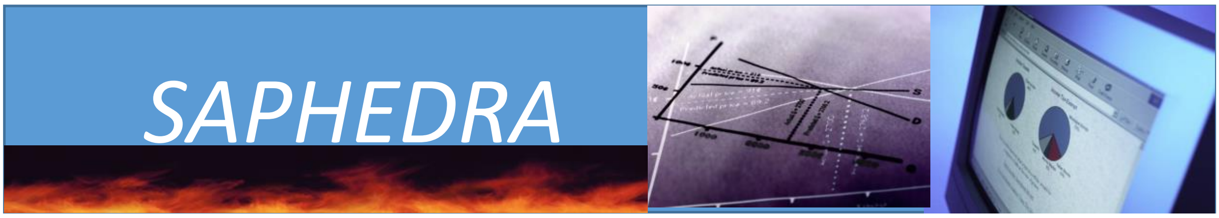
March
2016

S
A
P
H
E
D
R
A

SAPHEDRA - Building a European Platform for evaluation of consequence models dedicated to emerging risk

Review and analysis of previous model evaluation protocols

SAPHEDRA is a project funded in the framework of the ERA-NET SAF€RA "Coordination of European Research on Industrial Safety towards Smart and Sustainable Growth"



SAPHEDRA - Building a European Platform for evaluation of consequence models dedicated to emerging risks

The research leading to these results has received national funding in the framework of the ERA-NET SAF€RA "Coordination of European Research on Industrial Safety towards Smart and Sustainable Growth" (FP7 / 2007-2013 / GA n° NMP4-SA-2012-319092).

Report D3 "Review and analysis of previous model evaluation protocols"

Published by
Health and Safety Executive, acting through
Health and Safety Laboratory
Harpur Hill, Buxton
Derbyshire SK17 9JN
UNITED KINGDOM

HEALTH & SAFETY
LABORATORY

www.hsl.gov.uk

Authors:

Simon Coldrick, Harvey Tucker, Adrian Kelsey

Contributors:
SAPHEDRA Consortium

Buxton, March 17, 2016

Harpur Hill, Buxton
Derbyshire, SK17 9JN
T: +44 (0)1298 218000
F: +44 (0)1298 218986
W: www.hsl.gov.uk

HEALTH & SAFETY
LABORATORY

SAPHEDRA project work package 3: Review of model evaluation protocols

MSU/2015/22

Report Approved for Issue By:	**Charles Oakley**
Date of Issue:	**17-03-2016**
Lead Author:	**Simon Coldrick**
Contributing Author(s):	
Technical Reviewer(s):	**Adrian Kelsey**
Editorial Reviewer:	**Mat Ivings**
HSL Project Number:	**PH06328**

HSE Report
This report 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.

DISTRIBUTION

Harvey Tucker	HSE CEMHD 5
Adrian Kelsey	HSL Mathematical Sciences Unit
Mat Ivings	HSL Mathematical Sciences Unit
Simon Gant	HSL Mathematical Sciences Unit
Charles Oakley	HSL Mathematical Sciences Unit
Mike Wardman	HSL Major Hazards Unit

SAPHEDRA project partners:
INERIS
BAM
DEMOKRITOS
RIVM
TNO
UniBo

ACCESS CONTROL MARKING: Available to the public

Report Approved for Issue by:	**Charles Oakley**
Date of issue:	**17-03-2016**
Lead Author:	**Simon Coldrick**
Contributing Author(s):	
HSL Project Manager:	**Nat Winfield**
Technical Reviewer(s):	**Adrian Kelsey**
Editorial Reviewer:	**Mat Ivings**
HSL Project Number:	**PH06328**

ACKNOWLEDGEMENTS

The author would like to acknowledge the support and advice of Harvey Tucker (HSE) in preparation of this document and Adrian Kelsey (HSL) for informative discussions and suggestions.

EXECUTIVE SUMMARY

Objectives

The SAPHEDRA project is an EU-wide project on the evaluation of consequence models used in risk assessment studies for hazardous materials. The overall aim of the project is to derive an EU commonly agreed model evaluation protocol for consequence models and a series of example applications based on well-established experiments. SAPHEDRA is composed of seven work packages. This report forms the output of work package three which is to review existing model evaluation protocols and to make recommendations for the structure and content of a new evaluation method that can be broadly applied to models used in risk assessment studies for hazardous materials.

Main Findings

Model evaluation has been in existence since the early use of computer simulations and techniques were developed which could be applied to a wide range of fields. Much of the early work on model evaluation focussed on models where the outputs were used in support of policy decisions of some kind and a decision maker needed to be assured that the model output was a scientifically robust and reliable description of the actual process.

Atmospheric dispersion is an area where there has been significant activity in model evaluation. The main drivers in this area were the need to assess the risks from spills of hazardous substances and the introduction of air quality laws which led to a requirement to model air pollution. In both cases, the underlying need was to demonstrate that results were robust and reliable to a decision maker independent of, and far removed from, the modelling process. Numerous model evaluation protocols and model comparison exercises appeared as a result.

Fewer model evaluation studies exist in other areas of consequence modelling, such as fire, explosion and source term models. That does not mean that model evaluation does not take place in these fields, but that it takes a different form as quality assurance of simulations is important in sectors such as the nuclear industry. For fire modelling, standards and benchmark studies are more prevalent than model evaluation protocols. Computational modelling of explosions is also not as well established as dispersion modelling and many of the techniques are still in development rather than in routine use for consequence assessment calculations.

One of the main findings of the review was that model evaluation protocols fall into two categories; they are either very generic and can be applied to any consequence modelling area, or they are very specific and have a particular area of application. Those that are very generic tend to need a high level of effort on the part of the evaluator in tailoring them to their specific application. Those that are very specific require less effort by the evaluator, but have a fairly narrow area of application.

A second finding of the review was that the number of published applications of model evaluation protocols is fewer than the number of protocols. This may be partly due to the fact 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.

Recommendations

Based upon previous experiences of model evaluation derived from the literature, as well as experience of creating and applying model evaluation protocols, this report recommends that a model evaluation protocol for consequence models should follow an established structure of:

- Pre-evaluation tasks
- Scientific assessment
- User-oriented assessment
- Verification
- Validation
- Sensitivity and uncertainty analysis
- Post evaluation tasks

The report also makes a number of recommendations on how each of the stages may be undertaken.

CONTENTS

1 INTRODUCTION

Within the European Union, the Seveso Directive is the main legislation dealing with the control of on-shore major accident hazards involving dangerous substances. The Seveso III Directive came into force on 1 June 2015 and requires a detailed risk assessment to be undertaken when the estimation of the consequences of major accidents is an input for decision-making. Predictive models are used in this estimation and therefore directly influence the decision making process. For this reason, decision makers relying on these models need to understand their performance and limits of applicability. Model evaluation is a process that can be used to provide assurance of the robustness of predictions and to guide improvements in the modelling techniques. Consequence model evaluation has been a significant area of activity for many years and there have been several European initiatives on harmonisation and evaluation.

The use of materials and technologies constantly evolves, leading to new scenarios for which current models and tools may not have been evaluated. Adopting a harmonised approach to model evaluation can build on existing experience and at the same time, update procedures for new and emerging technologies and materials. A need for such an approach was recognised and a project initiated under EU funding to provide a means of objectively assessing the performance of models and related simulation tools. The overall aim of this project is to derive an EU commonly agreed model evaluation protocol and a series of test cases derived from well-established experiments. The project, titled “SAPHEDRA”, involves a consortium of seven European partner organisations and consists of regulators, research establishments and academic institutions:

INERIS, France (coordinator)

BAM, Germany.

DEMOKRITOS, Greece.

Health and Safety Laboratory, UK.

RIVM, Netherlands

TNO, Netherlands.

UniBO, Italy.

The project is composed of seven work packages assigned to the partners as follows:

WP1 - Identification of the existing tools for modelling hazardous phenomena, led by TNO.

WP2 - Gap analysis of existing modelling tools in emerging risks management, led by UniBO.

WP3 - Review and analysis of previous model evaluation protocols, led by the Health and Safety Laboratory.

WP4 - List of experimental campaigns and information available to be used to evaluate existing tools or new tools, led by BAM.

WP5 - Definition of a complete, new and robust procedure to evaluate modelling tools, led by INERIS.

WP6 - Application of the new procedure to evaluate dispersion, fire and explosion models to a case study, led by DEMOKRITOS.

WP7 - Project coordination, led by INERIS.

This report forms the output of work package three, the review and analysis of previous model evaluation protocols. Section two of the report introduces the origins, concepts and terminology surrounding model evaluation. The remainder of the report is divided into two main activities; Section three contains a review of existing evaluation protocols and Section four contains lessons learned and makes recommendations for the new evaluation procedure.

2 CONCEPTS AND TERMINOLOGY

The literature on model evaluation employs a number of standard terms involved in the various stages of model evaluation and the following Sections provide a review of those terms. Many of the topics are research activities in their own right and a significant amount of material has been published in the different areas. The model development and evaluation process collects together these areas in a structured way with the goal of an overall assessment of model performance.

2.1 MODELLING AND MODEL EVALUATION

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 even impossible to achieve by other means. In consequence and hazard assessment, the ability to predict things, rather than measure them or wait to learn from an accident, is an essential tool. In the context of hazard assessment, the terms "model" and "simulation" are often used interchangeably because the model is often of little use without being implemented in software. However, since these models are a representation of reality, rather than reality itself, there is a need to examine the correspondence between the model and the real world. This is an element of computer simulation that has been studied since the early use of computer predictions. 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. The activity of model evaluation is used in part to help answer the question "how much confidence do we have in the predictions?" Model evaluation is not a single activity, but a collection of stages that need to be undertaken before that question can be fully answered.

One of the main reasons for evaluating models, according to the US General Accountability Office (US GAO, 1979) is to inform a decision maker, who may be 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.

The steps undertaken in model development are shown in Figure 1, taken from Ivings *et al.* (2007). 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 a higher fidelity model might involve a more complex set of equations which in turn require a more sophisticated solution method. In any case, model evaluation is present in each of the steps:

- The set of equations chosen must actually represent the process being modelled
- The solution method must function properly
- The solution must match that observed in reality

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.

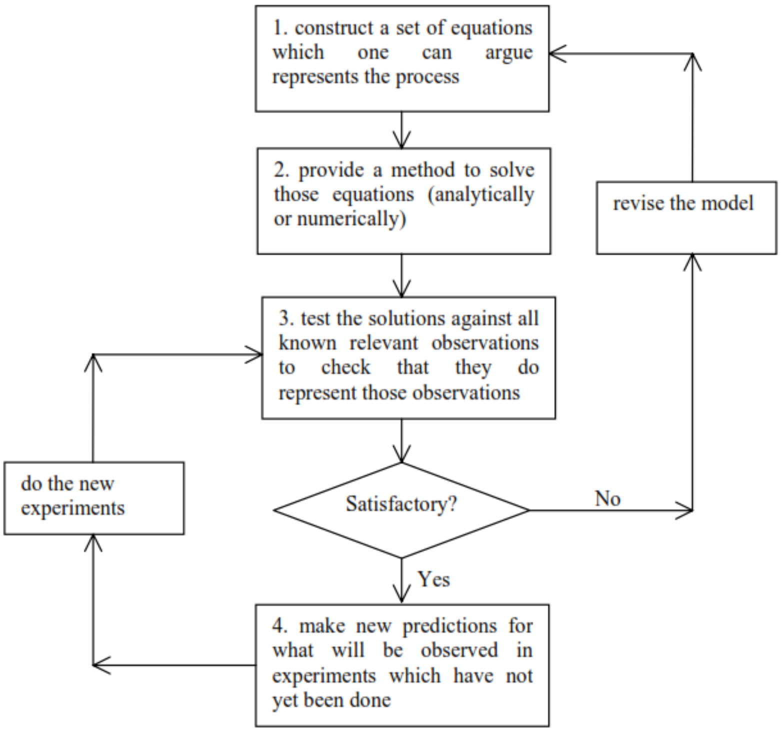

1. construct a set of equations which one can argue represents the process
2. provide a method to solve those equations (analytically or numerically)
revise the model
3. test the solutions against all known relevant observations to check that they do represent those observations
do the new experiments
Satisfactory?
No
Yes
4. make new predictions for what will be observed in experiments which have not yet been done

Figure 1 The steps of model development

2.2 CLASSIFICATION OF MODELS

A vast array of models is available for carrying out consequence assessment and these range from simple one-dimensional phenomenological models to sophisticated three dimensional Computational Fluid Dynamics (CFD) simulations. At the simplest end of the scale, a model may not even require solution by a computer, but the steps outlined in Figure 1 will still apply, and a model evaluation can still be carried out. For a more complex model or simulation, the evaluation process will be much more involved. This means that model evaluation must not only be broadly applicable, but the specific activities need to be tailored to the particular model type. Different model types will have different capabilities and this is an important part of the evaluation process. Therefore, some general classifications of models are given in the following Sections. These classifications are not fixed, because as noted by Duijm and Carissimo (2001), there are different methods of classification, such as the number of spatial dimensions or the solution technique. It is also worth noting that some software packages incorporate a suite of models and sub-models so that the extent of the evaluation needs to be clearly defined.

2.2.1 Correlations and phenomenological models

These models relate one quantity to another empirically. An example is a model for the decay of concentration with downstream distance in a gas jet, which in its simplest form could be a curve fit achieved by adjusting some parameters. The predictive capability of this model is limited by the available experimental data. A more sophisticated version could include some description of the physics of the jet so the model could be tuned to one dataset and applied in a predictive sense, providing the physical basis of the model is sound.

2.2.2 Integral models

A definition of an integral model is given by Ivings *et al.* (2007) as one composed of a few, partly phenomenological, equations to describe overall properties (the integral properties) of a flow. For example, an integral model for dispersing gas clouds may include a simple description of the atmosphere, along with relationships for how the cloud moves downwind and entrains air. Many integral models include a series of entirely separate sub-models, where the output from one sub model is fed into another. Integral models often reduce the process to a single dimension and typically consist of ordinary differential equations. These equations may require advanced solution techniques requiring input from the user.

2.2.3 Shallow layer models

Shallow layer models are intermediate in terms of complexity and dimensions, typically involving solutions of partial differential equations averaged over one dimension. They can be thought of as a simplified two-dimensional CFD model where the equations describe the variation of a property over an area, but with properties averaged over the depth dimension. An example would be the model for a spreading liquid pool or heavy gas cloud, where variations in terrain height can be accounted for. This classification of models is aimed principally at dispersion and, in common with CFD models, shallow layer models involve the numerical solution of the equations over a discrete grid designed by the model user.

2.2.4 CFD models

CFD involves the solution of partial differential equations and sub-models in several dimensions. Typically they are the time and spatially varying fluid flow equations over a discrete grid covering the domain of interest. This highly generalised approach means that CFD modelling is heavily user-dependent. Setting up a typical CFD model requires the user to construct geometry and to generate a grid within this geometry using the available meshing techniques. They must then decide which flow equations are to be solved for their particular problem and the numerical settings needed to solve those equations accurately and efficiently. The flexibility and wide range of applicability of modern CFD codes also mean that the user is presented with numerous other controls which can affect the final solution or how it is arrived at. To develop a model for a simple gas jet using CFD would still require the same stages as outlined in Figure 1, but the process of evaluation would become much more involved.

2.3 CLASSIFICATION OF USERS

Model development and evaluation may involve the interaction of several different groups, each of whom is responsible for a particular area. Before looking at the model evaluation process, it is useful to see how these groups are defined in relation to it. Their definition depends also in part on the type of model, its intended use and its level of complexity and is by no means clear cut. Balci (1986) provides a simple distinction between model builders and model users in terms of the risk of making certain errors. The model builder is at risk of rejecting results when they are sufficiently credible and the model user is at risk of accepting results that are not sufficiently credible. Roache (1998) refers to three distinct teams involved in CFD analysis of engineering problems: code developers, physical model experts and users. Code developers work on the mathematics, software and physical model development. Physical model experts examine the correspondence between the model and what it represents. Users specialise in the applications of the code, but are not necessarily experts in code or model development. The roles of these teams may be interrelated and there is the potential for significant overlap. For the case of the simple phenomenological model, it may be that the model developer and user is the same person. On the other hand, for complicated, general purpose CFD software, there may well be distinct teams responsible for the development and testing of the code. The user is then a third party who trusts that the vendor has done everything to ensure that the code meets its specification. For this review, it is possible to define two groups along the lines of Balci (1986), who are model developers and model users. In a regulatory setting, two other groups could be added which are model evaluators and decision makers and a possible set of interactions between the groups is shown in Figure 2.

2.3.1 Model developers

Model developers define and construct the model and implement it in software. They provide assurance that the model has been correctly implemented. Model developers may also be software vendors. In the case of CFD software, the developers provide a pre-compiled kit of tools, models and sub models.

2.3.2 Model users

Model users may also be known as practitioners or analysts. They use the models or software provided by the developers, but are not necessarily experts in development aspects. They may instead be experts in model application for particular physical scenarios. Again, for CFD software, the users apply the kit of tools and models to their particular scenario. This aspect means that CFD falls into the area of overlap in Figure 2, because the user is constructing a model and therefore takes on some of the responsibility of a developer.

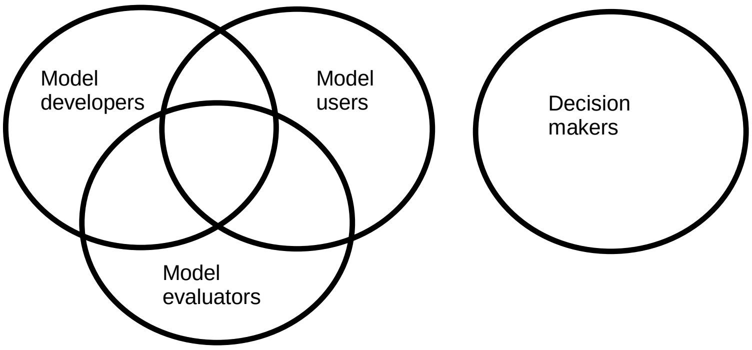


Figure 2 Interaction between modelling groups

2.3.3 Model evaluators

The aim of the model evaluator is to carry out an assessment of a model and report the results in a format appropriate for the target audience. In some cases, the model evaluator may be the model developer, because development involves some of the steps in the evaluation process, and because evaluation is an essential part of model development. Model users may also need to carry out some evaluation activities to provide themselves with assurance of their results. A model evaluator may also be a third party who is separated from the model development process and who is carrying out the evaluation on behalf of a decision maker.

2.3.4 Decision makers

Many of the early publications on model evaluation (Gass, 1977, for example) suggest that the context of model evaluation is one in which an independent analyst evaluates a model and makes a recommendation to a decision maker as to whether the model can be used with confidence. The decision maker is then a party who is some distance removed from the model development and application process, but who must be aware of the model results and limitations. The concept of independent evaluators and decision makers is perhaps more relevant in a regulatory framework where there is a need to communicate results to non-experts.

2.4 VERIFICATION AND VALIDATION

Verification and validation are two important elements within model evaluation that have appeared since the early publications by authors such as Van Horn (1971) and the US GAO (1979). Even recently, with the profusion of literature on the subject, there is still some confusion and debate over what each relates to. Balci (1986) cites sixteen commonly used terms under the umbrella of model credibility, but offers the following succinct differentiation:

Verification: building the model right

Validation: building the right model

The following sections give what are generally considered to be the accepted meanings within the mathematical modelling community:

2.4.1 Model 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 SQA (software quality assurance).

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 CFD because code verification is an activity that might be expected of the developer, or software vendor. Vendors are supplying 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.

2.4.2 Model validation

Gass (1977) describes validity as an umbrella term for correspondence of a model with the real world and refers to "predictive validity" as being the relationship between actual outcomes and the predicted outcomes generated by the model. Van Horn (1971) defines validity as the process of building an acceptable level of confidence that an inference about a simulated process is a correct or valid inference for the actual process. He also notes that validation will not necessarily prove that the simulation is a correct or true model of the real process. This is highlighted by Ivings *et al*. (2007) who 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". A validated model is therefore one where tests have been performed which could have shown it to be invalid, but which failed to do so. However, the idea that validation is a process is also supported by Roache (1998) who suggests that verification is something that can be complete, but validation is something that can be ongoing. Roache (1998) offers numerous definitions for validation, extracted from the literature, but notes that validation ultimately involves the comparison of code predictions with physical experiments. Where physical experiments or reality are the ideal means for comparison, it is also worth remembering a definition of validation offered by the US GAO (1979): "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. Processing experimental data and its influence on the validation process is discussed further in Section 4.2.8.

2.4.3 Calibration

Calibration is an activity that regularly appears in connection with evaluation, verification and validation. Calibration bears some resemblance to validation in that it involves checking the output of a model against physical test data. For experimental work, calibration is normally taken to mean determining the accuracy of measuring instruments (Roache, 1998). But calibration of a model is not necessarily determining its accuracy; rather it can be seen as an adjustment of model parameters needed to fit experimental data (Roache, 1998). Derwent *et al.* (2010) note that, in air quality modelling, if all model parameters were geophysical constants then there would be no need for calibration. Unfortunately, this is not the case and some calibration is required in most types of modelling. Calibration can therefore be equated to model tuning. Derwent *et al.* (2010) highlight one particular issue with calibration that is relevant to model evaluation. This is that the steps taken by model developers to calibrate and tune model parameters can become lost in the mists of time. Then a model can end up being "validated" in part against the data which was used to calibrate it in the first place. Determining whether or not this is the case may be difficult because a model may be optimised over a range of experiments, rather than tuned individually against single experiments.

2.4.4 Benchmarking

Benchmarking is often taken to mean the comparison of the results of different models against each other, perhaps when applied to a particular test. Often an experiment is referred to as a benchmark experiment when it is used in an intercomparison exercise. 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. 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. A need to carry out an indirect validation might arise because of unavailability of the original experimental data. A potentially complicated problem with indirect validation is with the variables used to make the comparison. One could imagine an example of an experiment where only concentration was measured and we want to validate a new model that predicts only temperature. Suppose we have a benchmark model which is validated against these experiments and which predicts both concentration and temperature. Would it be an acceptable indirect validation of our new model using the benchmark when temperature was not measured in the original experiment? There is no clear answer to this question as it may be that an indirect validation is the only feasible approach for a particular scenario.

2.4.5 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. A key issue is that experimental data may only be available at reduced scale, but that the model will be used at full-scale. The model physics needs to be correct in order to have confidence in scaling-up the model from the experiments to reality. Validation may be associated with determining numerical values which represent the goodness-of-fit of a model with the experimental data. 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. 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 can only be gained in a model 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 data sets. In these cases, most of the evidence may be provided by the scientific assessment. Such an approach was recommended by Webber *et al*. (2009) (See Section 3.12).

2.4.6 Qualitative and quantitative criteria

Qualitative and quantitative criteria are terms associated with scientific assessment and validation and are used to help define whether or not a model is “acceptable.” Qualitative criteria are often a list of features that are required in a model for a particular application. In dense gas dispersion modelling, the criterion of “model accounts for gravity driven spreading” is an essential feature and therefore a qualitative criterion that must be met. Other qualitative criteria may list features that are desirable, but not essential and a judgement made by the evaluator that a model missing those features can still perform adequately. Quantitative criteria are associated with validation where performance metrics are calculated, comparing the model output with experimental data (Statistical Performance Measures, SPM, are described in Section 4.2.9). The criteria are set by attaching acceptance values to the performance metrics, for example that a certain fraction of the predictions are within a factor of two of the observations. Setting these criteria is not straightforward and requires judgement, experience and information on how a “good” model can perform. It may also require an understanding of how much inherent uncertainty there is in the experimental data.

3 REVIEW OF EXISTING MODEL EVALUATION PROTOCOLS

3.1 MODEL EVALUATION IN GENERAL

The importance of model evaluation has been identified since the early use of computer simulations. Van Horn (1971) discusses methods of validation[1] of computer simulations of systems which may be economics, human behaviour, management science, or engineering and physicial processes. Irrespective of the type of simulation, Van Horn (1971) notes that such systems are characterised by:

- The structure and parameters of the process are determined by the environment, not by the modeller
- Part of the process depends on physical phenomena
- People are part of the process either as information processors or as decision makers

The evaluation activities should therefore take into account these characteristics, which could equally apply to consequence modelling. Van Horn (1971) lists a three stage approach suggested by Naylor and Finger (1967) which he notes appears to capture the major ways to build confidence in a model, namely:

1. Constructing a set of hypotheses and postulates for the process using all available information- observations, general knowledge, relevant theory and intuition
2. Verifying the assumptions of the model by subjecting them to empirical testing
3. Comparing the input-output transformations generated by the model to those generated by the real world

Van Horn (1971) also notes that while statistical tests are important, the overall evaluation process should encompass much more.

In a paper titled “Evaluation of complex models” Gass (1977) lists the central activities as verification and validation, using concepts that are highly relevant today. Verification is defined as ensuring that the mathematical relationships, computer program and data elements represent the desired model under all anticipated conditions. Gass (1977) suggests that the verification process is the jurisdiction of model developers, but a review of this verification process should be part of the evaluation activity and must be reported by the evaluators. Gass (1977) uses the term “validity” to encompass a number of activities, where the task of “model validity” is the correspondence between the model assumptions and hypotheses with the problem environment being modelled. This task may well be viewed as a form of scientific assessment.

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 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). The document was

[1] *In this context, “validation” is taken to mean “evaluation.”*

aimed at models informing social, economic, political and military programmes, which the authors note have the following characteristics:

- They are models developed to assist decision makers
- They are mathematical models of complex systems
- They are large scale models

In "Guidelines for Model Evaluation", 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 following five steps:

1. Documentation
2. Validity
    a. Theoretical validity
    b. Data validity
    c. Operational validity
3. Computer model verification
4. Maintainability
    a. Updating
    b. Review
5. Usability

The report also highlights the importance of not only judging a model against certain goals, but also considering its purpose and the manner and environment in which it is to be used. In effect an assessment of its applicability.

Perhaps due to the increasing use of computer models around this time, there was also an increasing concern in the evaluation of gas dispersion models. According to Dickerson and Ermak (1988), the evaluation of atmospheric dispersion models had already been of interest for some time, partly driven by laws and regulations concerning air quality (Fox, 1981). Many of the early evaluation studies were driven by the atmospheric science community where the focus was on the dispersion of tracer gases from an air quality perspective. These early evaluation studies were supported by extensive experimental datasets involving releases of tracer gases at both ground level and from elevated stacks. Around 1980, there was an increased interest in evaluating consequence models, in particular, models used for emergency response planning (Dickerson and Ermak, 1988). Evaluation of emergency response models has a slightly different emphasis than for consequence models used in risk assessment because they may be used in real-time and therefore there is a need to also consider the input meteorological data.

As a result of the various model evaluation programmes and the needs identified by those programmes, there were a number of large scale experiments carried out in both the US and Europe. An extensive review of these datasets is provided by Hanna *et al.* (1988). According to McQuaid (1979) at the time the predictive models gave widely varying results and there was little experimental information on which to base the choice of predictive method. While most of these early campaigns were concerned with producing model validation data, rather than evaluation methods, there was a considerable improvement in the models and a reduction in variation between models (Duijm and Carissimo, 2001).

In the rest of this Section, particular model evaluation protocols, or methods, are reviewed. The Section concludes with a summary and a table showing the main features of the protocols reviewed (Table 1).

3.2 ERMAK AND MERRY (1988)

In the late 1980s, the US Air Force was becoming interested in examining the effects of releases of hazardous chemicals. This arose in part from the need to assess the effects of potential spills of fuels and rocket propellants such as Dinitrogen tetroxide and Hydrazine which are highly toxic and unstable. Specific models such as the "Ocean Breeze" and "AFTOX" models were in use for the purpose of calculating the dispersion of these materials and numerous other models were available. The US Air Force required a methodology for evaluating these models, with the specific requirement of it being quantitative and statistical in nature so that a relatively objective evaluation of model ability could be made. Other desired attributes of the methodology were ability to help identify limitations of models and to estimate the level of confidence with which a validated model could be used. A methodology aiming to meet these requirements was developed by Ermak and Merry (1988) and also summarised in Ermak (1988). While being quantitative and statistical, the methodology re-iterates that the goal of model evaluation is to gain the confidence to apply the model both within and beyond the range of observations used to test the model. Therefore it has many of the attributes of an evaluation protocol through specification of tests including:

- Evaluation of the accuracy of the input meteorological data and the concentration data used for comparisons

- Examination of model structure including the accuracy of the mathematical framework, the realism of the model representation of important physical processes, and the appropriateness of assumptions used in the model when applied to real situations

- Sensitivity analysis of the model to uncertainties in the input meteorological and source data

- Testing of the model predictions against observations including both laboratory and field scale experiments

In addressing the second aspect, Ermak and Merry (1988) suggest the examination of model structure should include the evaluation of the theoretical submodels used to describe the various physical processes of concern. The evaluation should also cover the numerical approach used to implement these submodels, though it is noted that this area of model evaluation is generally performed by the model developer as the model is being created and improved.

The main body of the Ermak and Merry (1988) methodology covers statistical methods for model evaluation and describes a number of different measures of dispersion. The use of bootstrap procedures to estimate confidence intervals is also included. Ermak and Merry (1988) advise that the selection of model performance measures needs to be guided by the intended model application, the nature of the emissions and the associated hazard. They also conclude that such measures can provide misleading guidance unless they are interpreted in light of the accuracy of the experimental data and the physical processes included within the model.

3.3 THE METHOD OF HANNA ET AL. (1988)

In parallel with the work of Dickerson and Ermak (1988), and Ermak and Merry (1988), the US Air Force and the American Petroleum Institute sponsored a further two part project to produce a framework for performing model evaluation and estimating model uncertainty. This project was also based on assessing the impact of spills of fuels and aimed to provide a systematic quantitative method for determining the uncertainty associated with the model predictions.

The first part of this project titled "Hazard response modeling uncertainty (a quantitative method)" by Hanna *et al.* (1988) aimed to review the literature on hazard modelling uncertainty, develop a framework for accounting for model uncertainty and apply the procedures to several models. The second part, titled "Hazard response modeling uncertainty (a quantitative method) volume II evaluation of commonly-used hazardous gas dispersion models" by Hanna *et al.* (1991) aimed to develop software and apply this to a group of models. The second part is more commonly referenced as it contains details of how the experimental data were processed and how the models were evaluated against the data. The main aims of the project were to answer the following questions:

- Do suitable data sets exist for use in evaluating hazardous response models?
- What are the errors in the data used for input to models?
- Is it possible to obtain a number of current models for evaluation purposes?
- Can a model evaluation framework be developed that accounts for all the components of model uncertainty, including stochastic fluctuations?
- Can the models properly handle the effects of sampling and averaging times and distances of concentration measurements?
- What are the confidence bounds on model evaluation statistics such as the mean square error? Are they small enough to permit the relative performance of two or more models to be distinguished?

In part one of the project, Hanna *et al.* (1988) identified three sources of uncertainty in model predictions for emission, transport and dispersion of hazardous gases, namely:

- Errors caused by model physics assumptions
- Random variability (turbulence)
- Errors generated by data input errors

The framework developed in the project aimed to provide a method to attach values to these errors using statistical techniques for estimating the magnitude of each error source in turn.

To enable the uncertainty quantification exercise to be undertaken, Hanna *et al.* (1991) assembled a dedicated database which is the subject of the second part of the project. In addition to the methods used to assemble the database, part two also details the evaluation of fourteen commonly used dispersion models.

The database was named the "Modeler's Data Archive" (MDA) and consisted of data for eight experimental campaigns, with input data sufficient to set up and run models. The experimental campaigns covered both dense and passive dispersion experiments. One of the features of the work by Hanna *et al.* (1991) is that they define the criteria for selecting experimental datasets and also explicitly set out the methods used to process the raw experimental data. This enables the same methods to be applied to other datasets for use in other evaluation studies. For example, Coldrick *et al.* (2009) applied the same data processing methods for a model evaluation database which will be discussed in Section 3.11.

Havens (1992) provides a criticism of the evaluation method of Hanna *et al.* (1991) which is that the sensor location used for the concentration predictions for the DEGADIS model was different to where the measurements were made, and for dense gas dispersion, this can have a significant effect on the results. From the statistical analysis results, Hanna *et al.* (1991) found that the Gaussian plume model performed better than a dense gas dispersion model (DEGADIS). However, Havens (1992) suggests that the Gaussian plume models frequently appear to perform satisfactorily with the experimental data for small to medium size releases, but the results diverge sharply as the release size increases. Therefore good performance against the field data is a necessary, but not sufficient, condition for model acceptance. It is worth noting that Havens (1992) makes use of the term “model evaluation protocol” and lists specific experiments and how a model should be compared with them.

This highlights the issue of getting “the right answer for the wrong reason”, and demonstrates the importance of scientific assessment in model evaluation. Hanna *et al.* (1988) do incorporate an assessment of model physics error by examining the effects of individual model components on the model variance. The idea behind this method was to assess whether increasing model complexity by adding additional physical components is worth the uncertainty those extra components introduce. However, the approach by Hanna *et al.* (1988) does not set out to be a model evaluation exercise in the framework of US GAO (1979) or Van Horn (1971) for example. Instead, it is a method of systematically determining which models best fit the data and the levels of uncertainty in the predictions. Use of such a method in isolation, as in this example, may not uncover important physical aspects such as scaling where a model performs well against the data for a particular experiment but does not incorporate the correct physics to be used at some other scale.

3.4 ZAPERT ET AL. (1991)

The US Environmental Protection Agency (EPA) had an ongoing programme of model evaluation, using model performance measures recommended by the American Meteorological Society (AMS). As part of this programme, a report titled “Evaluation of dense gas simulation models” was produced by Zapert *et al.* (1991). The overall aim of the exercise was the comparison of a number of dense gas dispersion models with data from three field scale experimental campaigns. The report contains a short description of each model and each dataset and describes the release characteristics and whether those features are present in the models.

The authors provide details of how the data were processed, the input conditions and how the models were run. They also describe the statistical comparison and provide tables of statistical results. This report is perhaps similar to Hanna *et al.* (1991) in that it is a systematic statistical analysis of model results but does not include many of the other steps used in model evaluation. For this reason, it can be seen as a validation exercise, rather than model evaluation. The authors do conclude that an equitable “hands off” evaluation is difficult to achieve in practice as many proprietary models have limited documentation and require considerable user experience to be applied effectively. This may be because the evaluation method concentrated on model use and application, and did not consider the user or documentation aspects, which are an important part of model evaluation.

3.5 THE MODEL EVALUATION GROUP

Within the EU, the regulatory framework surrounding the Seveso Directive requires that major hazard accidents are identified and assessed in safety reports, and also that consideration is given to the potential effects on land surrounding plants and installations (“land use planning”). For both aspects, predictive modelling forms an essential part of the process and in the early 1990s, there was a concern that many of the models had not been formally evaluated (Petersen, 1999). Furthermore, around that time, the Joint Research Centre of the European Community had organised a Benchmark Exercise on Major Hazard Analysis with the aim of assessing the state of the art in risk analysis. The exercise highlighted the need for harmonisation of modelling approaches and areas for improvement in the modelling techniques (Contini *et al.*, 1991). A Model Evaluation Group (MEG) was subsequently set up to address these issues relating to model quality and also to identify areas for research into major industrial hazards. The model evaluation group initially produced a very generic model evaluation protocol (MEG, 1994a) and then formed several working groups to generate more specialised versions in the following areas:

- Dense gas dispersion
- Pool fires
- Gas explosions

Each working group was composed of people with more in-depth knowledge specific to their area of application and were therefore able to consider the requirements for a model evaluation protocol (MEP). In addition to the generic model evaluation protocol, the model evaluation group produced a set of “guidelines for model developers” (MEG, 1994b). Of the three working groups formed by the MEG, only two resulted in adapted versions of the MEP. These were the Heavy Gas Dispersion Expert Group (HGD) and the gas explosion group, or Model Evaluation Group Gas Explosions (MEGGE).

3.6 THE MODEL EVALUATION PROTOCOL

The generic model evaluation protocol (MEG, 1994a) is a very brief document and the steps it recommends have much in common with the process set out in “Guidelines for Model Evaluation” (US GAO, 1979). Part of the reason for the simplicity was because the method was intended to be refined and tailored to the specific areas of application. The document outlines the following basic steps:

- Model description
- Database description
- Scientific assessment
- User oriented assessment
- Verification
- Validation

In addition to these steps, the document contains appendices with an example model description, relevance of the database and example database structure along with guidance for each.

3.6.1 Model description

The model description gives details of the purpose of the model, its origins and references to any supporting documentation. By specifying a model's intended area of application, the evaluation process can be tailored accordingly, to avoid misrepresentation of the model. The description also includes the version number, which is an important element, because models are often in a constant state of development. The evaluation can only be relevant to the stated version and any changes or improvements mean the model must be re-evaluated.

3.6.2 Database description

The aim of the database description is to describe the data to be included in the evaluation. This appears to cover any data used in the evaluation and does not distinguish between those used in verification (e.g. derived from analytic solutions) and those used in validation (e.g. field test data). The database description also requires the evaluator to identify, if possible, the data used to tune, or calibrate the model parameters during development.

3.6.3 Scientific assessment

The first activity in the scientific assessment is a statement of its aims. The reason for this is that the assessment may cover only a given aspect of a model if only that part is to be evaluated. This may arise for large models where only a particular component is of interest. In this case, the method of isolating that part of the model should also be included (this may be difficult to achieve in practice). The scientific assessment then proceeds with a model description, an assessment of the scientific content, a statement of the model's limitations and limits of applicability. A final aspect is an indication of the potential areas for improvement of the model.

3.6.4 User oriented assessment

The user oriented assessment follows a similar pattern as the scientific assessment and also allows for only a partial assessment to be made, if it can be justified. The main aspects of the user oriented assessment are to examine the documentation, usability, help system, computational costs and to identify possible improvements. Usability can impact significantly on model results due to unintentional user errors or misunderstanding of input or output. Thus a model may be well verified and validated but give erroneous results through poor usability.

3.6.5 Verification

Verification is covered briefly in that "assessors should ensure that code is producing output in accordance with the model specification." The MEG protocol also suggests that verification is a task for the model developers, and should be recorded in the documentation. The evaluators need to be satisfied that this task has been properly undertaken. The same process was suggested by the US GAO (1979) but the exact approach may depend on the type of model, especially for CFD simulations, where verification may well fall into the grey area between developers, users and evaluators as discussed in Section 2.3.

3.6.6 Validation

The first activity in validation is a statement of its aims and which model parameters are to be tested. Validation then consists of selecting appropriate data and parameters and making an assessment of the uncertainty of both the model and the data. How the model is to be compared with the data appears to be covered by "selection of validation parameters", leaving the choice to the evaluator to decide which technique to use, and whether to adopt a statistical method for example. The final aspects of validation are forming conclusions and making recommendations, and these activities may result in a need to revisit certain aspects as deficiencies may be found in particular datasets.

The very generic format of the MEP means that it is not tied to any particular physical phenomena or modelling technique. This "goal setting" rather than "prescriptive" approach is very similar to the guides by Van Horn (1971), US GAO (1979) and Gass (1977) and arose because the intention was to create a framework that could be adapted by the various working groups. It is an approach that places emphasis and responsibility on the evaluator, rather than the producer of the protocol.

The "Guidelines for model developers" (MEG, 1994b) document sets out requirements for the description of models. The aim of this document was to identify the important features that should be included in documentation from the point of view of both users and evaluators to enable them to judge the model for their own purpose.

3.7 THE HEAVY GAS DISPERSION EVALUATION PROTOCOL

The heavy gas dispersion group produced a final report on their activities in 1998 which contains a completed version of the protocol (Mercer *et al.*, 1998). An earlier seminar publication by Cole and Wicks (1994) also contains a version of the protocol. The tasks of the heavy gas dispersion expert group were:

1) To draw up a list of heavy gas dispersion models

2) To identify datasets

3) To review and adapt the MEG documents

4) To arrange an open exercise to test the protocol

The heavy gas dispersion evaluation protocol was structured according to the MEG MEP and was also based on one produced for the REDIPHEM project (Nielsen and Ott, 1996), which was another CEC sponsored project on model evaluation. A protocol was not delivered at the end of the REDIPHEM project, because the same protocol was issued under the responsibility of the MEG. The heavy gas dispersion MEP defines three classes of models, namely:

1. Phenomenologial models in which the dispersion behaviour is described by a series of nomograms or simple correlations

2. a. Intermediate models (box models or integral models)

   b. Shallow layer models

3. 3D CFD models

The introductory notes suggest that the evaluation is preferably carried out by the model developer or a user. If the evaluation is undertaken by an independent party, then the model developer needs to be informed how the model is to be used so that they can provide input on how to correctly apply the model. This is an aspect that is closely related to the quality of the documentation, which can also inform the evaluator on model use. Sufficiently good documentation may mean that there is little need to interact with the developer, a position that may not happen in practice.

Unlike the evaluation studies by Ermak (1988), Hanna *et al*. (1988, 1991) and Zapert *et al.* (1991), the heavy gas dispersion MEP does not specify a particular set of experiments to be used for the evaluation. Under the section of "Database Description" the text suggests that it is up to the evaluator to identify the relevant data sets and to justify why those are being used. The evaluator must describe properties of the data which may limit it, or make it particularly useful. It is recognised that much useful data already exist in databases but the evaluator may need to consider what quality assurance activities have been undertaken on the data. These may include aspects such as accuracy as well as uncertainty in the measured quantities. The heavy gas dispersion MEP does not specify a list of experiments. Since one of the group's tasks was to identify datasets, the final report (Mercer *et al,* 1998) does contain a list of relevant data. This list of experiments corresponds to the data available in the REDIPHEM database at the time.

The heavy gas dispersion MEP covers verification fairly briefly, noting that it is an extremely tedious task and many developers take a less rigorous approach. The 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.

For validation, the guidance of the MEG is restated, and extended by referring extensively to the statistical methods given in Hanna *et al.* (1991) as an objective means of validating a model. No guidance is given on acceptance criteria, or what would constitute a "good" or "acceptable" model. Instead, it is 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. The section on validation also requires that a quantitative assessment is undertaken on the uncertainty in the input and output data for the model, for example, mesh and timestep sensitivity for numerical models. It also suggests that some estimate should be made for the dependent and independent variables in the data. In practice, this is not straightforward when dealing with historic or "second-hand" data which is far removed from the original experiment and the accuracy of sensors etc. may never be known.

3.7.1 Supporting appendices

The heavy gas dispersion MEP contains several supporting appendices which give additional information relevant to the scientific assessment and the validation stages. Appendix I discusses applicability aspects of dispersion models and considers restrictions on their use in the near and far fields. The purpose of this is for the evaluator to be able to check that the model can be used with confidence for the distances under consideration. In the near field, source effects may dominate and in the far field, atmospheric effects may dominate. Appendix II discusses the features of the different classes of models with the aim of providing the evaluator with checklists of important physical aspects of the models. Some guidance is provided on the coupling of source term and dispersion models where it is emphasised that sometimes it is difficult to distinguish between the release and dispersion of contaminant. In these cases, two models are effectively being evaluated in series[2].

The final two appendices cover the selection of relevant variables and appropriate statistical methods for comparing the model output with the experimental data. Important considerations when selecting relevant variables are the response time of the instrument and the averaging time used to process the data as well as the averaging time implicit in the model.

A method of processing a generic dispersion data set is described, and this is based on the methods set out by Hanna *et al.* (1991) for determining concentrations and plume widths from the raw measurement data. Some discussion of the use of wind tunnel scale data is provided, which focuses on the ability of models at smaller scales and the application of relevant scaling parameters. For the statistical analysis, reference is again made to the methods described by

[2] *A method of evaluating combinations of models is presented in a protocol by Gant (2012), for release, near-field and far-field dispersion of dense phase carbon dioxide. The protocol by Gant (2012) has not been included in this review as it is currently work in progress.*

Hanna *et al.* (1991) and some alternative methods for calculating a “goodness of fit” are suggested.

3.8 MEGGE

The MEGGE protocol (MEGGE, 1996) takes the headings from the MEP and expands each section to be relevant to the different techniques for modelling gas explosions. It also begins by defining what the protocol does and does not cover. The protocol covers the subject area of gas explosions where a flame is accelerated by the presence of many obstacles[3] and in which the boundaries of any confinement do not fail. Blast wave propagation and loading of structures are explicitly not covered, though it is noted that many of the models do provide input to these problems. A short time after the production of the MEGGE protocol, a further document was produced by an Explosion Model Evaluation Group on structural response to explosions (Worth, 1997). This project was a continuation of the MEGGE document, which aimed to provide guidance on how to evaluate a model for computing the response of a structure, following an explosion which gives rise to significant structural loading. This document, while it is aimed at structural response models, follows the content of the MEGGE MEP very closely.

The MEGGE MEP begins by defining the types of models used for gas explosions, and classes them as:

- Empirical models
- Phenomenological models
- Numerical simulators
- Experimental scaling

The first three of these broadly fit the descriptions given in Section 2.2 of this document, but the fourth, experimental scaling class is interesting. This is because the technique amounts to a geometrically equivalent reduced scale experiment where the reduction in scale is accounted for by using a more reactive gas mixture. Clearly, this technique will require a different approach to evaluation than would be applied to theoretical or computer models. MEG (1994a) specifically includes physical modelling with scaling as a possibility. Today we are more likely to validate and use a computer model.

The MEGGE MEP shares much of its wording with the heavy gas dispersion MEP, both having been derived roughly simultaneously. Therefore the comments in the previous Section are also relevant. The use of supporting appendices follows the same pattern as the heavy gas dispersion MEP, but is tailored towards explosion modelling. Here the selection of variables focuses on overpressure and comparison of rise time.

3.9 SMEDIS

The SMEDIS (Scientific Model Evaluation of Dense gas DISpersion models) project was a continuation of the MEG and HGD work and was also part sponsored by the CEC. As previous evaluation studies had dealt only with releases over flat unobstructed terrain, real release scenarios from process plants would involve more complex problems such as aerosol sources, obstacles and complex terrain. A need was therefore identified to address these scenarios (Daish *et al.*, 2000). The SMEDIS project aimed to produce a model evaluation protocol to address these “complex

[3] *This means that the protocol would also be applicable to large “open air” explosions like in the Buncefield Incident in 2005 (Buncefield Major Incident Investigation Board, 2005)*

effects.” Rather than a tool for ranking models in terms of performance, the SMEDIS project intended to encourage continual model development and leave in place a protocol and database for use by future model developers. Although the SMEDIS protocol was based on the MEG and HGD protocols, it was designed to be much more specific to the three complex effects scenarios. This was because the HGD protocol was considered to be not explicit enough in its description of the evaluation procedure. Another area where SMEDIS differs from the MEG protocols is that a significant aspect of it is concerned with development and refinement of the protocol, or versions of it, for specific uses. While the protocol is based on the structure of the MEG protocol, there are additional activities which are specific to evaluation of the protocol, as much as evaluation of the models. Therefore, it is based around a five stage process (CERC, 2000):

1. Pre-evaluation tasks
2. Carry out scientific assessment exercise
3. Carry out verification exercise
4. Carry out validation exercise
5. Post-evaluation tasks

3.9.1 Pre-evaluation tasks

Pre-evaluation tasks are defined as the setting-up activities that are required before the actual model evaluations can begin (CERC, 2000). These activities consist of: defining the models and who will be responsible for the various parts of the evaluation of each one; setting the desired parameters of the evaluation; ascertaining whether this protocol is adequate for the requirements; modifying this protocol, if necessary, to meet those requirements; and finally modifying the database of validation data sets, if necessary, to meet those requirements.

The protocol notes that, if the user is content that the protocol is already satisfactory for their requirements, then the only pre-evaluation tasks are to define the models and who will be responsible for the various parts of the evaluation of each one. For example, if a dense gas dispersion model is being evaluated, then the protocol as it stands may well be sufficient without modification.

The SMEDIS protocol requires a “model proponent” for each model, that is, someone who has access to a model and an understanding of its function. In many cases the proponent may be the model developer. The proponent may be assigned some or all of the evaluation tasks and therefore there is scope for self-evaluation, as well as evaluation by an independent third party.

3.9.2 Scientific assessment

The SMEDIS protocol makes use of the concepts of active and passive evaluation. An active approach is where some new information about a model is generated, for example during the validation exercise. A passive approach makes use only of pre-existing information, such as that contained in documentation and references. It is recognised that an active approach may not be possible in many cases, without direct interrogation of the model developer. SMEDIS therefore treats the scientific assessment passively. The principal stages are to obtain information on the model, to carry out the scientific assessment according to the protocol and to record the findings in a Model Evaluation Report or MER. The main source of information on the model is the questionnaire which is provided in an appendix and is completed by the model proponent/developer. The scientific assessment is then carried out using the returned questionnaire, along with the relevant documentation, under the following headings:

(0) Evaluation information

(1) General model description

(2) Scientific aspects

(3) User-oriented aspects

(4) Verification performed

(5) Validation performed

(6) Conclusions

3.9.3 Verification

SMEDIS treats verification passively and evidence is therefore sought during the scientific assessment. The reason for this is the labour intensive nature of verification and the practicality of carrying it out in full for each model.

3.9.4 Validation

Validation is the only part of the SMEDIS protocol that is treated actively. It involves running the model against the test cases listed in the SMEDIS database, or the experiments identified in the pre-evaluation stage and computing the statistical performance measures. Since the protocol is concerned with complex effects, it is recognised that not all models are able to take these effects into account. Therefore, it allows the user to select a subset of the data to run the model against.

3.9.5 Post-evaluation tasks

The post evaluation tasks focus on assessing the suitability of the protocol and feeding the results of the evaluation back to the model developer or proponent. This gives an opportunity for the developer to comment on how the model has been applied. Post evaluation also involves making recommendations for refinement or improvement of the protocol in light of the evaluation exercise.

3.9.6 Supporting appendices

As with the MEG MEP, there are a number of supporting appendices, including a model evaluation questionnaire, a template model evaluation report and details of the model validation parameters and how to compute them from the experimental data. This approach, as with the MEG MEP allows a standardised method to be applied to model evaluation and provides a means of comparison of results for future evaluations.

3.9.7 Model evaluation report template

The template model evaluation report provided in SMEDIS is a very detailed document that consists of a number of check-boxes in the various sections, together with spaces to allow the description of various features to be expanded. The template is tailored to assessing the capabilities of models to account for the complex effects that are the main focus of SMEDIS.

3.9.8 Limitations

The SMEDIS protocol recognises that models may be linked together, for example a jet source model may be used as input to a dispersion model. The general model description makes allowance for this with a section titled “Relationship with other models.” However, this does not consider how to evaluate the coupled system of models that may be used to make a particular prediction.

While model performance is quantified using a range of performance measures, there is no guidance on acceptance criteria or for ranking of models according to their performance.

Sensitivity and uncertainty are considered only in the scientific assessment. The SMEDIS protocol does not involve the in-depth statistical analysis of model results used by Hanna *et al.* (1991) for example. In this sense, the SMEDIS protocol places much more emphasis on the scientific assessment than the numerical analysis of results.

3.10 THE MODEL VALIDATION KIT

The Model Validation Kit (Olesen, 2005, Olesen and Chang, 2010) is a collection of four experimental data sets accompanied by software for model evaluation[4]. The kit is based on short range passive atmospheric dispersion and was originally intended to provide a practical tool that could be used as a common frame of reference by modellers. The kit can be seen as a validation tool, rather than evaluation system and therefore has much in common with the work of Hanna *et al.* (1988, 1991). For this reason, the kit was not reviewed as part of this project. However, numerous useful and relevant lessons have been learnt from its application through various workshops and these lessons are discussed further in Section 4. In the context of the Model Validation Kit, the American Society for Testing and Materials (ASTM) produce a standard D6589-05 (ASTM, 2015) “Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance.” As with the Model Validation Kit, ASTM D6589-05 is concerned with techniques for validation, rather than model evaluation as a whole. Standards for model evaluation are reviewed further in Section 3.17.

3.11 LNG MEP

In the US, the Federal Energy Regulatory Commission (FERC) is responsible for siting of liquefied natural gas (LNG) facilities and requires applicants to demonstrate compliance with a set of regulations produced by the Pipelines and Hazardous Material Administration (PHMSA). These regulations adopt portions of the NFPA 59A standard produced by National Fire Protection Agency. This standard includes a risk based approach to LNG plant siting in which modelling the dispersion of LNG vapours is an important step. The 2001 version of NFPA 59A was prescriptive in its requirement to use either DEGADIS (integral) or FEM3A (FE CFD) for vapour dispersion modelling but limitations of DEGADIS and difficulties using FEM3A led to a need for alternative models to be allowed. Furthermore, the increase in use of LNG and the number of siting applications would lead to an increase in proposals to use other models. This led to a need to evaluate other models specifically for their ability to predict LNG vapour dispersion. Three projects were subsequently commissioned:

1. A dispersion model evaluation protocol, reviewed in this Section
2. A dispersion model validation database, referenced in this Section
3. A source term model assessment protocol, reviewed in Section 3.12

The model validation database (Coldrick *et al.*, 2009) arose from recommendations in the model evaluation protocol that a database of dispersion experiments be created in a similar way to that recommended in the SMEDIS protocol.

[4] http://www.harmo.org/kit/ *(accessed 25-09-2015)*

The dispersion model evaluation protocol is set out in Ivings *et al.* (2007) and summarised in Ivings *et al.* (2013). It follows much of the structure set out in SMEDIS, but was adapted to specifically account for the physics seen in the dispersion of LNG vapours, namely:

- Formation of a (cold) dense cloud due to the low boiling temperature of LNG
- Gravity-driven spreading
- Advection by the ambient wind field
- Reduction in turbulent mixing due to the (resulting) stable density stratification
- Dispersion influenced by atmospheric stability

Other factors could also affect the dispersion of vapour. These include the presence of obstacles or vapour fences and heat addition or removal by condensation, evaporation or contact with the ground. The protocol was therefore designed around evaluating models in a way which would account for these key physics. The protocol also recognises that how a model is used is at least as important as the choice of model itself. To aid the evaluator, a number of introductory pages are included which detail the key physics of LNG dispersion and a further section gives best practice advice on how to apply models consistently.

The protocol divides LNG dispersion models into four classes:

1. Workbooks and Correlations
2. Integral models
3. Shallow layer models
4. CFD models

One of the principal aims of the protocol is that it is applicable to all classes of model and is not biased to any particular type of model. This was achieved through defining an evaluation process that would account for the features of all the model classes, and by recognising the capabilities and limitations of each model type. Although specific to LNG dispersion, the protocol adopted the stages of scientific assessment, verification and validation as set out by the MEG and SMEDIS. To aid the process, the evaluation is carried out against a set of qualitative criteria (for scientific assessment) and quantitative criteria (for validation) which a model needs to meet to be considered acceptable.

3.11.1 Scientific assessment

The LNG MEP recommends that the model scientific assessment should be carried out by an independent third party who has the necessary expertise. The concept of a model proponent has been taken from SMEDIS; that is someone who may or may not be the model developer but who has intimate knowledge of the model. The information needed to carry out the scientific assessment is obtained from the proponent using a questionnaire supplied as an appendix to the protocol and also using documentation such as user manuals, published papers and reports.

3.11.2 Verification

The approach to verification follows that set out in SMEDIS, where it is treated passively, by reviewing evidence collected during the scientific assessment. Verification is recorded in a model evaluation report (MER), but not included in the qualitative assessment criteria. The reason given for this is that the absence of information or evidence of verification would not be a sufficient reason to reject a model. Also the judgment that needs to be made on whether a model has been verified is subjective as well as being reliant on claims made by the model developer/proponent which are impractical to substantiate.

3.11.3 Validation

The validation procedure again adopts the approach set out in SMEDIS. In this approach, careful consideration is given to identifying the key physics and variables involved in LNG dispersion and selecting appropriate test cases to cover the range of target scenarios. An alternative approach to validation would be to amass a large quantity of test data and run the model against as many scenarios as possible. However, because validation is extremely time consuming such an approach would be unfeasible and would also carry the risk of not testing the model correctly. In other words, the emphasis is on matching the domain of validation with the domain of application of the model.

Once the target scenarios have been defined, Ivings *et al.* (2007) recommend the following additional steps:

- Identification of suitable validation datasets
- Selection of specific cases from these datasets so as to cover the range of target scenarios
- Definition of physical comparison parameters (PCP) that are measured or derived from measurements and which form the basis of comparisons with model predictions
- Selection of statistical performance measures (SPM) that allow a quantitative comparison of predictions against measurements
- Review and definition of quantitative assessment criteria that define the acceptable numerical range of the SPM which result from applying this validation procedure

Each of these steps is addressed in turn in the protocol, which makes a series of recommendations and lists the test cases to be included.

The specification and construction of a validation database was not part of the scope of the evaluation protocol, but was carried out as a separate project. This project produced a validation database and an associated guide document (Coldrick *et al.*, 2009) which contained detailed information about each test, the data sources and how the test data were processed into a format suitable for model evaluation.

3.11.4 Model evaluation report

The outcomes of the evaluation stages are recorded in a Model Evaluation Report (MER), for which a template is provided. The model evaluation report contains the results of the assessment of the model against the various qualitative assessment criteria and quantitative criteria for validation. A final comment must also be made on the suitability of the protocol for assessment. This stage is important as it feeds into one of the recommendations by Ivings *et al.* (2007),that the protocol may need to be reviewed and refined in light of experience.

3.12 LNG SOURCE TERM MODEL EVALUATION

One of the final recommendations of Ivings *et al.* (2007) was that there was a need for a separate evaluation of LNG source term models. This was because many dispersion calculations are based on a source term model and assessing their effectiveness was seen to be an important but complex problem. The result of this recommendation was that a model evaluation protocol was produced, specific to source term models for LNG dispersion calculations. Webber *et al.* (2009) set out a refinement of the MEG approach (MEG, 1994a), in the same way as was done for dispersion modelling. However, they note that some features of gas cloud modelling are fundamental to the way such procedures have been constructed and these features are not necessarily present in source term modelling. There are many gas dispersion models and they all do roughly the same thing – they predict concentrations at distances from the source. There are also a relatively large number of experiments against which to compare the models. Even in this case, refinement of the generic evaluation procedure to a particular class of model is far from trivial (Webber *et al.*, 2009). Source models, by their nature, vary tremendously as they need to take into account a large number of factors, for example:

- The thermodynamic state (temperature, pressure) of the liquid within the containment
- The shape, size and location of the breach
- Liquid jet, liquid spray, two phase jets
- Jet impact on a surface, the ground or water
- Jet penetration of the water surface
- Rainout
- Pool spreading, including the effects of waves and vessel movement for spills on water
- Rapid Phase Transitions (RPTs)
- Water ingress into the LNG containment
- LNG vapour escape directly from the confinement (as in the case, for example, of roll-over)

Source term assessments may need to account for some or all of these factors, but the models may be specific to only one or two aspects. Evaluation of source term models is therefore not possible in the same way as it is for gas dispersion models. Much of this due to the lack of experimental data at the right scales, which can also be heavily substance dependent. There is also some uncertainty over the applicability and scaling of laboratory data to field scale. For these reasons, Webber *et al.* (2009) adopted an approach to source term model evaluation which follows the same structure as for dispersion models, but with a greater emphasis on “best practice” and scientific assessment.

3.12.1 Verification and validation

Verification is treated as for Ivings *et al.* (2007) and it is suggested that the job of the evaluator is to determine that the verification carried out by the developer is sufficiently adequate based upon the published details.

For the reasons outlined above, validation of source term models for LNG dispersion using pre-defined datasets was not deemed possible. The approach suggested by Webber *et al.* (2009) was that modellers should use as much data as possible to gain an overall picture of model performance, and this should be quantitative where possible. They suggest that validation should be done in the first instance by the model developer and the evaluator should, as with verification, assess the adequacy of this. In effect, this would constitute "passive validation." Only when adequate data were available, could an "active validation" be undertaken by the evaluator.

The LNG source terms MEP highlights the difficulty in producing an MEP specific to a particular area when the area covers a wide range of physical scenarios. In particular, the identification of validation datasets and methods of verification have to be made very generic.

3.13 DEFRA MEP FOR AIR QUALITY MODEL EVALUATION

DEFRA (Department for Environment Food & Rural Affairs) is the UK government department responsible for policy and regulations on environmental, food and rural issues. DEFRA uses air quality models in support of policy formulation and assessment and relies on its air quality modelling contractors to provide evidence that each model is fit-for-purpose. DEFRA developed an air quality model evaluation protocol (Derwent *et al.*, 2010) for use in a regulatory environment and to advise contractors on what would be considered as "best practice" in air quality model evaluation. There were several aims behind the development of the protocol; to document the level of performance of models, to provide a judgment on the performance and to help build a long term programme of model development and improvement. The protocol focuses on models for the dispersion of three substances, namely; models for ozone transport, models for the deposition of acidic agents and models for the transport of nitrogen oxides. The overall framework of the protocol was to address the following questions:

1. Is the scientific formulation of the model broadly accepted and does it use state-of-the-art process descriptions?

2. Does the model replicate observations?

3. Is the model suitable for answering policy questions and fulfilling its designated tasks?

The protocol also makes reference to a fourth step of probabilistic evaluation (sensitivity analysis) which was not considered in the scope of the study. It is noted that probabilistic evaluation is particularly relevant where dispersion models are used in support of policy decisions.

3.13.1 Scientific evaluation

To address the first question, a scientific evaluation is carried out which aims to enable an independent reviewer to decide on the appropriateness of the model and its fitness for the intended purpose. Some information on the model is obtained via a basic information questionnaire but it is noted that this information is mainly for documentation. The scientific evaluation stage then proceeds by asking a set of questions relevant to models for each of the three substance groups. To do this, the particular characteristics of each substance have been defined and the requirements for those characteristics identified, in addition to more general requirements such as atmospheric dispersion features.

3.13.2 Operational evaluation

The operational evaluation aims to answer the second question, i.e. how well does the model replicate real-world behaviour? The evaluation protocol does not specifically refer to this section as being a validation exercise, but lists the key features of one. Instead, definitions of verification and validation are given in a previous section where it is suggested that policy makers will look to see whether or not a model has been verified or validated. The document provides a section on computing some statistical performance measures but also warns that a model can be shown to be faulty by these measures, but cannot be necessarily shown to be valid. Values of acceptance criteria are listed, and a model should be considered faulty if these are not met. The justification for these acceptance criteria are given on the basis of experience of modelling.

3.13.3 Diagnostic evaluation

The third question is answered by a diagnostic evaluation stage. Where the operational evaluation is concerned with predicting absolute values, the diagnostic evaluation is a model's ability to predict responses in air quality to changes in inputs. This step appears to aim to evaluate how well a model operates in regions outside the test data, i.e. do we have confidence in its predictive abilities? This stage may be more relevant to the passive type air pollution models covered by this protocol, where the interest is in predicting dispersion of different emission rates under different atmospheric conditions. The approach may be seen as a form of sensitivity analysis and is slightly different to one which relies on the scientific assessment to give confidence in predictive ability.

The model evaluation protocol does not list specific datasets to use to carry out the evaluation, but gives several possible data sources in an appendix with indications of data quality and uncertainty.

3.14 COST ACTION 732

COST Actions are a European networking support initiative for researchers, engineers and academics to improve cooperation and coordination of nationally funded research. In 2005, under COST Action 732, a workshop was set up as it was recognised that despite their increasing use, many microscale meteorological models had never been the subject of rigorous evaluation. The workshop aimed to bring together experts in the field of urban meteorology to review present practices, to review available data and to recommend how to improve and assure the quality of models. The results of COST Action 732 were documents outlining the background, (Britter and Schatzmann, 2007a) the guidance and protocol, (Britter and Schatzmann, 2007b) a set of example case studies, (Schatzmann *et al.*, 2010) and best practice guidance (Franke *et al.,* 2007).

COST Action 732 is aimed at microscale meteorological models, particularly for use in urban areas, where the length scales range from a few tens of metres to a few hundred metres i.e. some way between the engineering and meteorological perspectives. There is particular emphasis on CFD, but the protocol also states its applicability to other types of model such as integral, empirical and lagrangian models. The basic protocol follows the pattern set out by SMEDIS and the MEG, including the established stages of:

- Scientific Evaluation
- Verification
- The provision of appropriate and quality assured validation data sets
- A Model Validation Process in which model results are compared with the experimental data sets
- An Operational Evaluation Process that reflects the needs and responsibilities of the model user

3.14.1 Scientific assessment

Information used in the scientific assessment is obtained from a questionnaire and it is recommended that ideally, the assessment is completed independently of the model developer or user. The authors note that in practice this can be quite difficult because the developer or the users are often in the best position to provide a sound understanding of the model attributes and limitations.

3.14.2 Verification

The section on verification recognises that non-CFD and CFD codes may need to be treated differently in verification, while there are some aspects that are common to both. For non-CFD codes, some simple checks are recommended and for CFD codes, there is a description of code and solution verification. These appear to be guidelines only and formally the protocol only requests that the code developers or users provide information about the strategies that have been used to ensure satisfactory model verification.

3.14.3 Validation

A requirement for validation is that several data sets are used and there is some discussion on the merits of using a combination of both lab scale and field scale datasets. The argument for this is that lab scale data are often better controlled and have reduced experimental uncertainty, but some physical effects are only present at field scale. Such effects might be non-neutral stability, thermal effects or deposition which are not easily included in the wind tunnel. The protocol provides a number of possible test cases and gives ways in which the data may be processed and compared with the model results. A possible baseline approach to validation is suggested including model validation metrics and acceptance criteria. An option during the validation stage is to carry out a sensitivity analysis carrying out an ensemble of simulations and varying input parameters.

3.14.4 Operational evaluation

The operational user evaluation is based on information obtained from the questionnaire, based in part on the SMEDIS protocol. A number of guidance paragraphs are given on features that should be present in terms of user operation, in effect, listing good practice. The protocol suggests that the operational evaluation questionnaire should be completed by the user.

3.14.5 Summary

The COST Action 732 protocol follows very closely the framework set out in previous protocols, but is specific to small scale atmospheric dispersion modelling. The key difference between it and other protocols is the way that various best practice approaches to each stage of the evaluation are included within the protocol itself, and the actual protocol requirements are succinctly stated at the end of each section. Overall, this leads to a non-prescriptive approach which appears to be applicable to other physical phenomena.

3.15 COST ACTION ES1006

Following on from COST Action 732, a second COST Action was recently completed, ES1006 on “Evaluation, improvement and guidance for the use of local-scale emergency prediction and response tools for airborne hazards in built up environments.” Part of this Action was to produce a model evaluation protocol for such tools and models. In addition to the protocol (COST ES1006, 2015a) there are also best practice guidelines (COST ES1006, 2015b) and example case studies (COST ES1006, 2015c) following the framework of COST Action 732.

The COST ES1006 protocol follows the same process as set out in COST 732, however, specific requirements are not set out with each section. Instead, each section lists its purpose and ways in which that might be achieved.

The fundamental difference between COST ES1006 and 732 is that sensitivity and uncertainty analysis is explicitly included in the ES1006 protocol. This is specified as characterisation of model uncertainties, propagation of input parameter uncertainties and assessment of the model's response to changes in input data or to in-model parameterisations and methods of solution. The assessment and quantification of uncertainties was seen to be an important step because the protocol is oriented towards the support of responsible authorities and stakeholders in emergency decision making processes. In these conditions, the communication of model results to decision makers needs to be accompanied with an estimate of their uncertainty.

3.16 SUSANA

The Fuel Cells and Hydrogen Joint Undertaking (FCH JU) is a European initiative supporting research and development activities in fuel cell and hydrogen energy technologies. The three members of the FCH JU are the European Commission, fuel cell and hydrogen industries represented by the new Industry Grouping and the research community represented by Research Grouping N.ERGHY. A report titled "Prioritisation of Research and Development for modelling the safe production, storage, delivery and use of hydrogen" (Baraldi *et al.*, 2011) was produced for FCH JU. The report was based on the outcomes of a literature review and a workshop attended by recognised experts in the field of hydrogen safety. This gap analysis found that a model evaluation framework and associated database did not exist for hydrogen safety modelling, using CFD in particular. The "SUpport to SAfety ANalysis of Hydrogen and Fuel Cell Technologies" (SUSANA) project aimed to support stakeholders using CFD for safety engineering design and assessment of fuel cells and hydrogen (FCH) systems and infrastructure through the development of a new model evaluation protocol. The project started in September 2013, is currently ongoing and involves seven partners across Europe, including stakeholders from research organisations, universities, industry and regulators. The main elements of the project are:

- A review of the state-of-the-art in CFD, physical and numerical modelling applied to safety analysis in FCH technologies
- Developing verification and validation procedures for CFD models/codes/simulations applied to hydrogen safety
- Compiling a best practice guide in numerical simulations of problems specific to safety of FCH technologies
- Developing a CFD Model Evaluation Protocol for assessment of the capability of CFD models to accurately describe the relevant physical phenomena and the capability of CFD users to follow the correct modelling strategy
- Creating the infrastructure for implementation of the CFD Model Evaluation Protocol, which includes:
    - A database of problems for verification of codes and models against analytical solutions, designed to demonstrate capability of CFD codes to numerically solve the governing equations
    - Model Evaluation Database of experiments for validation of simulations covering a range of phenomena relevant to FCH safety

- A benchmarking exercise for codes and models, (further advanced during the project) to be continued beyond the project

- A project website to provide open access to the databases, the best practice document, benchmark exercise specifications, available benchmark results and the MEP through a public access area

- Establishing an experts' group at an early stage of the project. The experts' group aims to complement the knowledge of the project partners and provide external feedback on the development of the project, and the implementation of the Model Evaluation Protocol

- Dissemination of project results to stakeholders through different channels, including the project website, an expert workshop, a dissemination seminar, publications and conference presentations

The protocol and its associated supporting documents are based around modelling aspects of a number of physical phenomena:

- Release and mixing of gaseous hydrogen, including permeation
- Release and mixing of liquid hydrogen
- Ignition
- Fire
- Deflagration
- Detonation
- Deflagration to detonation transition (DDT)

At the time of writing, the SUSANA protocol is in a draft form and much of the supporting structure such as the state of the art review and best practice guidance is in place. The protocol is based on an adaption of the SMEDIS protocol but its main difference is the listing of specific cases for code verification for the various physical phenomena and a number of validation datasets. The creation of the validation database is a large element of the project and the collation of sufficient data to cover the different phenomena has been challenging. The database differs from the MDA of Hanna *et al*. (1991), the SMEDIS database and the LNG model validation database of Coldrick *et al.* (2009). These databases contain processed experimental values that can be used directly in model validation. The SUSANA database where possible contains a detailed description of each experiment and, in many cases, the unprocessed experimental data. One potential advantage of this is that the evaluator may choose how to process the data to provide the most appropriate comparison with the model, but it means that significantly more effort is required in the validation stage. The approach also leaves the data interpretation to the evaluator, making model intercomparisons between different evaluators more difficult because the data processing steps need to be recorded precisely in the evaluation report.

3.17 STANDARDS AND FIRE MODEL EVALUATION

Standards for verification and validation of CFD models exist, such as the ASME V&V 20-2009 standard (ASME, 2009). This focuses on providing techniques for verification, validation and treatment of sensitivity and uncertainty, rather than model evaluation as a whole. ASME V&V

20-2009 shows that elements of model evaluation are being developed in other areas as it is one of a series of standards including structural, medical devices and nuclear system behaviour[5]. What sets model evaluation apart are the scientific assessment elements and the regulatory environment.

Modelling of fires is an area where experimental data exist in abundance and there has been a vast amount of work in comparing models with experiments. However, model evaluation protocols do not exist in the same way as they do for dispersion modelling and yet model users and decision makers still require assurance that fire models are producing acceptable results in regulatory environments.

An exception is a report by Hume (1992) titled "Evaluation of Fire Models" which aimed to provide relevant authorities with information on the accuracy of fire models due to the increasing use of these models. The report describes two phases; a qualitative phase which contains elements of scientific and user oriented assessments and a quantitative phase which involves comparison of the models with various fire scenarios. In addition to these stages, the authors recommend that a sensitivity analysis should be carried out to assess the impact on the results of changing input parameters. They also stress the importance of independent review and the scrutiny of computer code and stating the limitations of the model.

Fire modelling in particular is an area where there has been activity in standards. ISO standard 16730, titled "Fire safety engineering — Procedures and requirements for verification and validation of calculation methods" (ISO 16730-1, 2015) follows a similar structure to the ASME V&V 20-2009 and details methods for verification, validation and sensitivity analysis of predictive codes. The standard is accompanied by example applications to a fire zone model (ISO/TR 16730-2:2013) and a CFD model (ISO/TR 16730-3:2013).

ASTM also produce similar standard, ASTM E1355-12 (ASTM, 2012) "Standard Guide for Evaluating the Predictive Capability of Deterministic Fire Models." As its title suggests, while this standard is consistent with ISO 16730, more emphasis is placed on evaluation and the document has much in common with an MEP. The scope of the standard sets out a four stage process:

- Defining the model and scenarios for which the evaluation is to be conducted.
- Assessing the appropriateness of the theoretical basis.
- Assessing the mathematical and numerical robustness.
- Quantifying the uncertainty and accuracy of the model predictions.

The first stage is in essence a general model description and together with the second stage can be seen as defining a scientific assessment as it requires that the model is reviewed by recognised experts who were not involved in the production of the model. The review covers the documentation, the basis of the model and approaches and assumptions used. The third stage of assessing the mathematical and numerical robustness of the model contains tasks associated with verification – analytical tests, code checking and numerical tests, consistency and stability. The final stage covers model sensitivity and validation activities, though these are listed under "model evaluation." This activity includes comparison with standard or full scale tests which have either been carried out previously or commissioned specifically. In support of this, the standard gives several methods for comparing predicted and measured values. The model evaluation is

[5] *This effort can be viewed in a wider environment of uncertainty quantification and use and application of computer models.*

completed with an evaluation report which describes how all the above stages have been carried out.

Where ASME V&V 20-2009, ISO 16730 and ASTM D6589-05 are aimed at verification, validation and sensitivity analysis, ASTM E1355-12 describes a model evaluation process. The terminology is different to that used in many of the model evaluation protocols, but the stages involved are the very similar.

3.18 PROTOCOL SUMMARY

The previous Sections have shown that many model evaluation studies have been carried out in the field of atmospheric dispersion and a large number of protocols have also been produced. In other areas, such as gas explosions, source terms and fire modelling, there have only been a few attempts at producing model evaluation protocols. For example in the Model Evaluation Group, the gas dispersion group was the most active and the adaption of the MEP to the other areas was perhaps less successful. One of the reasons for this may be that dispersion modelling is widely used in a regulatory setting and this has driven experimental programmes. There are a relatively large number of publicly available datasets which can be accessed by model evaluators. In other areas, there are comparatively few experimental datasets, perhaps in part due to the quite specific nature of those tests. An experiment for the dispersion of a dense gas can be widely applicable in model evaluation, but the release of a flashing liquid tends to be more specific to that scenario and only of interest to the development of models in that area. However, a lack of experimental data does not necessarily mean that model evaluation cannot be performed, it means that reliance is placed on scientific assessment which is perhaps given less weight than it should.

So far, this report has described the main elements in model evaluation and how those elements have been combined into protocols for a range of applications. Table 1 provides a summary of the main features of each protocol. Generally, all the protocols have a number of common main elements and other features are added depending on the application and how specific the protocol is to a particular application. One exception is the evaluation study by Hanna *et al.* (1988, 1991) which was intended to be a validation study focussing on the quantitative metrics, rather than an evaluation method. The majority of MEPs that have been developed are for gas dispersion and relatively few have been developed in other areas. That does not mean that such activities do not take place in those areas, but possibly that quality assurance of models in those areas is of a different form. Within the EU and the US, a large number of regulatory decisions are based on dispersion modelling and this may explain the prevalence of protocols in that area. However, fire modelling, particularly in the nuclear sector, has very stringent quality requirements, yet this is not reflected in the number of model evaluation protocols in that area. Certainly, benchmark studies, verification and validation have been widely employed in the nuclear sector for many years. One of the reasons for the necessity for model evaluation protocols in dispersion modelling may be that it is an area of physics which places emphasis on scientific assessment. This is often the only way to demonstrate that a model which works at one scale is applicable at another scale.

The number of features in a given protocol is a reflection of how specific the protocol is to a given area. The MEG protocol (MEG, 1994a) is a framework and is intended to be adapted for specific uses. For this reason, it contains relatively few features, listing the stages that need to be undertaken. The LNG MEP of Ivings *et al.* (2007) set out with a narrow remit and this has enabled the MEP to have many more features. These were necessary for the project to deliver a complete model evaluation package that could be applied without carrying out extra activities. Generalised model evaluation protocols require a significant amount of extra work to tailor them to a given application.

Table 1 also shows that the “standard” tasks of scientific assessment, verification and validation form only a small part of model evaluation. To be able to carry out a model evaluation in full requires numerous ancillary tasks which are discussed in the next Section.

Table 1 Model evaluation protocol summary

Protocol	*Hanna et al (1988, 1991)*	*MEG*	*HGDEP*	*MEGGE*	*SMEDIS*
Topic	Atmospheric dispersion	Consequence modelling	Dense gas dispersion	Gas explosions	Dense gas dispersion
Pre-evaluation tasks					x
Questionnaire				included as appendix but missing	x
Scientific assessment		x	x	x	x
User oriented assessment		x	x	x	x
Verification		x	x	x	x
Validation	x	x	x	x	x
Specified datasets	x		x		x
Database	x		REDIPHEM data		x
Qualititive criteria					listed as topic of interest
Quantitative criteria	x				
Responsibility for evaluation			developer or user	developer or user	third party/developer/ proponent
Supporting documentation	x		x	x	x
Best practice guidance					
Sensitivity/uncertainty	x	x	x	x	x
Report template					x
Case studies	x		x (open exercise)		x
Known applications					UDM example

Table 1 continued

Protocol	***LNG MEP***	***DEFRA MEP***	***COST 732***	***COST ES1006***	***SUSANA***	***ASTM 1355***
Topic	LNG dispersion	Atmospheric dispersion	Atmospheric dispersion	Atmospheric dispersion	CFD analysis in hydrogen safety	Fire
Pre-evaluation tasks						
Questionnaire	x	x	x	x	x	
Scientific assessment	x	x	x	x	x	x
User oriented assessment	x		x	x	x	x
Verification	x	x	x	x	x	x
Validation	x	x	x	x	x	x
Specified datasets	x	potential datasets listed	potential datasets listed	requirements listed		
Database	x				x	
Qualititive criteria	x				x	
Quantitative criteria	x	x	x	x	x	
Responsibility for evaluation	proponent	independent reviewer	independent reviewer	developer/third party	independent reviewer	developers/user/ third parties
Supporting documentation	x	x	x	x	x	
Best practice guidance	x		x	x	x	
Sensitivity/uncertainty	x		listed as optional	x	x	x
Report template	x				x	minimum requirements listed
Case studies	x	x	x	x		
Known applications	five					three

4 LESSONS LEARNT AND RECOMMENDATIONS

4.1 APPLICATIONS OF MODEL EVALUATION PROTOCOLS

The previous Section has highlighted that the number of model evaluation protocols outnumbers published records of their application. In some cases, it is clear that the originators of the protocols have included example applications or case studies as part of the development of the protocol. This is useful for a number of reasons. It enables the developers of the protocol to identify deficiencies and areas for improvement at an early stage. A demonstration application also guides the evaluator through the process – i.e. demonstrates that it is possible.

The Model Validation Kit was discussed briefly in Section 3.10. While not a model evaluation protocol, the KIT has been applied numerous times and was the subject of a number of workshops (Olesen, 1996). The theme amongst much of the work surrounding the Model Validation Kit was the establishment of a common framework for evaluation of air quality models within Europe. Olesen (1994) highlights a question posed at one of these workshops; Why is model evaluation difficult?

- The appropriate evaluation method cannot be uniquely defined.
- Input data sets are limited - they reflect only a 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 (calibration and validation).
- There are inherent uncertainties.

The above list represents a fairly general set of problems which have arisen from model evaluation in short range atmospheric dispersion. This is an area that could be considered relatively well-covered by experiments, particularly in comparison to other areas such as source term modelling for example, where even fewer experiments exist. Olesen (1996) makes the following recommendations:

- Datasets should be well organised, carefully checked, and with their pitfalls and peculiarities well documented.
- An array of various model evaluation methods and corresponding software must be developed and be freely available.
- Protocols for specific applications should be developed and their usability thoroughly tested.

Ascertaining the usability of model evaluation protocols is relatively difficult, because, aside from the case studies included with some protocols, there are relatively few published examples of applications. Those publications that do exist tend to focus on the suitability of the model being tested, rather than the protocol and many of the case study applications also focus on presenting model results e.g. Hanna *et al.* (1991). Both SMEDIS and the LNG MEP of Ivings *et al.* (2007) (which was derived from SMEDIS) protocols have a section specifically for comments on the suitability of the protocol, so that this information may be used to improve them. Following the SMEDIS project, a sample model evaluation was carried out on the DNV Unified Dispersion Model (UDM) and an evaluation report was published (CERC, 2002). This reports that the protocol was suitable for assessment, but does not provide any further information.

A task in the MEG project was to arrange an open exercise to test the protocol (Mercer, *et al*., 1998) and this is reported in Cole and Wicks (1994) which gives comments on the protocol and general issues for discussion with the following headings:

- Comments on the protocol
- Choice and presentation of data sets
- Choice of test parameters
- Choice of performance measures
- Regulatory perspectives
- Future activities

A significant concern over this exercise and future evaluations were competing requirements for independent evaluators and the lack of funding.

Following the issue of the LNG MEP of Ivings *et al.* (2007), the US National Association of State Fire Marshals (NASFM) commissioned a review of the MEP by an independent panel of experts and this resulted in further guidance on obtaining approval for alternative models. This guidance was issued by the Pipeline and Hazardous Materials Safety Administration (PHMSA) in Advisory Bulletin ADB-10-07 (PHMSA, 2010). One outcome of the review was that the requirements in ADB-10-07 were modified to include that the model evaluator undertake a sensitivity analysis in relation to the various model inputs. Sensitivity analysis was mentioned in the MEP, but the new requirements were much more explicit, namely:

a. An uncertainty analysis that accounts for model uncertainty due to uncertainty in the assumption of input parameters specified by the user.

b. An uncertainty analysis that accounts for model uncertainty due to uncertainty in the output used for evaluation.

c. An uncertainty analysis that accounts for experimental uncertainty due to uncertainty in the sensor measurement of gas concentration, where known.

d. Graphical depictions of the predicted and measured gas concentration values for each experiment with indication of the experimental and model uncertainty determined from the analyses described above.

e. Calculation of three additional SPM (see Section 4.2.9).

f. Calculation of the SPM in the MEP for each experiment and data point in addition to the average of all experiments.

The LNG MEP of Ivings *et al.* (2007) has seen several applications and these are reported in published documents. One of the reasons for this is because of the requirement to apply the protocol before a model can be used in LNG siting applications in the US. Publication of the results also allows software vendors to promote their software as having been “approved.” GexCon describe evaluation of their FLACS CFD software in Hansen *et al.* (2010) and DNVGL describe evaluation of the Phast UDM integral dispersion model in Witlox *et al.* (2013). The available reports for these applications focus on describing the models and the validation results, rather than evaluation of the protocol itself, but one outcome of the initial applications of the MEP was that any errors in the datasets used in the validation database were located at an early stage. These applications of the LNG MEP were carried out by the applicants and the results were reviewed by the authorities in the US. Two further independent applications of the LNG MEP are for the DEGADIS integral model by the Federal Energy Regulatory Commission (Kohout, 2011a) and for the Fire Dynamics Simulator (FDS) CFD software (Kohout, 2011b). These applications also tended to focus on the suitability of the models, rather than the protocol.

The author of this report was one of the developers of the model validation database and also performed an evaluation of the UK Health and Safety Executive's DRIFT dispersion model (to be published) using the LNG MEP. As a database creator, the experiences echo that of Olesen (1994), including:

- Processing the experimental data is extremely time consuming and often subjective
- The potential for errors in the data is always present and independent checks are important
- The accuracy of experimental data sources must be questioned

These points are discussed further in Section 4.2.8. As a user of the model evaluation protocol, it was found that a large proportion of the time was spent on setting up and running the model against the test cases. Where possible, automating the process helped, but care is needed when running many cases in "batch" as errors or warnings during the model runs need to be acted upon and for this reason it was preferable to run the cases individually. In terms of the overall suitability of the protocol, it was found that, even though the protocol is specific to LNG, the evaluators found that the questions remained very generic in nature.

One of the main recommendations of Ivings *et al.* (2007) was that the protocol should be revisited in light of experiences of its application. This has happened following the review by the NASFM, which resulted in additional requirements over the original version. However, one of the main limitations of this approach is that it relies on the availability of funding to enable modifications to take place, and often this is not possible. Including trial applications of the protocol as case studies within the original project is therefore recommended, as it allows this experience to be gained before a project is completed.

Fire modelling in nuclear as well as non-nuclear applications has seen a movement from a prescriptive approach towards a risk based approach. In the US, since the 1990s, it has been Nuclear Regulatory Commission (NRC) policy to use risk-informed processes in regulatory decision making where possible (NRC, 2007). As with LNG terminal siting applications, the risk-based approach places a reliance on predictive modelling and there are particular requirements for the models used. These are set out in NFPA Standard 805 (NFPA, 2015) which states that "only fire models acceptable to the Authority Having Jurisdiction (AHJ) can be used in fire modelling calculations" and "fire models shall only be applied within the limitations of the given model, and shall be verified and validated" (NRC, 2007). Following an established standard is one way to meet these requirements and for this reason, there are a number of openly published applications of ASTM E1355-12 (ASTM, 2012).

NRC (2007) documents the evaluation of five fire models using methods which they state are in accordance with ASTM E1355-12. This is a substantial undertaking and is published in several volumes each of which detail the evaluation of a particular model covering simple zonal models to three-dimensional CFD. While the authors aimed to follow the process set out in ASTM E1355-12, they note that there is a challenge in implementing the tasks due to the fact that some fire scenarios cannot be modelled, or data do not exist to validate a model for those scenarios. Two specific limitations are therefore given as a lack of modelling capability or a lack of data.

McGrattan *et al*. (2015a) describe the evaluation of FDS following the general framework of ASTM E1355-12. This document is one of four volumes which together form the technical reference manual for FDS. The documents are based in part on the methods outlined in ASTM E1355-12 because the emphasis is on model description, verification and validation, rather than the complete evaluation process described in the standard.

CFAST (Consolidated Model of Fire Growth and Smoke Transport) is a zonal model used to compute the distribution of smoke, fire gases and temperature in compartments during a fire. The basis of a zonal model is that structures are divided into compartments and compartments into layers, where the layers result from the accumulation of hot gases. Like FDS, CFAST is developed by the National Institute of Standards and Technology (NIST). Also like FDS, the technical reference guide to CFAST (Peacock *et al*., 2013, Peacock and Reneke, 2013) follows the model evaluation framework in ASTM E1355-12. These two documents set out the evaluation of CFAST in accordance with the four stages listed in Section 3.17, in effect a full evaluation of the model carried out by its developers. The evaluation focuses on the model and does not provide feedback on the Standard (there is no requirement in the Standard to do this) but it is clear that full evaluation of a relatively simple zonal model is not a trivial task.

4.2 RECOMMENDATIONS

4.2.1 What should the overall structure be?

The majority of model evaluation protocols roughly follow a similar pattern and include at least the stages of:

- Scientific assessment
- Verification
- Validation

Many of the protocols also include:

- User-oriented assessment
- Sensitivity and uncertainty 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 a model evaluation protocol should therefore be designed to detect them. Omitting any of these stages has to be based on the assumption that some attention has been paid to the other ones. Statistically based model validation studies that do not consider the other stages assume that the models being tested have previously been shown to be scientifically robust and correctly implemented.

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). Some form of user-oriented assessment is mentioned in all the model evaluation protocols, with the exception of Hanna *et al*. (1988, 1991). 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 are the documentation and help aspects providing guidance on how to operate the model. Model results may be meaningless if they are not able to be correctly interpreted, or if variables are given confusing or obscure names.

Sensitivity analysis of models is an area which has recently seen more interest and this is partly due to the increase in computing power which has enabled multiple model runs to be easily carried out. However, the need to consider sensitivity analysis was recognised very early on. For example, Gass (1977) specifies that the evaluator needs to consider it in the evaluation process. Sensitivity analysis allows an evaluator to learn a great deal about the function of a model and the process of carrying it out can highlight faults which might be hard to detect otherwise. The fact that varying an input parameter has no effect on the output may be for a genuine reason, or due to a shortcoming in the model or its implementation. Sensitivity analysis should therefore form the fifth stage in the evaluation process and is discussed further in Section 4.2.12.

Adopting this five stage process also allows some attempt at model evaluation to be carried out in the absence of experimental data, a situation identified by Webber *et al.* (2009). When model validation has to be based on limited data, the evaluator will still need to satisfy themselves of the predictive capabilities of a model. In these cases, the emphasis has to be placed on scientific assessment, verification and sensitivity analysis.

The SMEDIS protocol (CERC, 2000) includes two additional activities which are:

- Pre-evaluation tasks
- Post-evaluation tasks

These were described in Section 3.9. Pre-evaluation tasks can be seen as addressing the "why" and "how" questions, meaning that a protocol is not applied 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 suitability of the protocol. In the LNG MEP of Ivings *et al.* (2007), the post evaluation task was set out in the model evaluation report as an assessment of the suitability of the protocol (see Section 3.11.4). The post evaluation tasks are connected with testing the protocol and a recommendation on this is made in Section 4.2.13.

4.2.2 Who should carry out the evaluation?

Most of the literature on model evaluation agrees that it is something that should be carried out by a third party to provide an objective or independent review. However, it is unlikely that an independent reviewer will have the same level of knowledge or expertise as the model developer. This leads to the possibility of incorrect model application or interpretation of the results. As previously noted in Section 2.4, some activities such as verification are the responsibility of the developer as they are in a position to do this as part of the model development. The evaluator as a third party can then look for evidence it has been carried out. The approach taken by SMEDIS, where a model proponent carries out some or all of the evaluation, allows for different aspects of the evaluation to be undertaken by different parties. In two applications of the LNG MEP of Ivings *et al.* (2007), the evaluation was carried out by the model developers and the completed evaluation reports were assessed by the authorities. This situation may arise when the authorities do not have the necessary resource to undertake an independent evaluation themselves and the model evaluation report is used to provide the evidence of the suitability of the model.

In France, a working group on three-dimensional atmospheric dispersion modelling has been set up and their guidance (INERIS, 2015) on model validation adopts another approach. The guidance specifies that a user cannot rely on model validation provided by the developer and must undertake the validation test cases themselves.

4.2.3 How should the information required to carry out the evaluation be obtained?

When the evaluation is not undertaken by the developer, the information needed can be obtained by a questionnaire completed by the developer or someone who has intimate knowledge of the model. The most comprehensive example of a questionnaire is included in the SMEDIS MEP but this was possibly because of the specific focus of SMEDIS on dense gas dispersion.

4.2.4 How should scientific assessment be done?

There is no single way of carrying out a scientific assessment and the method ideally needs to be tailored to the physics in question. The MEG MEP (MEG, 1994a) lists six stages which need to be considered:

1. Model description
2. Assessment of the scientific content
3. Limits of applicability
4. Limitations and advantages of the model
5. Any special features
6. Possible improvements

These generic stages may be applied to any model and it is the second aspect which can be made specific to a particular modelling field. In SMEDIS, the assessment of scientific content was expanded to include a set of "Topics for Interest," specific to dense gas dispersion which should be given consideration. In the LNG model evaluation protocol, the physical requirements were listed as quantitative criteria that must be met for a model to be deemed suitable. ASTM E1355-12 adopts a similar approach to scientific assessment, but places emphasis on evidence of peer review of models in the open literature. However, McGrattan *et al*. (2014) suggest that publication in scientific journals is often not the best way to report validation studies, partly because the articles become nothing more than a collection of routine verification and validation exercises. What is needed for model evaluation is evidence that the scientific basis of the model is sound.

4.2.5 How should verification be done?

Verification of models was introduced in Section 2.4.1. It is clear that verification of computer models is an area of study in itself and a pragmatic approach may be needed in model evaluation. Most of the existing model evaluation protocols adopt this approach, requiring that the evaluator looks for evidence that the model has been verified, rather than carrying out any verification themselves. The heavy gas dispersion MEP (Mercer *et al.*, 1998) does note that it may be possible for the evaluator to carry out some internal consistency checks. When the evaluation is carried out by the model developer, a more in-depth verification is possible, as demonstrated by McGrattan *et al*. (2015b).

4.2.6 How should validation be done?

Validation is an area where most model evaluation protocols agree – it is the comparison of model predictions and experimental data. In some cases (e.g. ASTM E1355-12) the definition is stretched slightly to include comparison with proven benchmark solutions, provided they have been evaluated for the scenarios of interest. In others (e.g. ASME V&V 20-2009), the definition is restricted to the points where measurements are available. The areas of validation that are less well specified are choosing the datasets and determining appropriate metrics for comparison with the data which are covered in the following two Sections.

4.2.7 How should datasets be chosen?

There are two possible approaches to specifying the validation datasets. One is for the protocol to define specific experiments, the second is for the protocol to specify that appropriate datasets should be chosen, but not specify which ones. Clearly, the first approach is only feasible when the protocol is for a particular physical scenario and a more generic protocol can only take the second approach. Specifying particular datasets places responsibility on the developer of the protocol and not specifying them places the responsibility on the evaluator. The LNG MEP of Ivings *et al.* (2007) is an example where specific datasets were chosen by the authors of the protocol as it was necessary to examine particular model features and provide the MEP as a package which required no further work other than running the models through it. The downside to this approach is that it does not readily allow new data to be incorporated into the protocol. If no datasets are specified in a protocol, then the evaluator must make the decision which to use, but this allows new and emerging experimental datasets to be considered. It also provides the opportunity for “cherry picking” of datasets to suit. There is the danger that experimental databases become a repository, filling up with more and more data, but without any thought given to the relevance of the experiments or the quality of the data. Carrying out the model runs for a validation exercise is time consuming and makes the case for being selective about which experiments to use.

In fire model validation, McGrattan *et al.* (2014) provide an argument to the opposite effect, 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.

Of the protocols surveyed, only Hanna *et al.* (1991), SMEDIS and the LNG MEP specified datasets in the protocols. The remainder give guidelines on selecting datasets. The heavy gas dispersion MEP provides guidelines on specifying the content of a database and choosing datasets, including a statement specifying why the dataset is acceptable for model evaluation and any properties that limit its usefulness. One of the most important aspects is to try to choose datasets that span the range of conditions that the model is designed to cover.

4.2.8 How should data be processed?

As noted by Olesen (1994), processing the input data is far from trivial. Often, the raw measurements must be converted into a format suitable for model evaluation. 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. Because of its importance, many of the protocols reviewed included some guidance on data processing (Hanna *et al*. (1991), CERC, (2000), Britter and Schatzmann, (2007b), Ivings *et al.* (2007).

One issue raised on numerous occasions (Britter and Schatzmann, 2007b, CERC, 2000) is the treatment of zero values (or non-detects) in experimental data. In many physical phenomena, the measurements of a particular variable will be a time series. In some cases, the time series will contain zero values, but more realistically will consist of very small or even negative values which may arise from the sensor inaccuracy or drift. This often occurs around the Limit of Detection (LOD) for many sensors. If it is only the peak or time averaged values that are of interest then consideration does not need to be given to this step. However, if the time series are of interest, then appropriate consideration needs to be given to how the values below the LOD are treated as this can have an influence on the Statistical Performance Measures (SPM – See Section 4.2.9), particularly those which cannot accept zero values. The way in which experimental data are interpreted can have a large impact on comparisons and several issues to this effect were raised by Cleaver (2012) when carrying out model runs against the LNG MEP database of Coldrick *et al*. (2009). In particular, the issue of zero values arises frequently in dispersion modelling when making comparisons at the cloud edges (because centreline values tend to be well defined).

Two common methods of dealing with values below the LOD are to remove them altogether or to replace them with a fixed value. The latter option was used in SMEDIS (CERC, 2000) where concentration measurements of zero were replaced by a value of 10^{-3} units. The authors acknowledged it would be preferable to base the thresholds on sensor sensitivity but such information was not always available. Helsel (2005) argues that both removal and substitution are overly simplistic. If the measured and predicted values below the LOD are both set equal to the LOD, or some fraction of it, datasets with many low concentrations would make a model appear to perform overly well, because the measurements and predictions would be identical in many cases.

An alternative method is to use Maximum Likelihood Estimation (MLE). This involves fitting a probability distribution to the measurements/predictions (excluding those below some low threshold value). This distribution is then used to replace values below the low threshold value with values obtained by sampling the fitted distribution, i.e. extrapolating the distribution to lower levels. An important consideration in using this approach is to ensure that the MLE distribution fits the measurements well. The method becomes more prone to error as the proportion of very low values increases, since less data is available to fit the distribution. Although the MLE method has advantages, it does not completely resolve the problem of zero or low measured values. It can still result in the generation of small observed and predicted values, which may cause problems in computing SPM. In any case, it is worthwhile identifying which data points represent “genuine” data points and those which have been generated by substitution or the MLE method. The use of data quality indicators is recommended by Olesen (1996) as a measure of the reliability of data points.

The above methods of dealing with zero values may apply at sensor locations when nothing (or a small value) is measured, but something is predicted. An equal and opposite situation occurs when something is measured, but nothing is predicted. Some statistical techniques (Section 4.2.9) cannot account for zero values and the temptation is to discard instances when nothing is predicted. However, on aggregate, this will result in a model that appears to overpredict the quantity as some instances of underprediction have not been considered.

If model comparisons are not to be made on peak values or time series, then consideration needs to be given to the technique used to average the data. The effect of averaging technique is particularly important where a model predicts a steady-state value which needs to be compared with inherently time varying experimental data such as atmospheric dispersion of dense gases. For atmospheric dispersion data, the aim of time averaging is to capture a relevant snapshot of a cloud over a particular duration or “averaging window”. Setting appropriate averaging windows can be highly subjective and usually one of two approaches is used. The first is a mechanistic one of the type used by Carissimo *et al.* (2001) in which the arrival and departure of a cloud is set by determining when 10% and 90% of the total dose (the concentration-time product) respectively are reached. The second approach is to visually assess the concentration-time series to determine the periods over which the data should be time averaged. This technique was used by Coldrick *et al.* (2009) for processing certain time varying atmospheric dispersion datasets into steady state values. For other scenarios, similar techniques may be adopted.

When processing experimental data, the decision must be taken whether to do this manually or automatically. The advantage of manual data processing is that the evaluator must look at the data and therefore gain insight into it. Manual processing may also be appropriate for small datasets where the time taken to write a processing algorithm may not be justifiable. On the other hand, adopting an automatic mechanistic approach introduces more rigour and repeatability and lends itself to large datasets where the time spent writing an algorithm is small in comparison to the time spent processing data.

Whatever procedure is used to analyse and process the data, it is important that it is transparently documented so that the process may be repeated and it is clear exactly what steps have been undertaken. A further recommendation is for protocols to include an appendix on data processing, or provide references to established methods.

4.2.9 How should models be compared with experiments?

In parallel with the data processing, the methods used to compare the experiments and model must be decided as it is important to consider validation metrics when processing the data and vice-versa. 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. Such analysis “by eye,” while essential and informative, may become subjective or introduce variability.

For a more rigorous evaluation, a procedural quantitative approach can be adopted. Statistical performance measures 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. A number of different SPM have been suggested among the protocols reviewed and an overview is given by Duijm *et al.* (1996) who suggest the following requirements for a set of SPM:

- They should give an indication of the model’s ability to predict on average, i.e. whether it under- or over-predicts.

- They should give an indication of the level of scatter i.e. the deviation from the average.

- Equal weight should be given to all measurements/predictions regardless of their absolute values.

Chang and Hanna (2004) suggest 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 SPM 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 with values of SPM that constitute a model that is performing well (Section 4.2.10). Commonly used SPM in model evaluation are the Mean Relative Bias (MRB), Mean Relative Square Error (MRSE), Geometric Mean (MG), Geometric Variance (VG) and Factor of n (*FACn*).

MRB is based upon the difference between measured (C_o) and predicted (C_p) values, but to meet the requirement for equal weight given to all measure/predicted pairs, the values are normalised by the average of the two:

$$MRB = \left\langle 2\frac{\left(C_p - C_o\right)}{\left(C_p + C_o\right)} \right\rangle \tag{1}$$

Where the angle brackets $\langle \ldots \rangle$ denote an average over all the measured/predicted pairs. MRB gives an indication of a model's ability to predict the measured values on average, and its sign indicates whether the model is under- or over-predicting. A perfect model would result in an MRB of 0, but under- and over-predictions cancel each other out and a model may appear to perform well for the wrong reason. Therefore, MRB is paired with MRSE which sums the squares of the errors and therefore gives an indication of the scatter in the predictions:

$$MRSE = \left\langle 4\frac{\left(C_p - C_o\right)^2}{\left(C_p + C_o\right)^2} \right\rangle \tag{2}$$

MG and VG similarly follow MRB and MRSE but are based on the logarithms of the ratio of the measurements and predictions (and are therefore multiplicative). This means that equal weight is given to all the pairs and the logarithm also acts to draw in outliers so that the SPM are not dominated by a few extreme values, which can occur in atmospheric dispersion modelling. MG and VG are defined as follows:

$$MG = exp\left\langle ln\left(\frac{C_o}{C_p}\right) \right\rangle \tag{3}$$

$$VG = exp\left\langle \left[ln\left(\frac{C_o}{C_p}\right)\right]^2 \right\rangle \tag{4}$$

A perfect model would result in MG and VG equal to 1. A final SPM based upon the ratio of measured to predicted values is *FACn*, which is:

$$FACn = fraction\ of\ predictions\ where\ 1/n \leq \frac{C_P}{C_O} \leq n \tag{5}$$

FACn is easily visualised and *n* is often 2.

The additive measures, MRB, MRSE and *FACn* are robust in that they can accept zero values for measurements or predictions, whereas MG and VG cannot. This may cause problems in dispersion simulations in which a model may not predict any concentration at a particular sensor location. As discussed in the previous Section, threshold values are sometimes used to avoid this problem, but can give erroneous results. The geometric measures, while they cannot accept zero values, are useful in that they can accommodate large ranges.

The above SPM have been used in many of the protocols, for example the MEG MEP (Mercer *et al.*, 1998), SMEDIS (CERC, 2000), LNG MEP (Ivings *et al.*, 2007), COST 732 (Britter and Schatzmann, 2007b), COST ES 1006 (COST ES1006, 2015a), mainly in applications for dispersion modelling. Other SPM have also been used such as index of agreement, figure of merit and correlation coefficients. However McGrattan *et al.* (2014) argue that sophisticated comparison metrics may be more trouble than they are worth and cite a study in which several metrics were abandoned in favour of a simple one because the choice of metric did not significantly affect the results. In many consequence modelling applications, the comparisons are based on peak, averaged or steady-state values and simple comparison metrics may be suitable. For some applications, for example fire or explosion modelling, the comparison may need to be made on conditions other than peak or steady state and a different approach may be required. ASTM E1355-12 (ASTM, 2012) provides guidance on comparing model and experiment in these cases, using vector norms.

SPM can be computed for individual experiments and the results reported on a per-experiment basis. When there are a large number of experiments, presentation of results on a per-experiment basis may not be very informative. A single score over a number of experiments may help to answer the question of how good the model is *on aggregate*. However, careful consideration needs to be given to how the average score is obtained so that individual results do not cancel each other out and result in what appears to be good model performance. An alternative is the approach adopted by Ivings *et al.* (2007) where the model performance over groups of experiments is assessed. These groups may be defined according to a particular physical aspect of the experiments, for example the presence of obstacles or whether the experiment was undertaken in a wind tunnel. The advantage of this approach was that a distinction could be made between models able to account for these effects, but that models would not be penalised if they could not. The evaluation report would then state the limitation of the model in not being able to account for a particular effect.

4.2.10 What is an acceptable model?

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 of 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.

While the SMEDIS protocol did not define any acceptance criteria, the associated 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, Hanna *et al.* (2004) evaluated the performance of the FLACS model in terms of a “good” or “acceptable” model. Ivings *et al.* (2007) used the results of these studies to suggest model acceptance criteria for LNG dispersion models. The MEG (Mercer *et al.*, 1998) do not provide any acceptance criteria, but suggest 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) were based on atmospheric dispersion and their values reflect its stochastic nature. There is the danger that the same acceptance criteria are adopted for other scenarios which do not have the same level of inherent uncertainty. For example, FERC (2013) use these acceptance criteria for solid flame models of pool fires, yet the physical processes are entirely different and may mean that completely different acceptance criteria should be used. Another 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 would be appropriate to adopt a much 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 an internal explosion in a sealed vessel, the maximum overpressure may be straightforward to predict but the arrival time less certain.

4.2.11 What should a user-oriented assessment consider?

A comprehensive user-oriented assessment is set out in SMEDIS, consisting of:

- User-oriented documentation and help
- Installation procedures
- Description of the user interface
- Internal databases
- Guidance in selecting model options
- Assistance in the inputting of data
- Checks on use of model beyond its scope
- Computational aspects (e.g. time taken for the model to run)
- Clarity and flexibility of output results
- Possible improvements
- Planned user-oriented developments

This list follows that given in the MEG protocol and the heavy gas dispersion protocol where it is also seen as addressing “fitness-for-purpose” and “ease-of-use.” Similar activities are also defined in the COST Action 732 protocol (Britter and Schatzmann, 2007b) under the heading of “operational user evaluation.”

4.2.12 How should sensitivity and uncertainty be included?

The importance of sensitivity analysis in model evaluation is evident because it is required, or referred to, in almost all of the protocols. Uncertainty analysis is a related but different concept which is often used in conjunction with sensitivity analysis because it is closely related. Whereas sensitivity analysis can be seen as examining the effect on model results due to the variations in input parameters, uncertainty analysis is examining the range of model outcomes for a given set of inputs. An example of this is in a study by Gant *et al.* (2013) in which dispersion calculations of dense phase carbon dioxide were performed using the DNV Phast model. The sensitivity of the model to various input parameters was tested and it was found that varying the wind speed between 0.5 m/s and 50 m/s had little effect on the dispersion distances. This was because the releases were dominated by the momentum of the jet at the concentrations of interest. Other sources of uncertainty are the assumptions about physics made in the model and the numerical solution used for the model. The sensitivity analysis may therefore be viewed as an initial test to find which model input parameters are important and their effect on output. The uncertainty analysis is a further step which quantifies variations in the input parameters and the effect on the output. SMEDIS (CERC, 2000) suggests that the evaluator should consider a number of sources of uncertainty, namely:

- Uncertainty from modelling of stochastic processes
- Uncertainty caused by model physics assumptions
- Uncertainty associated with the numerical method
- Uncertainty from errors in the input data, including identifying parameters to which the predictions are most sensitive

These (and other) sources of uncertainty can be classified into two types; aleatory uncertainty and epistemic uncertainty (Oberkampf and Roy 2010). Aleatory uncertainty is that which arises due to stochastic processes or the inherent randomness in processes such as weather or turbulence. Epistemic uncertainty arises from a lack of knowledge and is usually associated with modelling issues, or a lack of knowledge about the system of interest or its simulation. The assessment of uncertainty in the model has some overlap with verification, because the modeller may need to examine the effect of numerical parameters which alter the solution. COST 732 (Britter and Schatzmann, 2007b) provides guidance on the evaluation of CFD codes, stating that when performing validation simulations it is necessary to quantify and reduce the different errors and uncertainties originating from the following sources:

- Errors and uncertainties in modelling the physics
- Numerical errors and uncertainties

Most of the guidance given in the model evaluation protocols on uncertainty analysis tends to be generic as the exact form of the analysis depends on many factors and therefore it is impractical to go into sufficient detail on a case by case basis. Model evaluation is partly about the evaluator satisfying themselves that the model results replicate reality to a sufficient degree and having an estimate of the uncertainty of the results is an essential part of this process.

Numerous techniques are available for carrying out sensitivity analysis and these generally fall into two types; local sensitivity analysis and global sensitivity analysis. Local sensitivity analysis considers changes in the model output, for small changes in one of the inputs about some central point. Global methods consider variations in the model output for variations over the entire range of inputs. To do this, the inputs are assigned probability distributions and the model is run a number of times, sampling from the input distributions. Global methods tend to be more computationally expensive but yield significantly more information about the behaviour of the model than local methods.

The MEG protocol (MEG, 1994a) requires the evaluator to consider uncertainty in data/model inputs and this was expanded in the heavy gas dispersion MEP (Mercer *et al.*, 1998) to include a requirement for quantitative assessment of the uncertainty in the input and output data. The LNG MEP (Ivings *et al.*, (2007) follows this guidance, requiring users to consider the sensitivity of the model output to various factors such as the ground roughness length. ASTM E1355-12 (ASTM, 2012) is explicit in requiring a quantification of the uncertainty and accuracy of the model and mentions several methods of determining model sensitivity, including local and global methods.

Ultimately, it would be desirable for the evaluation to include some form of error bars on the model performance measures and this would need to be based on a rigorous sensitivity and uncertainty analysis. This could be similar to that suggested as a modification to the LNG MEP of Ivings *et al*. (2007) following review of that protocol. However, such analysis is not always straightforward, or even possible. Often models are validated against historic data for which detailed information on the accuracy of sensors, etc. is not available. In these cases, a more simplistic approach to sensitivity and uncertainty analysis may be required. Volume 2 of NRC (2007) suggests representative uncertainties based on analysis of a number of experiments.

4.2.13 How should the protocol be tested?

Trial application of a protocol is an important part of its development, and many of the protocols reviewed included example applications. This helps to identify shortcomings at an early stage and can be planned into the original task of creating a protocol. Reviewing a protocol following a "real world" application is also a desirable activity, because it allows improvements to be made as well as adjustments to quantitative criteria for model acceptance. This latter aspect would be extremely useful in areas for which there is little experience for what constitutes an acceptable model. Such experience can also be gained by organised benchmark activities, providing the emphasis of those activities is on model evaluation, rather than purely an exercise in modelling. One of the main barriers to testing of a model evaluation protocol outside of the original project is the funding of any additional activities.

4.2.14 What are the overall advantages and disadvantages of model evaluation protocols?

One of eight lessons learned, cited in McGratton *et al.* (2014) is that "model validation is not a blank check." In other words, the end user should satisfy themselves that the intended region of application falls within the parameter space of the validation. The same may be true of the evaluation process; evaluating a model for a particular purpose does not guarantee its applicability for every conceivable situation. The evaluation process should make it clear exactly what model has been evaluated, what the evaluation is for and the scenarios to which this applies.

Model evaluation protocols that are very generic have the advantage that they can be made universally applicable to different physical phenomena and different types of models. At this extreme, the task of assembling the protocol is simpler as many of the details do not need to be considered, such as assembling a validation database and its associated data processing. The MEG protocol is an example of a very generic protocol – it requires adaption to be useable. At the other extreme are very specific evaluation protocols that can be used "out of the box," the LNG MEP of Ivings *et al.* (2007) being an example. This protocol evaluates a model for a very specific range of scenarios, but means that the task of applying the protocol is simpler.

Recommendations for best practice are made in a number of the model evaluation protocols. Following best practice is a step in the model evaluation process and, especially for more complex models, there is a strong user dependence. Model evaluation can therefore be seen as part of ensuring model quality and consideration should be given to whether best practice guidance should be included.

5 CONCLUSIONS

The aims of this report were to review existing model evaluation protocols and to make recommendations for the structure and content of a new evaluation method, based on experiences gained from previous evaluations. Model evaluation has been in existence since the early use of computer simulations and techniques have been developed which could be applied to a wide range of fields. Much of the early work on model evaluation focussed on models where the outputs are used in support of policy decisions of some kind and a decision maker needs to be assured that the model output is a scientifically robust and accurate description of the actual process.

Atmospheric dispersion is an area where there has been significant activity in model evaluation. The main drivers in this area were the need to assess the risks from the loss of containment of hazardous substances and the introduction of air quality laws which led to a requirement to model air pollution. In both cases, the underlying need was to communicate reliable and robust results to independent decision makers who may be independent of and far removed from the modelling process. Numerous model evaluation protocols and model comparison exercises appeared as a result.

Fewer model evaluation studies exist in other areas of consequence modelling, such as fire, explosion and source term models. That does not mean that such activities do not take place, but that they take a different form as quality assurance of simulations is important in applications such as the nuclear industry. For fire modelling, standards and benchmark studies are more prevalent than model evaluation protocols. Computational modelling of explosions is also less well established than dispersion modelling and many of the techniques are still seen as being in development rather than in use for routine consequence assessment calculations.

On the whole, most of the model evaluation protocols follow a fixed format, which evolved early on, namely:

- Assess the scientific basis of the model and whether it represents the process being modelled (scientific assessment),

- Assess whether the computer implementation of the model matches its specification (verification),

- Compare the model outputs with experimental test data to determine whether the modelled process matches the actual one (validation).

This three stage approach arises naturally from the need of a decision maker to have confidence in model predictions. Often, some model evaluation must be done during the development stage, but falls under quality assurance.

In addition to the above three stages, most model evaluation protocols consider two further activities:

- A user oriented assessment (ease-of-use, or fitness-for-purpose),

- Sensitivity and uncertainty quantification.

User oriented assessment considers aspects such as model input/output, documentation and user-interface. It forms an important element in model evaluation because model users or evaluators are often different to the model developers and the usability of the model becomes critical to obtaining reliable predictions.

Sensitivity and uncertainty quantification has been linked with model evaluation since the early examples, but is becoming more explicitly required in more recent examples as techniques and computing power have developed. Sensitivity and uncertainty quantification allows evaluators to learn about a model and identify possible shortcomings. It also helps in the communication of model results to decision makers, to include an assessment of the uncertainty of the results. In developing a protocol, it may be beneficial to include pre- and post-evaluation tasks. These help to ensure that consideration is given to how a protocol is applied and how it is evaluated following the application.

One of the main findings of the review was that model evaluation protocols fall into two categories; they are either very generic and can be applied to any consequence modelling area, or they are very specific and have a particular area of application. Those that are very generic tend to need a high level of effort on the part of the evaluator in tailoring them to their application. Those that are very specific require less effort by the evaluator, but have a fairly narrow area of application.

Previous applications of model evaluation protocols were reviewed, to identify opportunities for "lessons learned" and the main finding was that the record of published applications is relatively small. One of the most frequently applied protocols is that for the dispersion of LNG vapour because its use is a regulatory requirement for any model to be approved for LNG siting applications in the US. This has led to a number of applications of the protocol. Application of this protocol is also relatively straightforward (though still a significant undertaking), because it is quite specific and the time consuming activity of data processing and assembling a validation database is already complete. Having their model "approved" is also beneficial for software vendors. Fire modelling in nuclear applications is also an area where the necessity for model evaluation is driven by regulations. This has resulted in several applications of the ASTM E1355-12 standard for model evaluation. Unlike the LNG dispersion MEP, ASTM E1355-12 is a fairly generic document and its application requires a significant amount of work. In other areas, the regulatory motivation for model evaluation does not exist and many studies report the results of validation. Showing that the model "fits the data" often bears more weight than showing it has been evaluated.

Based on the surveyed protocols, examples and experiences of previous applications, this report has made a number of recommendations about the structure of a future protocol and ways in which the stages of the evaluation process may be carried out.

The SAPHEDRA project aims to define a model evaluation protocol for consequence models used in emerging risk areas. Each of these areas will involve specific models which will require appropriate validation data, physical comparison parameters and evaluation criteria. However, the overall structure of the protocol may be common across all of these areas.

6 REFERENCES

ASME, (2009), ASME V&V 20-9009, Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer, The American Society of Mechanical Engineers, ISBN 978-0-7918-3209-7.

ASTM, (2012), ASTM E1355-12 Standard guide for evaluating the predictive capability of deterministic fire models, ASTM International, West Conshohocken, PA.

ASTM, (2015), ASTM D6589-05 Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance, ASTM International, West Conshohocken, PA.

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.

Baraldi, D., Papanikolaou, E., Heitsch, M., Moretto, P., Cant, R.S., Roekaerts, D., Dorofeev, S., Kotchourko, A., Middha, P., Tchouvelev, A.V., Ledin, S., Wen, J., Venetsanos, A. and Molkov, V.V., (2011), Prioritisation of Research and Development for modelling the safe production, storage, delivery and use of hydrogen, European Commission report EUR 24975.

Britter, R. and Schatzmann, M., (2007a), Background and justification document to support the model evaluation guidance and protocol, COST Action 732, Quality assurance and improvement of micro-scale meteorological models, Meteorological Institute, University of Hamburg.

Britter, R. and Schatzmann, M., (2007b), Model evaluation guidance and protocol document, COST Action 732, Quality assurance and improvement of micro-scale meteorological models, Meteorological Institute, University of Hamburg.

BS ISO 16730-1 (2014), Fire safety engineering - Procedures and Requirements for verification and validation of calculation methods - Part 1: General, draft for public comment.

Buncefield Major Incident Investigation Board, (2005), The Buncefield Incident 11 December 2005, The final report of the Major Incident Investigation Board, Volume 1, available from: http://www.hse.gov.uk/comah/buncefield/miib-final-volume1.pdf (accessed 25-09-2015)

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

Cambridge Environmental Research Consultants Ltd (CERC), (2002), Model evaluation report on UDM Version 6.0, Ref. No. SMEDIS/00/9/E.

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.

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.

Cleaver, P., (2012), Use of Model Evaluation Protocol, UKELG 49th meeting, United Kingdom Explosion Liaison Group, Use and validity of dispersion and explosion models, IGEM house, Kegworth, Derbyshire, UK.

Coldrick, S., Lea, C. J. and Ivings, M. J., (2009), Validation database for evaluating vapor dispersion models for safety analysis of LNG facilities: Guide to the LNG model validation database, The Fire Protection Research Foundation.

Cole, S. T. and Wicks, P. J., (1994), Model evaluation group seminar, the evaluation of models of heavy gas dispersion, Rauwse Meren (Mol) Belgium, European Commission report EUR 16146 EN.

Contini, S., Amendola, A. and Ziomas, I., (1991), Benchmark exercise on major hazard analysis volume 1 description of the project; discussion of the results and conclusions, Commission of the European Communities report EUR 13386 EN.

COST ES1006, (2015a), Model evaluation protocol, Evaluation, improvement and guidance for the use of local-scale emergency prediction and response tools for airborne hazards in built environments.

COST ES1006, (2015b), 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.

COST ES1006, (2015c), Model evaluation case studies: approach and results, Evaluation, improvement and guidance for the use of local-scale emergency prediction and response tools for airborne hazards in built environments.

Daish, N. C., Britter, R. E., Linden, P. F., Jagger, S. F. and Carissimo, B., (2000), SMEDIS: scientific model evaluation of dense gas dispersion models, International journal of environment and pollution, Vol. 14, No. 1 – 6, pp 39 – 51.

Derwent, D., Fraser, A., Abbott, J., Jenkin, M., Willis, P. and Murrells, T., (2010), Evaluating the performance of air quality models, UK Department for Environment, Food and Rural Affairs, available from: http://ukair.defra.gov.uk/assets/documents/reports/cat05/1006241607_100608_MIP_Final_Version.pdf (accessed 25-09-2015).

Dickerson, M. H. and Ermak, D. L., (1988), The evaluation of emergency response trace gas and dense gas dispersion models, Lawrence Livermore National Laboratory report UCRL-99348.

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.

Duijm, N. J., Ott, S. and Nielsen, M., (1996), An evaluation of validation procedures and test parameters for dense gas dispersion models, Journal of Loss Prevention in the Process Industries, Vol. 9, no. 5, pp. 323–338.

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.

Federal Energy Regulatory Commission (FERC), (2013), Recommended parameters for solid flame models for land based liquefied natural gas spills, Office of energy projects, Docket No. AD13-4-000.

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., (2012), Framework for validation of pipeline release and dispersion models for the COOLTRANS project, Third International Forum on the Transportation of CO2 by Pipeline, 20-21 June, Newcastle, UK

Gant, S. E., Kelsey, A., McNally, K., Witlox, H. W. M. and Bilio, M., (2013), Journal of Loss Prevention in the Process Industries, J. Loss Prev. Proc. Ind., 26, pp 792-802.

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.

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., Hansen, O. R. and Dharmavaram, S., (2004), FLACS CFD air quality model performance evaluation with Lit Fox, MUST, Praire Grass and EMU observations, Atmospheric Environment, Vol. 38, pp 4675 – 4687.

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.

Hansen, O. R., Ichard, M. and Davis, S. G., (2010), Validation of FLACS against experimental data sets from the model evaluation database for LNG vapor dispersion, Journal of Loss Prevention in the Process Industries 23, pp 857-877.

Havens, J., (1992), An evaluation of the DEGADIS dense gas (atmospheric) dispersion model with recommendations for a model evaluation protocol, US South Coast Air Quality Management District.

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

Hume, B. T., (1992), Evaluation of fire models, summary report, Central Fire Brigades Advisory Council, Scottish Central Fire Brigades Advisory Council, Joint Committee on Fire Research, research report 52, ISBN 0-86252-744-9.

L'Institut National de l'Environnement Industriel et des Risques (INERIS), (2015), Guide de Bonnes Pratiques pour la réalisation de modélisations 3D pour des scénarios de dispersion atmosphérique en situation accidentelle, Rapport de synthèse des travaux du Groupe de Travail National, INERIS report DRA-15-148997-06852A, available from http://www.ineris.fr/aida/liste_documents/1/86007/0, (accessed 25-09-2015).

ISO 16730-1:2015, Fire safety engineering - Procedures and requirements for verification and validation of calculation methods - Part 1: General.

ISO/TR 16730-2:2013, Fire safety engineering - Assessment, verification and validation of calculation methods Part 2: Example of a fire zone model.

ISO/TR 16730-3:2013, Fire safety engineering - Assessment, verification and validation of calculation methods Part 3: Example of a CFD model.

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., Lea, C. J., Webber, D. M., Jagger, S. F., and Coldrick, S., (2013), A protocol for the evaluation of LNG vapour dispersion models, Journal of Loss Prevention in the Process Industries, Vol. 26, no. 1, pp. 153–163.

Kohout, A. J., (2011a), Evaluation of DEGADIS 2.1 using advisory bulletin ADB-10-07, Federal Energy Regulatory Commission, Washington, DC 20426.

Kohout, A. J., (2011b), Evaluation of fire dynamics simulator for liquefied natural gas vapor dispersion hazards, Master of Science thesis, Graduate School of the University of Maryland.

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., (2015a), Fire Dynamics Simulator Technical Reference Guide Volume 1: Mathematical Model, NIST Special Publication 1018-1 Sixth Edition, available from http://firemodels.github.io/fds-smv/manuals.html (accessed 25-11-2015).

McGrattan, K., McDermott, R., Weinschenk, C., Hostikka, S., Floyd, J. and Overholt, K., (2015b), Fire Dynamics Simulator Technical Reference Guide Volume 2: Verification, NIST Special Publication 1018-2 Sixth Edition. available from http://firemodels.github.io/fds-smv/manuals.html (accessed 25-11-2015).

McQuaid, J., (1979), Dispersion of heavier-than-air gases in the atmosphere: review of research and progress report on HSE activities, UK Health and Safety Laboratory technical paper 8, ISBN 0-7176-0029-7.

MEGGE, (1996), Gas explosion model evaluation protocol, The Steel Construction Institute, Silwood Park, Ascot, Berkshire, UK.

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

Model Evaluation Group (MEG), (1994b), Guidelines for model developers, European Communities, Directorate-General XII, Science Research and Development.

Naylor, T. H. and Finger, J. M., (1967), Verification of computer simulation models, Management Science, Vol. 14, No. 2, pp B-92-B101.

NFPA 59A:2001 Standard for the production, storage, and handling of liquefied natural gas (LNG).

NFPA 805:2015 Performance-based standard for fire protection for light water reactor electric generating plants.

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.

Nuclear Regulatory Commission (NRC), (2007), Verification and Validation of Selected Fire Models for Nuclear Power Plant Applications, Volume 1: Main Report, U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research (RES), Rockville, MD, 2007, and Electric Power Research Institute (EPRI), Palo Alto, CA, NUREG-1824 and EPRI 10 11999.

Oberkampf, W. L. and Roy, C. J., (2010), Verification and validation in scientific computing, Cambridge University Press, ISBN 978-0-521-11360-1.

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.

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.

Olesen, H. R., (2005), User's guide to the model validation kit, Research Notes from NERI No. 226, National Environmental Research Institute, Ministry of the Environment, Denmark.

Olesen, H. R. and Chang, J. C., (2010), Consolidating tools for model evaluation, Proceedings of the 10th international conference on harmonisation within atmospheric dispersion modelling for regulatory purposes.

Peacock, R. D. and Reneke, P. A., (2013), CFAST – Consolidated Model of Fire Growth and Smoke Transport (Version 6) Software Development and Model Evaluation Guide, NIST Special Publication 1086r1 December 2012 Revision.

Peacock, R. D., Forney, G. P. and Reneke, P. A., (2013) CFAST – Consolidated Model of Fire Growth and Smoke Transport (Version 6) Technical Reference Guide, NIST Special Publication 1026r1 October 2011 Revision.

Petersen, K. E., (1999), The EU model evaluation group, Journal of Hazardous Materials, 65 pp 37–41.

Pipeline and Hazardous Materials Safety Administration, U.S. Department of Transportation (PHMSA), (2010), Advisory Bulletin ADB-10-07 Liquefied Natural Gas Facilities: Obtaining Approval of Alternative Vapor-Gas Dispersion Models, Federal Register Vol. 75, No. 168, pp 53371-53374.

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

Schatzmann, M., Olesen, H. R. and Franke, J., (2010), COST 732 model evaluation case studies: approach and results, COST Action 732, Quality assurance and improvement of micro-scale meteorological models, Meteorological Institute, University of Hamburg.

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.

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 25-09-2015).

Witlox, H. W. M., Harper, M. and Pitblado, R., (2013), Validation of PHAST dispersion model as required for USA LNG siting applications, Chemical Engineering Transactions, Vol. 31, ISBN 978-88-95608-22-8.

Worth, B., (1997), Structural response model evaluation protocol, European Commission report EUR 17682 EN.

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.