knowledge of the physical processes involved. In any of these cases, it is important to ensure that the level or complexity of the modelling is proportionate for its intended use and that the modelling is "fit-for-purpose".

Regardless of the level of complexity of the modelling, the process of model development typically follows a generic set of steps shown in Figure 1, taken from Ivings et al. (2007, 2016). This generic set of steps could be applied to any computer model of any physical process. The only difference between the different modelling approaches may be that higher fidelity models might require more complex equations which may in turn require more sophisticated solution methods. In any case:

- The set of equations chosen must actually represent the processes being modelled (scientific assessment)
- The solution method must function properly (verification)
- The model predictions must match reality to an acceptable level (validation)

Once we are satisfied that the above criteria have been met, we have increased our level of confidence in the predictive capabilities of the model. These three steps support the idea that a model has predictive capability (box four in Figure 1), but only for a particular scenario.

One approach to model evaluation is shown in the three steps above. The process of model evaluation is not new and to a certain extent, is an implicit and necessary part of the model development process. References to model evaluation as an activity in its own right started to appear in the late 1960s, in connection with the increasing use of computer models. For example, Van Horn (1971) reviews and discusses issues of simulation testing that date back to the late 1960s, using what have now become familiar terms for model developers. One of the main reasons for evaluating models may be to inform a decision maker, who is far removed from the modelling process, of the quality of the model results. Model evaluation also gives decision makers an indication of the applicability of a model to new problem areas. These two aspects are particularly important in a regulatory environment where there may be a need to provide evidence of the quality of model predictions, in addition to the suitability of the model for a given scenario. Duijm and Carissimo (2001) also suggest that a further benefit of model evaluation is that it can encourage model improvement through management of model quality. Model evaluation may also identify areas for improvement in models as well as shortcomings in experimental datasets.

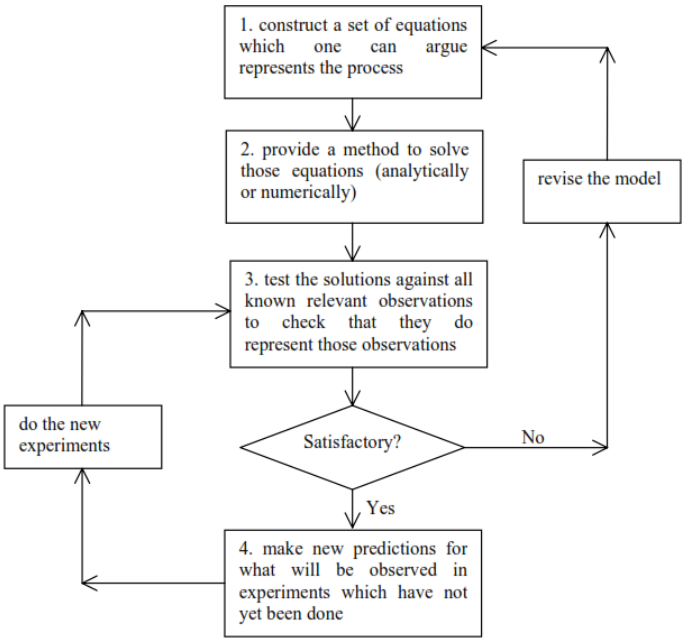

Figure 1 The model development process

In the late 1970s, the US General Accounting Office (GAO) became involved in model evaluation (Balci, 1986). This was because the use of complex models by many government departments was increasing and it was recognised there was a need for guidelines for use and interpretation of the models in decision making by senior management. The GAO subsequently organised a model evaluation review group involving developers and users in business, industry, government and academia. One of the outcomes was a report titled "Guidelines for Model Evaluation" (US GAO, 1979) which was summarised by an article in Operations Research (Gass and Thompson, 1980). In this report, the GAO set out a very general set of evaluation criteria, aiming to be independent of the subject matter or the modelling methodology. They suggest as a minimum, the evaluation process should include the steps of: documentation, validity, verification, maintainability and usability. In the following years, model evaluation became an important component of air quality modelling, driven by regulatory requirements (Fox, 1981). Evaluation frameworks and exercises also arose for consequence modelling in part from the need to assess the effects of potential spills of fuels and rocket propellants which are highly toxic and unstable. From these beginnings, further evaluation protocols were developed in the area of consequence modelling (e.g. Ermak and Merry, 1988; Hanna et al., 1988; MEG 1994; Ivings et al., 2007), including collections of data (e.g. Hanna et al., 1991; Nielsen and Ott, 1996; Carissimo et al., 2001; Stewart et al., 2016). Many of these protocols follow a similar format to that suggested by the US GAO in 1979 and include the basic steps of: scientific assessment, verification, validation, usability and sensitivity analysis. This structure arises naturally, because each stage is dependent on the previous one having been carried out. There is little point in validating a model which has been programmed incorrectly and there is little point in programming a model which is not scientifically robust. That does not mean that such faults do not happen in practice and the evaluation process should be designed to detect them.

In this paper, each of the steps of scientific assessment, verification, validation, usability and sensitivity analysis will be discussed in light of recent experience. It is possible to get into philosophical and/or semantic deliberations on the meanings of the various terms but this has already been covered in depth in various sources (e.g. Roache, 1998). Instead, we aim to revisit some of the topics reviewed by Olesen (1994), who provided some answers to the question "Why is model evaluation difficult?"

The challenge of model evaluation in consequence modelling

The techniques of model evaluation are well established where models are used in the product design process. This is because the design process is often an iterative cycle involving different types of modelling and prototype testing. That is not to say that the process is easier than in consequence modelling – rather the goal is different. The overall aim may be to use modelling to increase the efficiency of a piece of machinery by a few percent, or to assess the feasibility of new designs. Emphasis can be placed on model validation because models are very often being applied within a narrow region of operation and can be comprehensively validated for those scenarios.

In consequence modelling for hazard assessment in the process industries, important decisions are based on modelling results for complex, sometimes completely new scenarios, for which there may be little or no experimental data. The same reliance cannot be placed on model validation and other tests are needed to provide assurance that the predictions are

reliable. One of the reasons for this is that we are interested in a wide range of model types and operating ranges. It is therefore important for model users, developers and regulators to understand the key principles of what is required to demonstrate that a model is fit-for-purpose.

Scientific assessment

Scientific assessment may be one of the most important aspects of model evaluation. Although validation can be used to assess whether a model is "right," in order to have confidence in a model, it must be "right for the right reason." This is an aspect that is partly addressed by verification, because a model may give the right results but be incorrectly implemented in software. However, the underlying model must also have a sound physical basis. Venkatram (1988) noted that a good model should incorporate a realistic description of the physical processes being modelled as this provides the confidence necessary to apply the model beyond the range of observations used to test the model. This contrasts with correlations, when they are not based on appropriate scaling parameters, have not been subject to scientific assessment and therefore cannot be used in cases different from the experiments on which they have been based. Scientific assessment examines the model form and assumptions and whether they are consistent with physical principles. The US GAO (1979) refer to such activities as "Theoretical Validity" which, although not the now accepted meaning of validity, appears to convey the need to examine the theories and assumptions on which the model is constructed. This need was also identified in an air quality model evaluation workshop summarised by Fox (1981) where it was recorded that a scientific evaluation should also be included. In some cases, it was felt that scientific judgement might prove to be the only way to distinguish between models. Validation may be associated with determining numerical values which represent the goodness-of-fit of a model with the experimental data. While statistically-based validation can result in more objective and well-defined evaluations, Ermak (1988) suggests that this may be at the expense of understanding. Olesen (1994) defines this understanding as the "diagnostic (scientific) approach" which is complementary to the "operational (statistical) evaluation", i.e. that confidence in a model can only be gained through a combination of approaches.

For certain phenomena, Duijm and Carissimo (2001) suggest that there may be insufficient test data to provide proof of a model's quality and its capabilities for problems outside the range of the validation datasets. In these cases, most of the evidence may be provided by the scientific assessment. This scenario is discussed later.

Models, then, must incorporate an *appropriate* description of the physics. The problem for an evaluator is to determine if this is the case. Model evaluation documents and protocols often state that the science must be "broadly accepted" or conform to "state of the art". This implies a certain level of technical knowledge on the part of the evaluator and also that they have access to a full description of the model in the documentation. An alternative approach to the assessment is for the model evaluation protocol to contain a check-list of essential features for a given phenomenon and for the evaluator to complete the check list based on information obtained from the developer (possibly via a questionnaire). The evaluator then does not need to have the same level of expertise, but protocols and checklists would need to be developed for every conceivable phenomenon. It is difficult to see how this situation may be improved upon.

Verification

The main aspect that distinguishes verification from validation is that verification is not directly about the physical system being modelled – rather it involves checking that the computer implementation of a model is consistent with its mathematical basis. Verification encompasses a number of aspects which are all necessary to demonstrate that the computer implementation "runs as intended." The US GAO (1979) suggests that these include establishing that the computer program, as written, accurately describes the model as designed and also that software aspects are correctly implemented and debugged. More recently, in a review of verification and validation techniques, Oberkampf et al. (2002) divide verification into code verification and solution verification. They also suggest that code verification should be further sub-divided into numerical algorithm verification and Software Quality Assurance (SQA).

The applicability of the different types of verification depends on the type of model and whether the verification is being done by the model developer or model user. Gass (1977) suggests that verification is an activity of the model developer, rather than the evaluator. The definitions given by Oberkampf et al. (2002) are probably more relevant to Computational Fluid Dynamics (CFD) because code verification is an activity that might be expected of the developer, or software vendor. Vendors supply a set of pre-compiled algorithms and routines that need to function correctly. Solution verification is an activity that might be expected of the user. They need to be able to demonstrate that the model they have constructed is still faithful to the flow equations they are ultimately solving. The CFD code may be correct and error free, but the solution may not be mesh independent or properly converged. Accurate verification of CFD models is not a straightforward task, in part due to the "black box" nature of proprietary software and in part because of the complexity of modern codes.

Mercer et al. (1998) cover verification fairly briefly, noting that it is an extremely tedious[1] task and many developers take a less rigorous approach. Their view on verification is that the evaluator should appeal to the developer to provide information on what verification has been undertaken. It may also be possible to carry out some simple internal consistency checks such as mass and momentum balances and running the code against analytical solutions where possible. This is similar to the approach taken by the EU SUSANA[2] hydrogen model evaluation project (Baraldi et al., 2016) which provides an on-line database of verification problems and worked examples.

[1] This was also alluded to in 1839, by Charles Babbage (Buxton and Hyman, 1988), when referring to the in-depth checking of calculations.
[2] www.support-cfd.eu (accessed 22nd November, 2016)

Validation

The accepted definition of validation is straightforward enough – it is the comparison of model predictions against physical experiments. Ivings et al. (2007, 2016) propose that comparison with an experiment can never show that a model is "valid", the best it can do is fail to show that the model is "invalid". While validation may be easily defined, actually doing it in practice is somewhat more difficult when it comes to obtaining the data, doing the comparisons and deciding if the model results are "acceptable".

The lack of data suitable for model validation is an ongoing problem. Olesen (1994) pointed out that datasets are limited and that the luxury of independent datasets can rarely be afforded. This means that some models have been validated against experiments that were used to calibrate those same models. When one examines the possible number of scenarios needed for model validation (substance, scale, release mechanism, dispersion mechanism, weather, ignition mechanism, fire, explosion etc.), it becomes clear that Oleson's comments on lack of data are still valid, twenty years later. A further issue is that new and emerging risks require new data for validation of new (or reconfigured) models. As an example, in recent years significant research efforts have been spent on carbon dioxide experiments for validating models used in assessing Carbon Capture and Storage technologies, and efforts are continuing at present to examine issues relevant to Floating Liquefied Natural Gas (FLNG) operations and the hydrogen economy. Measurement and modelling techniques also advance over time. Datasets that were originally intended for validating one type of model may have to be applied to a new type of model with different requirements. In this case, an option is to revisit the old data and process it into the required form. In other cases, it may be desirable to repeat experiments with new measurement equipment. One good example of this is NOAA's project Sage Brush[3], which is revisiting the Prairie Grass passive dispersion experiments that were conducted in 1956 (Barad, 1958).

An interesting definition of validation is given by the US GAO (1979) as "Validation examines the correspondence of the model and its outputs to perceived reality." Here, the word "perceived" makes an important distinction, because the model predictions are compared to a measured and processed or simplified form of reality. For example, an experimenter may take a measurement of some quantity with an instrument, logging the data, averaging it, then printing the values in a graph in a report. Some years later, a modeller wishes to do a validation, but only having a hard copy of the experimental report, has to scan it and digitise the graph to obtain the data. Unless it was recorded at the time, all the information on the accuracy of the measurements and the steps used to process the data is lost. Nowadays, as computer storage is less of a problem, it is easier to achieve the goal suggested by Olesen (1996) of creating well organised and structured datasets, with their pitfalls and peculiarities well documented. This makes it possible to return to the original data at a later time and re-process it if necessary, correcting any errors that may have arisen. Creation, maintenance and upkeep of these datasets remains an issue, particularly when the original funding for projects expires and organisational structures are prone to change. Significant efforts have been spent recently, for example, in correcting errors and updating the database that accompanies the LNG Model Evaluation Protocol (Stewart et al., 2016).

Data processing

Olesen (1994) highlighted that processing the experimental data is far from trivial as the raw measurements must be converted into a format suitable for model evaluation and this must be done while preserving the essence of the data. For example, in dispersion modelling, inappropriate averaging may result in a cloud shape that never existed in reality and could not be modelled. The difficulties may be compounded by the need to turn a transient process into a steady state. User-variability in modelling is well documented and the same variability also arises when processing experimental data. Different approaches are possible when it comes to time averaging, treatment of zero or non-detect values etc. Ideally, the processing method should be consistent with the intended use of the data. If data are stored in an unprocessed form, such as time series, then it is possible to process the data accordingly. However, because this is a time consuming activity, there is a benefit to supplying pre-processed data in a database (such as Hanna et al., 1991; Stewart et al., 2016) which allows a common frame of reference to be used in a model evaluation study.

Comparing models and experiments

The methods used to compare model predictions to experimental measurements must be considered when deciding upon data processing methods, to ensure that the processed data is suitable. Qualitative evaluation of models can be undertaken by comparison of plots of the relevant variables and this can give a general indication of the ability of a model to predict a particular scenario. This exploratory data analysis is recommended by Chang and Hanna (2005) as a first step in model evaluation, possibly including scatter plots, quantile-quantile plots, residual (box) plots and conditional scatter plots. However, whilst such analysis "by eye" is essential and informative, it can also be subjective or introduce variability.

For a more rigorous evaluation, a procedural quantitative approach can be adopted. Statistical performance measures (SPMs) provide a means of comparing measured and predicted physical comparison parameters. They are non-dimensional and therefore the comparison made is independent of the units of any observed and predicted quantities. Chang and Hanna (2004) suggested that multiple performance measures should be applied as each measure has its advantages and disadvantages and there is not a single measure universally applicable to all conditions. A further consideration when selecting SPMs is that it is beneficial to be consistent with those previously used in other model evaluations. By doing so, it is possible to gain experience and understanding of SPM values allowing comparison of one model's performance to another and therefore get a feel for what "good" model performance looks like.

[3] http://www.noaa.inel.gov/projects/sagebrush/sagebrush.htm, accessed 2nd December,2016.

One of the fundamental problems of SPMs can occur when they are based on measurements and predictions at fixed locations in space. It might be the case that, due to the physics of the experiment or some other reason, nothing is measured at a particular location but something is predicted. Alternatively, something may be measured, but nothing predicted. SPMs based on ratios of predicted and measured values are meaningless in these cases and SPMs based on differences may falsely show under- or over-prediction. SPMs may be aggregated (averaged) values over multiple points. If either the measurement or prediction is zero at just one point, this can produce a single undefined or infinite value at that point, which then causes problems in the aggregated value for the whole SPM. The evaluator is then faced with a dilemma: should they discard the comparisons at these points (which would bias the results in one direction), or try and include them (which would also bias the results in another direction)? Another option is to explore the use of alternative metrics, as discussed by Yu et al. (2006). A similar problem occurs in experiments where very small values below the limit of detection are measured, where Helsel (2005) argues techniques such as removal and or substitution with fixed values are overly simplistic. The use of data quality indicators is recommended by Olesen (1996) as a measure of the reliability of data points, however, it may well be that acknowledging any shortcomings in the analysis is the only option. Gant et al. (2016b) describe an example where the configuration of sensors in a dispersion experiment meant that the plume passed between some of the sensors. Therefore, the way in which models are compared with these experiments can have a significant effect on the results of the validation exercise. In another example from Gant et al. (2016a), spray mists of oil were simulated with a CFD model, and due to the large number of simulations, an SPM approach was used to help identify the best performing model. In the experiments, measurements were only made where droplets were present and care was therefore needed not to draw incorrect conclusions from the analysis. As noted by Venkatram (1988) "it has now become fashionable to compute all sorts of statistical performance measures that are supposed to quantify performance. In my opinion, the reason for this is that it is easy to compute statistics and compare numbers (which are often meaningless). On the other hand, it is much more difficult to examine the science that governs models."

McGrattan et al. (2014) provide an argument for the use of less-than-perfect data, noting that thousands of experiments have been undertaken but some of these are missing vital information on how the tests were set up and the data processed. They suggest that it would be foolish to throw away all these datasets and the number of data points combined with appropriate statistical techniques can make up for a lack of quality in the data.

Other issues

Small-scale test data are often used in place of large-scale tests, particularly for dispersion experiments. The argument for this is that tests undertaken in the laboratory or wind tunnel are often better controlled, have reduced experimental uncertainty and are easier and cheaper to undertake. However, some physical effects such as non-neutral atmospheric stability, thermal effects and deposition and are not easily replicated in wind-tunnel tests. Whilst field-scale tests have advantages in this respect, there are difficulties in that the weather conditions at the test site may change over the course of the experiment and it may be impossible to perform multiple repeats of the experiment. Since atmospheric turbulence introduces a random element to dispersion behaviour, maximum concentration data from field-scale tests may represent just one sample of the range of possible outcomes, rather than a statistical average (this is particularly an issue for short-duration "puff" releases). It therefore becomes difficult to compare a model's prediction of the ensemble average to measurements of just one realization of the flow behaviour. For these reasons, there are merits in using a combination of both small-scale and field-scale datasets in a validation exercise. One of the issues with small-scale data is the uncertainty introduced by using scaling rules to infer their equivalence to field-scale data. In the recent edition of the LNG MEP (Stewart et al., 2016) modellers are required to simulate the wind-tunnel experiments at wind-tunnel scale, rather than equivalent field scale.

The above discussion assumes that some test data are available. What would be a suitable option if this were not the case? Comparison of a model with other models in the absence of experimental data is not validation – it can only be a model comparison. However, Roache (1998) suggests that an indirect validation can be carried out by comparing model results against those from another "benchmark" model which has previously been validated (i.e. validation by proxy). This is because the model results are still being compared to data, but in a second-hand way, one level removed from the original experiment. If a model evaluation has to be carried out in the absence of experimental data, a situation identified by Webber et al. (2009), the evaluator will still need to satisfy themselves of the predictive capabilities of a model. In these cases, more emphasis may need to be placed on scientific assessment, verification and sensitivity analysis. One of the benefits of comparing one model to another is that it provides a means of assessing the performance of a model against a baseline to help assess the effects of any model refinements. This approach was recently proposed in the context of passive dispersion modelling by Herring and Huq (2016).

User-oriented assessment

User-oriented assessment is an aspect that is linked with scientific assessment, but considers practical usage of a model to solve a given problem (CERC, 2000). The main aim of user-oriented assessment is to assess how information is input into a model and how the results are interpreted. Closely linked to this is the assessment of guidance on the operation of the model in terms of documentation and user support. Model results may be useless if they are not able to be correctly interpreted, or if variables are given confusing or obscure names. User-oriented aspects are important to bear in mind as they give an indication of the level of skill required of a modeller, and this can be assessed in the evaluation process. This is important since blind model validation exercises have shown many times that different users of the same model can produce very different results for the same scenarios (e.g. Rein et al., 2007).

Sensitivity analysis

In the verification and validation process, the user needs to consider the effects of errors and the level of uncertainty when using models. There are uncertainties in both experimental data and model predictions that need to be taken into account when evaluating a model. Sensitivity analysis may be viewed as an initial test to find which model input parameters are important and their effect on output. Uncertainty analysis is a further step which quantifies variations in the input parameters and the effect on the output.

There are a number of different types of uncertainty: model form, model setup, model solution and experimental uncertainty. Model form uncertainty is that which arises from whether the assumptions made in the conceptual model are adequate and validation can be used to provide information on the relevance of the model form. Model setup uncertainty relates to propagation of uncertainty arising from model input and boundary conditions and other model parameters. Model solution uncertainty is uncertainty due to the numerical methods used. This uncertainty can be reduced in the verification process. Experimental uncertainty should be considered during the validation process, but is very difficult to quantify, especially for older experimental data sets as discussed previously. Information on the type of sensors, their accuracy and calibration as well as other parameters, may become lost, or may never have been recorded in the first place. More recently, with comprehensive electronic databases and information systems, storing this kind of information and making it available should be less of a problem. Experimental uncertainty can be allowed for when using SPMs where the uncertainty in observations is not explicitly specified, but recognised by the range of SPM values a "good" model can take (this will be reviewed in the next Section). The Fire Dynamics Simulator (FDS) validation manual (McGrattan et al., 2015) provides another possible approach to this problem. Measurement uncertainties from well documented experiments (Hamins et al., 2006) are combined with engineering judgement to estimate uncertainties for experiments where that information is not available. This approach is only available where the experiments performed and the instruments used are similar.

How will the results of the evaluation be used?

Before carrying out an evaluation, it is necessary to consider how the results will be used. Will the model be declared "acceptable" against some criteria? Should some kind of safety factor to be applied to the model when it is used in practice? Should the assessor define certain applications where the model can and cannot be used?

One of the most difficult aspects of model evaluation is determining what constitutes an "acceptable" model or defining values for qualitative and quantitative criteria. In some ways, defining qualitative acceptance criteria is the simpler of the two, because if a particular feature is missing from a model, then it may be deemed unable to model a particular phenomenon. Determining absolute values of quantitative criteria is more difficult because it relies, to a certain extent, on the results of previous model evaluations and on building up experience in a particular area. Atmospheric dispersion modelling is an area where there is a relatively large amount of experience as many of the evaluation studies report the results of statistical analyses. Examples are Zapert et al. (1991) and Hanna et al. (1993), the latter going on to suggest SPM values for better-performing models.

A validation exercise reported by Carissimo et al. (2001) included over 300 sets of model results and associated SPM values. Similarly, Chang and Hanna (2004) analysed the results of a large number of atmospheric dispersion model runs and made suggestions for values of performance measures expected of a "good" model. Following on from this, Ivings et al. (2007, 2016) used the results of these studies to suggest model acceptance criteria for liquefied natural gas (LNG) dispersion models. The MEG (Mercer et al., 1998) did not provide any acceptance criteria, but suggested that the evaluator may either draw their own conclusions from the statistical analysis, or when comparing numerous models, select the best performing model for their application. This is, of course, subjective.

The acceptance criteria proposed by Ivings et al. (2007, 2016) were based on atmospheric dispersion and their values reflect its stochastic nature. Care is needed in adopting the same acceptance criteria for other scenarios which do not have the same level of inherent uncertainty. An example would be in the assessment of dispersion indoors in still air which is not governed by atmospheric wind or turbulence. For this case, it may be appropriate to adopt a narrower range of acceptance for a model, to reflect the lower uncertainty of the process. However, selecting the range could be difficult depending on the evaluation data available. Acceptance criteria will also depend on the quantity that is being predicted. Different criteria may therefore be used for the same problem if more than one parameter is being compared. For example, in the case of an internal explosion in a sealed vessel, the maximum overpressure may be straightforward to predict but the flame arrival time may be less certain. Setting acceptance criteria for scenarios other than atmospheric dispersion is an area where relatively little progress has been made. To do so would require a systematic analysis of a large amount of experimental and simulation data which would be a significant undertaking.

The role of best practice

The issue of user dependence in modelling was mentioned in a previous Section. Following best practice guidance in modelling is a way of reducing this user dependence, especially for complex models, or those which require a high level of user-input, such as CFD. Numerous general best practice guides are available in this respect (e.g. ERCOFTAC, 2000; Lacome and Truchot, 2013; OECD, 2015). Model evaluation can be seen as part of ensuring model quality and therefore, a number of model evaluation studies have also included best practice guidance (Franke et al., 2007; COST, 2015; SUSANA, 2016). Best practice guidance is not necessarily a part of the model evaluation process, but along with user-oriented aspects, can help to address the issue of model input or "garbage in".

Putting it all together

So, to attempt to answer the question posed by the title, "How do we demonstrate that a consequence model is fit-for-purpose?", it could be argued that a fit-for-purpose model is one for which all the stages in the model evaluation process have been carried out, to a level in line with the intended use of the model. Therefore, a model would need to be scientifically robust, have been verified, validated for a particular use, with some idea of the uncertainty of the predictions, and have clear guidance on operation of the model presented to the user.

Part of the evaluation process is to address the "why" and "how" questions and consider any pre- and post-evaluation tasks (CERC, 2000), meaning that an evaluation is not carried out without forethought and planning. These tasks need not be particularly onerous, and may simply involve defining who is to carry out the various parts of the evaluation. Post-evaluation tasks involve the evaluator providing feedback on the evaluation protocol – an assessment of the method by an end-user. This is an important aspect as the details of evaluation methods are not necessarily set in stone and may need to be updated following an application to a model (Ivings et al., 2007, 2016).

Conclusions - where are we now?

The previous Sections have outlined the stages in the model evaluation process. Collecting these stages into model evaluation protocols provides a structured and, importantly, documented way of assessing a model. Recent activity in model evaluation has also made progress towards the recommendations made by Olesen (1996) that model evaluation methods and corresponding software must be developed and be freely available and that protocols for specific applications should be developed and their usability thoroughly tested. Two examples of recent projects are the EU SUSANA project (Baraldi et al., 2016) and the SAPHEDRA[4] platform for evaluation of consequence models. Despite the increasing number of evaluation activities and protocols, the number of published examples remains relatively small. One of the reasons for this may be that there is no regulatory requirement in the EU to evaluate models in the same way as exists in the US for Liquefied Natural Gas (LNG) vapour dispersion, or for fire modelling in nuclear applications. Model evaluation as an isolated activity is time consuming, with associated cost, especially when model development is on-going and this lends weight to the idea that evaluation is an ongoing process of model improvement, rather than a pass/fail test.

Acknowledgements

This publication and the work it describes were funded by the Health and Safety Executive (HSE). Its contents, including any opinions and/or conclusions expressed, are those of the authors alone and do not necessarily reflect HSE policy

References

Balci, O., (1986), Credibility assessment of simulation results: The state of the art, Technical Report TR-86-31, Department of Computer Science, Virginia Polytechnic Institute and State University.

Barad, M. L., (1958), Project Prairie Grass--A field program in diffusion, Vols I and II. Geophysical Research Paper No. 59, Air Force Cambridge Research Center, Bedford, Massachusetts, NTID PB 151424, PB 1514251.

Baraldi, D., Melideo, D., Kotchourko, A., Ren, K., Yanez, J., Jedicke, O., Keenan, J., Makarov, D., Molkov, V., Giannissi, S. G., Tolias, I. C., Venetsanos, A. G., Coldrick, S., Slater, S., Verbecke, F., Duclos, A., (2016), HYMEP The Model Evaluation Protocol for CFD analysis of hydrogen safety issues, presented at 21st World Hydrogen Energy Conference 2016. Zaragoza, Spain. 13-16th June.

Buxton, H. W. and Hyman, A., (1988), Memoir of the Life and Labours of the Late Charles Babbage Esq., F.R.S., MIT Press, ISBN: 0262022699.

Carissimo, B., Jagger, S. F., Daish, N. C., Halford, A., Selmer-Olsen, S., Perroux, J. M., Wurtz, J., Bartzis, J. G., Duijm, N. J., Ham, K., Schatzmann, M., and Hall, R., (2001), The SMEDIS database and validation exercise, International journal of environment and pollution, Vol. 16, No 1-6, pp 614 – 629.

Cambridge Environmental Research Consultants Ltd (CERC), (2000), SMEDIS Model Evaluation Protocol, Version 2.0, Ref. No. SMEDIS/96/8/D.

Chang, J. C. and Hanna, S. R., (2004), Air quality model performance evaluation, Meteorology and Atmospheric Physics, Vol. 87, pp 167 – 196.

Chang, J. C. and Hanna S. R., (2005), Technical descriptions and user's guide for the BOOT statistical model evaluation software package, Version 2.0, Boot Tech & User Guide V2.01.

COST ES1006, (2015), Best practice guidelines for the use of atmospheric dispersion models in emergency response tools at local-scale in case of hazmat releases into the air, Evaluation, improvement and guidance for the use of local-scale emergency prediction and response tools for airborne hazards in built environments.

Duijm, N. J. and Carissimo, B., (2001), Evaluation methodologies for dense gas dispersion models, in Fingas M. F., (Editor), Hazardous Materials Spills Handbook, McGraw-Hill, ISBN 0-07-135171-X.

ERCOFTAC, (2000), Special Interest Group on Quality and Trust in Industrial CFD: Best Practice Guidelines, Version 1.0.

[4] https://projects.safera.eu/project/14 (accessed 22nd November 2016)

Ermak, D. L., (1988), Field validation of dispersion models for dense-gas releases, Lawrence Livermore National Laboratory report UCRL-98139

Ermak, D. L. and Merry, M. H., (1988), A methodology for evaluating heavy gas dispersion models, Lawrence Livermore National Laboratory report UCRL-21025.

Fox, D. G., (1981), Judging air quality model performance, Bulletin of the American Meteorological Society, 62, pp 599-609.

Franke, J., Hellsten, A., Schlünzen, H. and Carissimo, B., (2007), Best practice guideline for the CFD simulation of flows in the urban environment, COST Action 732, Quality assurance and improvement of micro-scale meteorological models, Meteorological Institute, University of Hamburg.

Gant , S. E., Bettis, R., Coldrick, S., Burrell, G., Santon, R., Fullam, B., Mouzakitis, K., Giles, A. and Bowen, P., (2016a), Area classification of flammable mists; summary of joint-industry project findings", IChemE Hazards 26 Conference, Edinburgh, UK, 24-26 May.

Gant, S. E., Coldrick, S., Tickle, G. and Tucker, H., (2016b), Impact of alternative model validation methods: A case study on the LNG model validation database using Drift, 17th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (Harmo-17), Budapest, Hungary, 9-12 May.

Gass, S. I., (1977), Evaluation of complex models, Computers & Operations Research, Vol. 4 pp 27-35.

Gass, S. I. and Thompson, B. W., (1980), Letter to the editor guidelines for model evaluation: an abridged version of the U.S. General Accounting Office Exposure Draft. Operations Research 28(2):431-439.

Hamins, A., Maranghides, A., Johnsson, R., Donnelly, M., Yang, G., Mulholland, G. and Anleitner, R.L., (2006), Report of Experimental Results for the International Fire Model Benchmarking and Validation Exercise 3. NIST Special Publication 1013-1, National Institute of Standards and Technology, Gaithersburg, Maryland, May 2006. Joint Publication of NIST and the US Nuclear Regulatory Commission (NUREG/CR-6905).

Hanna, S. R., Chang, J. C. and Strimaitis, D. G., (1993), Hazardous gas model evaluation with field observations, Atmospheric Environment, Vol. 27 A, No 15 , pp 2265 – 2285.

Hanna, S. R., Messier, T. and Schulman, L. L., (1988), Hazard response modeling uncertainty (a quantitative method), Sigma Research Corporation, Final report.

Hanna S. R., Strimaitis D. G. and Chang J. C., (1991), Hazard response modeling uncertainty (a quantitative method), Volume II: Evaluation of commonly-used hazardous gas dispersion models, Sigma Research Corporation, Final report, Volume II.

Helsel, D. R., (2005), Nondetects and data analysis: statistics for censored environmental data. John Wiley and Sons, ISBN: 9780471671732.

Herring, S. and Huq, P., (2016), Assessing the performance of atmospheric dispersion models, 17th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Budapest, Hungary, 9-12 May 2016.

Ivings, M. J., Jagger, S. F., Lea, C. J. and Webber D. M., (2007), Evaluating vapor dispersion models for safety analysis of LNG facilities: Technical report, The Fire Protection Research Foundation.

Ivings, M. J., Gant, S. E., Jagger, S. F., Lea, C. J., Stewart, J. R. and Webber D. M., (2016), Evaluating vapor dispersion models for safety analysis of LNG facilities: Technical report, The Fire Protection Research Foundation.

Lacome, J-M. and Truchot, B., (2013), Harmonization of practices for atmospheric dispersion modelling within the framework of risk assessment, 15th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Madrid, Spain, May 6-9, 2013.

McGrattan, K., Peacock, R. and Overholt, K., (2014), Fire Model Validation – Eight Lessons Learned, Fire Safety Science - Proceedings of the Eleventh International Symposium, Christchurch.

McGrattan, K., McDermott, R., Weinschenk, C., Hostikka, S., Floyd, J. and Overholt, K., (2015), Fire Dynamics Simulator Technical Reference Guide Volume 3: Validation, NIST Special Publication 1018-3 Sixth Edition.

Mercer, A., Bartholome, C., Carissimo, B., Duijm, N. J., and Giesbrecht, H., (1998), CEC model evaluation group, heavy gas dispersion expert group, final report, European Commission report EUR 17778 EN.

Model Evaluation Group (MEG), (1994), Model evaluation protocol, European Communities, Directorate-General XII, Science Research and Development.

Nielsen, M. and Ott, S., (1996), A collection of data from dense gas experiments, Risø report Risø-R-845(EN), Risø National Laboratory, Roskilde, Denmark.

Oberkampf, W. L., Trucano, T. G. and Hirsch, C., (2002), Verification, validation, and predictive capability in computational engineering and physics, Foundations for Verification and Validation in the 21st Century Workshop, October 22-23, Johns Hopkins University/Applied Physics Laboratory Laurel, Maryland, USA.

OECD, (2015), Best Practice Guidelines for the Use of CFD in Nuclear Reactor Safety Applications – Revision, OECD Nuclear Energy Agency (NEA), Committee on the safety of nuclear installations report NEA/CSNI/R(2014)11.

Olesen, H. R., (1994), European coordinating activities concerning local-scale regulatory models, in: Gryning, S. and Millán, M. M. (Editors), Air pollution modeling and its application X, ISBN: 978-1-4613-5734-6.

Olesen, H. R., (1996), Toward the establishment of a common framework for model evaluation, in: Gryning, S. E. and Schiermeier, F. (Editors), Air pollution modeling and its application XI, ISBN: 978-1-4613-7678-1.

Rein, G., Empis, C. A. and Carvel, R., (Eds), (2007), The Dalmarnock Fire Tests: Experiments and Modelling, School of Engineering and Electronics, University of Edinburgh, ISBN 978-0-9557497.

Roache, P. J., (1998), Verification and validation in computational science and engineering, Hermosa Publishers, ISBN 0-913478-08-03.

Stewart, J. R., Coldrick, S., Lea, C. J., Gant, S. E. and Ivings, M. J., (2016), Guide to the LNG model validation database version 12, Final report for the Fire Protection Research Foundation, Quincy, Massachusetts, USA.

SUSANA, (2016), SUSANA D3.2, Guide to best practices in numerical simulations (Report of the SUSANA project, funded by Fuel Cells and Hydrogen Joint Undertaking (FCH JU). Grant agreement No. 325386).

US GAO, (1979), Guidelines for model evaluation, US General Accounting Office report PAD-79-17.

Van Horn, R., (1971), Validation of simulation results, Management Science, Vol. 17, No. 5, pp 247-258.

Venkatram, A., (1988), Topics in applied dispersion modeling, in: Venkatram, A. and Wyngaard, J. C., Lectures on air pollution modeling, American Meteorological Society, ISBN 0-933876-67-X.

Webber, D. M., Gant, S. E., Ivings, M. J., and Jagger, S. F., (2009), LNG source term models for hazard analysis, A review of the state-of-the-art and an approach to model assessment, Health and Safety Executive research report RR789, available from: http://www.hse.gov.uk/research/rrpdf/rr789.pdf (accessed 22-11-2016).

Yu, S., Eder, B., Dennis, R., Chu, S.H. and Schwartz, S.E., (2006), New unbiased symmetric metrics for evaluation of air quality models, Atmospheric Science Letters, 7(1), pp 26-34.

Zapert, J. G., Londergan, R. J. and Thistle, H, (1991), Evaluation of dense gas simulation models, US EPA report EPA-450/4-90-018 by TRC Environmental Consultants report under EPA contract No. 68-02-4399.