INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE CENTRATOM TRIESTE

UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION
INTERNATIONAL CENTRE FOR SCIENCE AND HIGH TECHNOLOGY
c/o INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS 34100 TRIESTE (ITALY) VIA GRIGNANO, 9 (ADRIATICO PALACE) P.O. BOX 586 TELEPHONE 040-224572 TELEFAX 040-224575 TELEX 460449 APH I

SMR/760-51

"College on Atmospheric Boundary Layer
and Air Pollution Modelling"
16 May - 3 June 1994

"Model Evaluation"

R. BRITTER
Department of Engineering
University of Cambridge
Cambridge, United Kingdom

Please note: These notes are intended for internal distribution only.

MODEL EVALUATION

Commission of the European Communities
Directorate - General XII
Science, Research and Development

THE EVALUATION OF TECHNICAL MODELS USED FOR MAJOR-ACCIDENT HAZARD INSTALLATIONS

by Dr. R.E. Britter

Environment

Major Industrial Hazards

Research Projects
Reference EUR 14774EN

CONTENTS

PREFACE

In order to calculate the possible damage to people, property on the environment from industrial accidents, analysts resort to mathmatic models of the accident. These models are often processed using a computer. Increasing numbers of models are emerging from research leaving the analyst with a large choice. There is a considerable possibility of models being simply wrong or of them being used outside their range of applicability.

As a regulator and also a funder of research in Major Industrial Hazards the CEC became concerned about the need to evaluate existing models and to improve the quality of what was being developed. Therefore in 1991 it commissioned Dr. Rex Britter to study the situation. The present report is the outcome of his work.

Dr. Britter recommended that the CEC establish expert groups to examine the models in detail. These activities have now been initiated. The groups are essentially open and interested scientists are invited to participate. Further information can be obtained from me.

Dr. S.T. COLE
CEC Brussels
DG XII-D-1
200, rue de la Loi
B - 1049 Brussels
Phone : +32 2 295 03 47
Fax : +32 2 296 30 24

Brussels, January 1993

EXECUTIVE SUMMARY

Models are used in many ways by individuals and organisations within the European Community. Within the context of major technological hazards decisions of considerable importance are based wholly, or partly, on these models.

The quality, that is the fitness-for-purpose, of these models will differ and there is obviously an interest in both the relative quality of one model against another and the absolute quality of any particular model.

This leads to the evaluation of model quality and to methods by which model quality might be improved.

It is noted that models are a principal method for the transfer of technology but also that the nature, role and limitations of models is frequently misunderstood.

There is a widespread demand for guidance on model quality and consequently on the development of evaluation techniques. The CEC is an appropriate body to have a role in the study of the quality and the management of quality of technical models.

The evaluation of the quality of technical models entails

(a) An assurance of correct coding of algorithms is required and this is probably straightforward.

(b) A statistical model validation is required and this entails comparison with experimental data sets. Considerable care is required in performing such a validation, in particular in determination of an appropriate protocol to ensure unambiguous conclusions to be drawn from the validation.

(c) A model assessment including a scientific review and other less objective aspects is also required and should be given equal weight to a statistical model validation.

Actions which may lead to an improvement in model quality are presented in the report.

1. INTRODUCTION

1.1 Models

Models are used in many and various ways by individuals and organisations within the European Community. In many cases decisions of some considerable importance are based, wholly or partly, on these models.

It will be appreciated that the use of models is pervasive throughout most fields of study as a means of communicating information. This study may be viewed in a wider context; however, its main object is the consideration of the use of *technical* models, particularly those used for assessing major technological hazards.

The Seveso Directive (82/501/EEC) obliges a manufacturer to foresee the consequences of a major accident to his installation by carrying out a risk and consequence analysis and to use this to determine means with which to combat the accident. Concern with major accident hazards has also led to the development of emergency response plans and to the need for advanced techniques of accident investigation. These activities involve analyses which, in turn, contain models.

The quality, that is fitness-for-purpose, of the outcome of these activities will depend, in part, on the quality of the technical input, e.g. technical models.

There is obviously a need to evaluate model quality since they are used to make important decisions about the likelihood and consequences of accidents and the steps which should be taken for their prevention and for mitigation.

Currently there a very large number of technical models available for treating each of the many different aspects of major technological hazards. This is particularly evident in the risk and consequence analysis which is commonly referred to as a 'safety report'.

The quality of models will differ and there is obviously an interest in both the relative quality of one model compared with another and the absolute quality of any particular model. Consequently one is led to the evaluation of model quality and to methods by which model quality, in general, might be improved when necessary, that is, the management of model quality.

The study takes note of a desire for increased harmonisation of the quality of the technical input to 'safety reports' and the possible need for a framework to ensure a minimum standard for that technical input.

Models are used in many ways. In decreasing order of importance, the roles of models are:

(i) As a *predictive* tool for various scenarios of interest that have not, themselves, been tested.

(ii) As a means of summarising extensive analytical and experimental results into a compact form in order to

  (a) assist in the efficient transfer of that knowledge;

  (b) focus on deficiencies in knowledge.

(iii) To provide knowledge in a form accessible to users with varying levels of sophistication in the subject area.

(iv) The model may be used to highlight the sensitivity of the output to the various input parameters. If this is a real physical sensitivity, then we have a means for directing research resources to the areas that show greatest sensitivity.

(v) There will always be a better model tomorrow. If we are concerned with model quality rather than excellence then the model and its evaluated performance can be used as a measure for the diversion of resources from one research area to another for which the benefit/cost ratio is higher.

The nature, role and limitations of models are frequently misunderstood. Models appear to be the principal means of information or technology transfer between a predominantly research community and an operational one. It is often not clear what assumptions are being made at the interface between these communities. The situation is further complicated by the existence of a third community between research and operation, e.g. consultants, software developers, etc., wherein technical model development may also take place. It is, of course, the case that these communities may exist within one organisation.

One frequently hears the comment from the research community that models should only be used by those who have a sound training in the relevant disciplines and understand the intricacies of the model well. This request is often at variance with the requirements of many users that they require a model to be used by staff who have many other, equally important, diversions and distractions and so are not able to develop a level of understanding expected by a research community.

The point to be made here is that it is not clear that the assumptions made by each community about the others are consistent, thus allowing the possibility of misunderstanding. It is also uncertain where the responsibility lies for improving the transfer of correct and appropriate technology and information in a manner that best suits some overall, rather than local, goal.

It is unwise to view a model in isolation from its working environment even if only to ensure an appropriate specification of the model's input and output format. The training and suitability of the user, which can vary markedly, must also be considered. Further to this point, we note that the user (in a broad sense) is also responsible for the correct selection of an appropriate model and it may be here that the greatest contribution to poor overall model performance is found.

This study addresses *technical* models. It does not explicitly concern itself with

(i) techniques for hazard identification;
(ii) techniques for assigning probabilities to various possible scenarios;
(iii) behavioural models;
(iv) models related to plant management;

except insofar as these include technical models. Technical models are most frequently found within the context of consequence analysis.

Evaluation of the quality of safety and risk analyses from a broader perspective has been considered elsewhere - see, for example, Suokas (1988), Rouhiainen (1990) and others.

1.2 Quality and quality management

It is important to realise that model users differ in their requirements and, as a consequence, models to address the same problem may also differ. In addition to a model's correctness, we must also consider its appropriateness for its intended use. Models with similar scientific intent may be quite different, for example:

(i) A model for emergency response may need real-time predictions or access to pre-calculated real-time output.

(ii) A model for planning or regulatory purposes may need to be run several thousands of times to cover a wide variety of source possibilities and environmental conditions.

(iii) A model used for post-accident investigation would be an advanced (research grade) model with little concern about computational cost or other resources required. An investigation can frequently lead to significant model development.

Thus, in discussing any evaluation of model quality, it is necessary to consider the intended purpose of the model. Broadly speaking, we might reduce this difficulty

by noting that (iii) allows appropriate post-accident selection and development while (i) is typically a reduced and simplified form of (ii). Thus we may consider the user of (ii) as the archetypal model user.

Of course, if within (ii) a particular aspect of operation was found to be critical, resort to (iii) for that operation might be considered. This procedure is widespread and should be mentioned. One model (a less sophisticated one) will be used for screening purposes while other models are used for a limited number of critical situations. That is, a user screens scenarios to best allocate resources to be used for further investigation.

It is commonly the case that it is not apparent, or, at least, not adequately documented, what management there is of the quality of the models. For example, in general there appears to be no attempt by model developers to provide an objectively assessed performance rating for their models, nor for that matter any attempt to determine what might be an appropriate or even obtainable performance rating prior to model development.

Presumably, the management is implicit and comes about from an interplay between interested parties who wish to propose or defend their point of view. These parties would include, for example,

government or regulatory authorities
industrial organisations
public groups
insurers
consultancies
researchers (be they from universities or from any of the above groups).

In addition, one would recognise that individuals or small groups from within the above would have immediate goals of seeking excellence, in addition to quality, in their activities.

The lack of more soundly-based and explicit quality management can and does lead to large differences in the predicted outcome of the same scenario.

1.3 The purpose of model evaluation

The general purpose of model evaluation is to:

(i) produce a measure of model quality that may be communicated to interested parties;

(ii) allow a referee to consider a user's use of a model and to assess in a (semi-) quantitative way what weight should be given to the model results;

(iii) through (i) and (ii) to encourage and assist the management of model quality;

(iv) through (iii) to ensure that:

(a) any distortion of outcome produced by the use of different models is reduced;

(b) when technology is transferred to less sophisticated users, it is the correct and appropriate technology that is transferred;

(v) provide a user with documented bounds of applicability;

(vi) provide a user with documented uncertainty in prediction when the model is used within its bounds of applicability;

(vii) identify what improvements are needed in a model.

Several of these points will assist model users (with their various levels of expertise) in the correct choice and implementation of a model and the interpretation of its output.

The consideration of model evaluation procedures may also lead to a framework or structure for determining research priorities by objectively assessing gaps in knowledge, user demand and determining what is implied by a 'state-of-the-art' model. Co-ordination or direction on the management of model quality may provide a useful device for encouraging appropriate model improvements in a cost-effective way.

1.4 What is evaluation?

Model quality or fitness for purpose will need to be evaluated and communicated to interested parties in a non-arbitrary and structured way in order to ensure acceptance and usefulness. We argue here that evaluation would have three components: assurance, validation and assessment.

Assurance here is used to describe whether the computational aspects of the model are correctly treated, i.e. it would not involve any consideration of the physics of the problem (other than how the physics will influence the computational techniques used). This subject refers to correct and efficient coding of algorithms and is, probably, the most straightforward aspect. It is also the area within which there has been the most extensive prior effort.

Validation, in a crude sense, concerns getting the model to predict the experimental results.

Thus the validation of technical models relies on the comparison of model predictions with experimental data. The usefulness of, and the ability to carry out, model

evaluations will depend upon the existence and extensiveness of appropriate data bases and the accessibility of the data bases.

Assessment would offer a more global view in which consideration was given to points such as

(i) Is the physics correct and is all the necessary physics included?

(ii) Are the assumptions/approximations appropriate?

(iii) Are the model limitations indicated?

(iv) Are estimates of model uncertainties provided?

(v) Is the model computationally expensive?

(vi) Is the model (financially) expensive?

(vii) Are the model input and output formats user-friendly?

(viii) Does the model have friendly *selectability*, i.e. will a user be aided in choosing or not choosing this model in a scenario in which the user's judgement is required for model selection?

The points (i), (ii) and (iii) are encompassed by the general terms of 'scientific review'.

We might note that assurance and validation should be objective exercises, whereas elements of a model assessment could be subjective.

The evaluation of model quality will be undertaken by model developers in order to support acceptance of their models. Evaluation will also be undertaken by users to ensure that they can defend their use of the models. Evaluation by independent third parties is of particular assistance to groups seeking advice or direction on model selection and who do not have available expertise to undertake evaluation.

There are obvious advantages if model evaluations performed by different parties produce similar results or, at least, use a similar structured evaluation protocol.

2. CURRENT SITUATION

In order to address the general questions of model quality and that of model evaluation, we consider here a technical perspective on models and their evaluation.

2.1 Models

The term (technical) model encompasses both mathematical and physical models.

2.1.1 Mathematical models: empirical

These models are basically:

(i) Dimensional correlations, which are a technique for simplifying or shortening the presentation of experimental results, or

(ii) correlations which may have been written in a non-dimensional form and, with appropriate caveats, can be extended to quite different materials in different flow situations, but remaining within the dimensionless parameter range of the original experiments. See, for example, various workbook formats such as Britter and McQuaid (1988).

(iii) With suitable argument the models in (i) and (ii) may be extended (extrapolated) outside the parameter range upon which they were based.

2.1.2 Mathematical models: based on the relevant fundamental equations

These models are based, to a greater or lesser extent, on the underlying equations of physics. In the context of turbulent fluid mechanics, only the technique of 'direct numerical simulation' actually attempts to solve these equations. This requires very substantial computer resources (of the order of days of dedicated super-computer time per run for very restrictive flow conditions).

Otherwise there are approximations made to the equations and the model is the solution to these simplified equations. There then follows extensive argument about the correctness or incorrectness of the simplifications. A component of the evaluation of technical models concerns the adequacy of these simplifications and their impact on the quality of the model, that is, a 'scientific review'.

2.1.2.1 Analytical models

Analytical models are, in essence, exact or approximate solutions to a set of equations in a closed form.

2.1.2.2 Numerical models

Numerical models are frequently referred to as computational models, though it must be recognised that there may be significant computational work within analytical models such as the evaluation of complex integrals. Numerical models include models of differing levels of sophistication and complexity.

2.1.3 Comments and perspective

A few comments might be noted:

(i) In a broad sense, there is generally an over-confidence in the ability of mathematical models to provide accurate prediction.

(ii) In particular, there is significant over-confidence in the ability of very large numerical models to provide accurate prediction.

(iii) There is no guarantee that the larger and more complex the model the better the prediction.

(iv) *Experience* has shown that the more complex models are user-sensitive, i.e. the same model, applied to the same problem, can produce different results.

(v) The more sophisticated and complex a model is, the more sophisticated a user must be in terms both of understanding the underlying physical problem and of understanding the solution procedure that is to be adopted.

To provide some perspective on these points, we note that:

(i) Top-flight research models might be solving unsteady (time-varying) problems on 5×10^7 grid points and require weeks of dedicated super-computer time per run.

(ii) Standard research models might be solving steady and unsteady problems on 10^5 grid points and require hours of dedicated super-computer time per run.

(iii) Advanced engineering models used in industry might be solving steady problems on 10^4 grid points and require an hour of dedicated computer time. This is about the level of the most complex models currently envisaged for use within consequence studies. These have not, to my knowledge, actually been used in any consequence study.

(iv) Models typically run in risk assessment activities typically require several minutes on a mini- or micro-computer. However, several different model modules may be required and very many (1000s) of possible scenarios considered.

This range of activities can lead to uncertainty as to what is a reasonable expectation from a mathematical model. As a consequence, there is often confusion when discussing the use of mathematical models and also the possibility of poor communication between the model developer and the model user. This is made worse by the frequent lack of adequate documentation (in the most general sense) and the lack of any external peer-review of the models and their validation (if this exists).

2.1.4 Physical Models

Physical models, sometimes called physical simulations, use wind and water tunnels to model specific scenarios. This technique is well advanced, has strong and weak points and specific limitations, and is aided by the existence of extensive documentation, e.g. Snyder (1981), on a required operating practice. However, this guidance is frequently not followed, leading to the existence of questionable (misleading) data. The technique requires experienced and sophisticated users.

When physical modelling is used to provide quantitative data concerning Major Technological Hazards, it is essential that documentation is provided to ensure the quality of the physical modelling procedure.

Physical models can also be made with no-scale reduction, i.e. the use of full-scale experiments but using similar, but non-hazardous, materials. This activity might also be interpreted as a 'proving' operation for technical models.

Experiments in the field and in the laboratory are often performed in order to provide data to enable the development of mathematical models. This activity is distinct from the use of physical modelling.

For the remainder of this report the term technical model will refer only to mathematical models.

2.2 Evaluation of models

Within the context of Major Technological Hazards evaluating and communicating the quality of a model is an area of study which has only recently been addressed and is, in general, neither simple nor straightforward.

From discussions with an, admittedly small, number of interested groups it would appear that there is no generally accepted, objective procedure in use for the evaluation of technical models used in the area of Major Technological Hazards. There is also no accepted technique for quantifying model quality nor for the effective communication of the outcome of an evaluation procedure to interested parties.

There are, of course, many procedures in use which are considered, quite correctly, to be appropriate for the particular user, though these and the results of their application are not necessarily openly available. These procedures are normally in use by large organisations rather than those with smaller resources.

As an example of an evaluation procedure, Hill (1989) provides an overview of all the technical models available at the National Radiological Protection Board in the UK for calculating rates of radionuclide transfer through the environment.

The NRPB procedure involves 'verification' and 'validation' where 'verification' refers to the process of showing that a technical model is a proper representation of the conceptual model on which it is based, and of checking that the mathematical equations involved have been solved correctly. For computer models verification also includes quality assurance of the computer code. NRPB recognises that different levels of verification may be appropriate for different uses, the most stringent verification being required for regulatory-type models.

The verification procedure is implemented internally by

(i) review of model structure and basic equations by staff other than those involved in developing the model;

(ii) checking computer codes to ensure that programming is correct;

(iii) comparing computed results with problem solutions obtained from other models.

These internal procedures are supplemented by taking part in UK and international model-model comparison exercises and by submitting model descriptions and example results to other organisations and experts for external peer review.

Validation, showing that the conceptual model and computer code provide an adequate representation of the problem, requires a comparison of model results with experiments that were not used in the model development. When independent experimental data bases are not available more qualitative techniques are used. The quantification of what is meant by an adequate representation is somewhat open and will depend on the intended model use.

Other large organisations will have some equivalent, similar evaluation procedure. What is of prime importance here is that an evaluation procedure has been clearly stated and is communicated to interested parties.

Although there is no general accepted procedure for the evaluation of technical models within the context of Major Technological Hazard, there does appear to be a demand for an improvement in the management of the quality of technical models, particularly through model evaluation procedures. Arguments supporting this view may be found in:

(i) Areas of similar activity currently addressing such matters, e.g.

(a) within the nuclear energy area (see della Logia and Royen (1988), Proceedings of 'Water-cooled reactor aerosol code evaluation and uncertainty' sponsored by the CEC DGXII and other activities at JRC/ISPRA);

(b) within the area of the atmospheric dispersion of pollutants - see, for example, workshops and recommendations arising from the American Meteorological Society and US Environmental Protection Agency, e.g. see Fox (1981), USEPA (1984);

(c) VAMP (1990) (Validation of Model Predictions) is an International Atomic Energy Agency/CEC co-ordinated research programme on validation of models for the transfer of radionuclides in terrestrial, urban and aquatic environments and acquisition of data for that purpose;

plus, I presume, many others. Not surprisingly, these activities tend to arise some time after the initial rapid development and proliferation of models.

(ii) Recent general requests from within the area of study, e.g.

a discussion document from Pineau (1990) concerning 'Research in Areas of Accident Prevention and Response' prepared for an OECD Workshop, Boston, May 1990, contains the following points:

"Various computer program are available to predict dispersal, but their validation has been mostly confined to simple configurations. Large-scale testing and the study of existing data should yield more detailed information about the phenomena."

"It is not enough to develop software for calculating the consequences of discharges or leakage on a full scale. It must be validated...and little has been done on the subject."

"How are research findings disseminated today? How can dissemination be improved?"

"Safety studies are now routinely performed and are partly automated (use of software and software packages). Is there not a risk that such studies, which ought to be adapted or adaptable to each situation, become less efficient? To what extent has validation occurred, in particular after major accidents?"

In the discussion subsequent to the discussion document it was noted that:

"One reason for failure to apply research was poor diffusion, e.g. information in a highly technical language may not be appropriate to a user – for example a fire brigade service which needs simple tools to be used under stress."

"Some simple ways to diffuse information were discussed:

- publish a list of research undertaken at a regular frequency;
- summarize the results and digest them within a computer package. In that case, a minimal validation of the content is an absolute necessity. What's the use of a large diffusion of out-of-date information?"

There seemed as much concern with the dissemination of correct information and a coherent structure for research activities as there was for identifying specific research needs, that is, ensuring that the research reflects the needs of the user and is structured and undertaken with the needs of the user in mind.

Technical models are, of course, the principal means whereby research finds its way to the user. Two corollaries of this are that information on the quality of the model should also be passed to the user and that different users will require models of different character.

(iii) Recent specific contributions from within the area of study:

(a) The CEC sponsored SHARE (Safety management and hazard assessment research co-operation in Europe) programme stated the following objectives:

1. develop an overview over research needs;
2. direct research at most appropriate problems;
3. optimise research capabilities and co-ordinate effort;
4. develop a communication network;
5. disseminate information;
6. promote practical use of research findings;
7. promote the development of integrated approaches to management of risk.

The question of quality management of technical models encompasses many of these objectives.

(b) 'User's guide to information systems useful to emergency response planners and responders, available in OECD European member countries'

(October 1990), produced by SERETE for CEC DGXI. This guide summarises, under specific categories, the many systems (typically a group of linked models) available and provides a brief comment on users' experiences. In some cases the names of the various constituent models are mentioned, in others not. What is specifically not addressed is the quality of the models themselves (except in one case where users doubted whether the models used were 'state-of-the-art'). The book by Hanna and Drivas (1987) is a similar exercise in which descriptions are supplied but not a model evaluation.

2.3 Experience of model evaluation

There is experience of model evaluation in a formal, statistical sense provided over the past decade with regard to atmospheric dispersion codes, sponsored by American Meteorological Society (Fox, 1981, 1984) and the United States Environmental Protection Agency (which has produced guidelines on evaluation procedure) (USEPA, 1984). This has led to a number of conclusions that can place this activity in context, that is:

(i) It has been increasingly widely accepted that a purely statistical assessment is not adequate and attention must be given to the quality of the physics that have been included in or excluded from the model. This is more difficult to objectively evaluate although it has been suggested that equal weight should be given to both activities.

(ii) Models should only be applied to physical situations similar to those used to derive them.

(iii) Although initially many model performance measures were used it became apparent that a reduced set is desirable to aid comprehension of the results.

(iv) It is often the case that operational models lag behind research grade models by decades.

(v) In addition, Knox (1985) stresses that a model cannot be improved unless there is a precise understanding of its performance in certain well defined, physical situations; knowledge of the model's sensitivity to input parameters is necessary in order to decide where emphasis should be placed in research effort. That is, care and comprehensiveness must be used when selecting experimental data sets for model comparison.

Recently a model validation exercise has been completed by Hanna et al. (1991) which compares in a statistical manner the results of 14 models when compared against

two sets of trials comprising 4 and 3 experiments respectively. Typically three measurements (concentrations on specific arcs) were used for each experiment. No assessment of the appropriateness of the physics within the models was attempted nor was a comment offered on the parameter range to which the experiments referred.

Hanna et al. (1991b) report on the results of a Model Evaluation Panel that addressed hazardous gas dispersion models. Their first recommendation regarding model evaluation was that "an independent international peer-review panel should be established".

The recent benchmark exercise on major hazard analysis, performed for the CEC (Contini et al., 1991), identified the large range (several orders of magnitude) of concentration and risk estimates arising when different groups of investigators consider the same plant. A major contributor to this range was the requirement of extensive engineering judgement necessary throughout the investigation.

A more restricted exercise, undertaken with well defined common boundary conditions, still resulted in an order of magnitude variation in concentration and greater variation on risk. These were attributed to differences in source and dispersion models, and their implementation.

Of relevance to our current study are the observations that

(i) many models currently in use produce concentration variations of an order of magnitude and not, as is sometimes assumed, a matter of 10% or 20%;

(ii) the use of the models may be as important as the models themselves, i.e. the interaction between models and engineering judgement is important;

(iii) there is a need for greater transparency as to the substance of the models;

(iv) there is a need for the development and validation of more consistent models.

Finally we note that even a comparative presentation of model attributes, by an independent evaluation, such as Table 1 taken from Hanna et al. (1991) which is a by-product of a model evaluation, is helpful to a model user.

TABLE 1

Attributes of models*

Attribute	Model													
	ADAM	AFTOX	ALOHA	B&M	CHARM	DEGADIS	EAHAP	HEGADAS	OB/DG	PHAST	SAFETI	SLAB	TRACE	WHAZAN
Surface roughness (z_0)	✓	✓ ($z \geq 0.005$ m)	Rural or urban	–	–	✓	✓	✓	–	✓	✓	✓	✓	✓ ($z_0 \geq 0.01$ m)
Avg. Time (min)	✓	✓	10 (rural) 60 (urban)	–	10	✓	10	✓	–	10	10	✓	✓	10
Height of wind (m)	✓	–	–	10	–	✓	10	✓	–	–	–	✓	✓	–
Receptor height (m)	0	✓	0	0	✓	0	1	0	0	0	0	✓	✓	0
Release:														
Continuous	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Instantaneous	✓	✓	–	✓	✓	✓	✓	–	–	✓	✓	✓	✓	✓
Variable rate	–	–	–	–	✓	✓	✓	✓	–	–	–	–	✓	–
Aerosols	✓	–	–	–*	✓	–*	✓	–*	–	✓	✓	✓	✓	✓
Flash fraction	✓*	–	–	–*	✓*	–*	✓	–*	–	✓	✓	–*	✓	✓
Jet mixing	✓	–	–	–*	–	–	✓	–	–	✓	✓	✓	✓	✓
Relative humidity	✓	–	–	–	✓	✓	✓	✓	–	✓	✓	✓	✓	✓
Ambient pressure	✓	–	–	–	✓	✓	–	–	–	✓	–	–	–	–

*Explanations of terms are given in Section 2.3. A check (✓) indicates that the model accounts for variation in that attribute. A dash indicates that the model does not account for variations in that attribute. A number indicates the values assumed by the model. An asterisk indicates that we accounted for this attribute in our model initialization assumptions.

Table 1. A comparison of model attributes (from Hanna et al., 1991)

3. THE QUALITY, AND QUALITY MANAGEMENT, OF MODELS

3.1 The quality of models

Evaluation of the quality of models was introduced in §1.4 and we now extend that discussion.

Assurance of correct and appropriate coding is an essential part of the evaluation of the quality of the model. This should be straightforward and will not be considered further.

3.1.1 Types of model evaluation

Several techniques have been used with the intent of providing an evaluation of models and/or their use. We note three commonly used approaches:

(i) comparison between prediction and experiment

(ii) code comparison exercises

(iii) benchmark exercises

and these have quite distinct goals.

A comparison between prediction and experiment is currently referred to as a validation exercise. Comparisons have been made on both raw and dimensionless variables, performed by the model developer or a neutral agent and based on experimental data which has or has not been used in model development.

A preferred arrangement would be a comparison using raw variables, performed by a neutral agent using data not available to the model developers. Model validation is discussed further in the next section.

Code comparison exercises require no experimental data and compare directly the output from various codes for a range of scenarios. This exercise allows no absolute assessment of the models but can be useful when determining whether a consensus exists amongst model developers or in providing a measure of the lack of consensus currently existing in the activity. With no other information available, model performances showing large lack of consensus for certain scenarios are likely to encourage further work in improving the understanding of the relevant physics dominant in those scenarios. The code comparison exercise is particularly useful when considering the performance of models outside the parameter range for which experimental data exist.

The term code-comparison is also used for the statistical comparison of the the performance of two models against the one experimental data set. The technique allows determination of the statistical significance of differences in model performance against the data set.

Benchmark exercises occur in several forms. The simplest are essentially a code comparison exercise wherein a specific problem scenario is stated and models are used to provide predictions which are then compared.

A far more enlightening exercise will require the users to also be responsible for the correct selection of appropriate scenarios, e.g. see Contini et al. (1990). This allows a comparison of the broader methodologies and, of relevance here, the correct selection and implementation of a model and the correct interpretation of the output. Contini et al. (1990) noted that very significant variations occurred due to inappropriate model selection: variations comparable to any variations between the models themselves.

Finally, we note a somewhat different exercise, noted in §2.2, which is really a user's guide comparing models on the capabilities described by the developers and, possibly, including users' reactions to the models. Similar exercises are sometimes undertaken by journals for assessing software. These rarely address the physics of the models but do serve a useful purpose when considering the more subjective aspects of an assessment procedure.

3.1.2 Techniques of model validation

The validation of air quality models has been an active area of study since the early 1980s, driven in part by regulatory requirements on operational models.

We can expect that the development of validation procedures for models concerning major-accident hazard installations will follow a similar structure, or at least the satisfactory parts of it.

A validation procedure must be designed to determine how well a model is able to predict certain scenarios and allow communication of that information to the user.

This superficially simple requirement covers a number of difficult questions.

However, if we restrict ourselves to cases where

(i) an appropriate model has been selected for comparison with observational results;

(ii) the model is operated correctly and the output correctly interpreted;

(iii) the model output data is consistent with the form of the observational data, e.g. the data is based on the same averaging time;

(iv) the observations are representative of the scenario that we wish to model;

(v) the observations have not been used in the development of the model, and

(vi) the spatial location of the observation and prediction are coincident,

then progress can be made.

Thus we have a number of pairs of observations and model predictions and these may be statistically compared.

The statistical measures for comparison may be very extensive although it is realised that "a reduced set of performance measures is desirable to aid comprehension of results" (Hanna, 1988).

A suggested simple set has been

$$(\overline{C_0 - C_p})/0.5(\overline{C_0 + C_p})$$

and

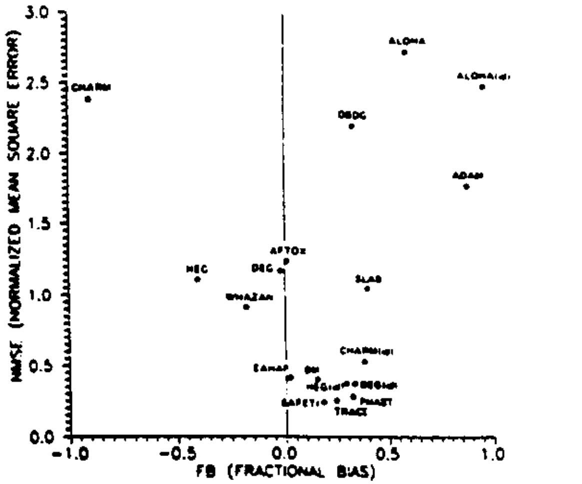
$$\overline{(C_0 - C_p)^2}/\bar{C}_0\bar{C}_p$$

where the $\bar{}$ sign indicates an ensemble average. Such a simple set has been found to be very useful - see Figure 1 from Hanna et al. (1991) in a preliminary validation of dense gas dispersion models.

It is not intended to elaborate here on other possible statistical measures that might be used other than to note that various measures can be chosen and rather simply evaluated.

Of more concern is the need for a clearly defined protocol under which an unambiguous statistical model evaluation may be made. Hanna (1991b) addresses some of the issues relevant to developing a suitable protocol. This is not a simple task but is of prime importance if the results of the evaluation are to be usable, and to allow the evaluation to be used for quality management.

Item (vi) above is sometimes relaxed, that is, identical pairing in space may be an unnecessarily severe restriction on the comparison; see for example Hanna (1988). Such a relaxation may be less permissible when considering Major Technological Hazards (than for conventional air pollution problems), where population centres or possible ignition sources may have to be specifically included in a consequence assessment.



Weighted average fractional bias, *FB*, and normalized mean square error, NMSE, for concentration predictions for each model , for Desert Tortoise and Goldfish data. The "id" suffix indicates that the initial dilution assumption has been applied to that model. Note that if a model tends to overpredict, its *FB* will be less than 0.

Figure 1. Taken from Hanna et al. (1991)

3.1.3 Uncertainties

A model validation, based on a comparison of model results with experimental data, requires care when the phenomena under study are stochastic in nature. Concern about uncertainties arises in two somewhat distinct ways.

In model validation there will be uncertainties associated with the stochastic nature of the problem and with data errors. The comparison of model and experiments will include these effects in addition to the inaccuracy of the model itself. Stochastic effects and data errors will place a restriction on what quantitative results can be deduced from a model validation.

This view of the problem may be inverted when a model is used to predict the consequences of a particular scenario and stochastic effects and data errors will lead to uncertainties on the model output. Information on the uncertainty of the model prediction may be as valuable to a user as the model prediction itself. The consequences of known uncertainty on model input parameters may be directly evaluated by running a model for the range of input parameters.

Model uncertainty is a well developed area of study with its own literature: see, for example, Hanna (1988). For example, the total model *uncertainty* might be measured by $\overline{(C_p - C_o)^2}$ where 'o' indicates observed and 'p' indicates predicted. three areas contribute to the total model uncertainty.

These have been categorised (Hanna, 1988) by expanding observed and predicted data, as

$$C_0 = \bar{C}_{0a} + C_0' + \Delta C_r$$
$$C_p = \bar{C}_p + C_p' + \Delta C_p.$$

The observed concentration C_0 consists of

$\bar{C}_{0a}$ is the actual ensemble average in the absence of instrument errors

C_0' is the observed stochastic random variability

ΔC_0 is the data error in C_0

$\bar{C}_p$ is the predicted ensemble average

C_p' is the predicted random variability; this is frequently assumed to be zero and thus leads to an underestimation of the total model uncertainty

ΔC_p is the error due to data input errors.

Thus if there is no correlation among the components then the total model uncertainty is given by

$$\overline{(C_p - C_0)^2} = \overline{(\bar{C}_p - \bar{C}_{0a})^2} + \underbrace{\sigma^2_{c_0} + \sigma^2_{c_p}} + \underbrace{\overline{\Delta C_0^2} + \overline{\Delta C_p^2}}$$

Total model uncertainty	Model physics error	Stochastic uncertainty	Data errors

That is, the total model uncertainty is made up of:

(i) the model physics errors;

(ii) stochastic uncertainty due to the turbulent (stochastic)nature of the phenomena, and

(iii) input and output data errors.

It will be apparent that the existence of the last two terms limit the extent to which the model may be improved by reducing the model physics error. This might be reinterpreted by suggesting (Hanna, 1988) that there is an optimum degree of model complexity that minimises model error.

The interest of this report is, in essence, within the first term, i.e. the model physics error, although the last two terms will be relevant in

(i) their influence on the assessment of the model physics errors;

(ii) the uncertainty in the predictions of a model, and

(iii) reflecting the sensitivity of the model to lack of knowledge of the model inputs, e.g. the relevant meteorological conditions, the source release rates etc.

Specifically we note that even if the model physics error was zero there is an underlying or inherent uncertainty due to the stochastic nature of the problem and to data errors. An argument can be made for attempting to assess this inherent uncertainty prior to any model development, thereby providing an estimate of what is a reasonable goal at which a model developer should aim.

Similarly, estimates of the influence of errors in data input parameters provide a useful perspective on the problem.

The CEC currently supports research into uncertainty analysis, though within the context of radiological problems, e.g. see Elderkin and Kelly (1990).

A related technique is the use of expert judgement for uncertainty analysis (see, for example, Cook, 1991) in which the knowledge of experts is directly used to determine the uncertainty. This may take the general form and concern judgement about the accuracy of a prediction using whatever tools or methods that may be available. The

technique is also used when a particular model structure is used and expert judgement is sought concerning the values of 'constants' or parameters within the model. The CEC currently supports a study on the applicability of methods of structured expert judgement in assessing the uncertainty of accident consequence models in the European Community's COSYMA code.

3.1.4 Further points on model evaluation

Our basis for model evaluation must rest in large measure on a comparison between experimental observations and predictions. However, several questions will require consideration prior to the implementation of any evaluation procedure, such as:

(i) Do we require that the physics is correct or is the sole criterion to be the provision of the correct answer?

Experience indicates that an evaluation of the appropriateness of the physics is required, particularly if the model is to be applied outside the parameter range for which it has been validated or if the independent data for validation is limited. However, how an assessment of the physics may be incorporated into a model evaluation protocol is uncertain.

(ii) How do we evaluate the items, some subjective, under the heading of assessment, as distinct from validation?

Questions of assessment, including the appropriateness of the physics, would seem to be best handled by peer review, either in its conventional form of the published literature, or in a more specific and structured format where, for example, neutral, competent parties were used.

(iii) If some output variables, e.g. concentration, are more important than others, e.g. relative humidity within a cloud, should the model's ability to predict the more important variables be given greater weight?

Mercer (1988) notes the large number of variables that may be used for a model evaluation procedure. A pragmatic solution will be to perform an evaluation on a large number of variables, consistent with resources, and to separately weight the evaluation guided by an intended use of a model. Further reference to air pollution model evaluation shows that a weighting system is to be agreed upon by interested parties based on anticipated problems at a particular site (Hanna, 1988).

(iv) Do we rank a model that exactly reproduces the experimental results higher than one that has a consistent error of 10%, if

(a) the experimental error is 20%, or

(b) the simplifications to the physics within a model could not provide better than 20% prediction?

There are statistical techniques available (Efron, 1982) for determining whether the results of one model are significantly (in a mathematical sense) better than another or whether the difference in observed performance is not statistically significant.

3.2 The management of the quality of models

Evaluation of model quality and the communication of that evaluation to interested parties may be seen as a method for managing the quality of models, that is, a reactive approach.

An active approach requires a consideration of the process of model development. In the broadest sense the main contribution to the management of the quality of models is likely to be the awareness that there *is* a process of model development and the *understanding* of what that process entails.

It is not intended to consider this under-studied area here except to note the basis on which a model development will take place and the resulting compromises faced by a developer.

It is recognised that the fundamental equations will not be exactly solved except in relatively few, very specific, cases. As a consequence, technical models are developed based on approximations to the fundamental equations. We note that there will be disagreement and argument amongst experts concerning

(i) the appropriateness of the approximations;

(ii) the appropriateness of the conversion of the equations and the boundary conditions into a form to be solved, e.g. into a finite difference form, and

(iii) the appropriateness of the technique used to solve the equations.

In the context of Major Technological Hazards the first of these will have the largest influence on model quality and produce the greatest disagreement. These disagreements may be scientific or may be based on experience. They may also reflect the different, valid, intentions of the developer.

In addition model development will need to consider the intended use of the model, as outlined in §1.2, and the questions that a model user will ask, such as:

(i) Does the model produce the *right* answers?

(ii) Are the model inputs easily obtained?

(iii) Is it clear what are the model's limitations?

(iv) Is it clear what accuracy might be expected for the model results?

(v) What level of extrapolation is implicit in the model away from data used for model development and what are the likely effects?

(vi) Is it financially expensive?

(vii) Is it computationally expensive?

As a consequence, a model developer will have to face various trade-offs and these should be recognised as an essential part of model development.

(i) *Local versus general*

A model allowing very general use will be more complex and expensive, require greater computer resources and be likely to be less accurate than one whose goals are more modest. If the user's problems are always quite similar then the latter choice would be preferred, the model here tending to act as a rather complex correlation.

(ii) *Single or modular*

A similar question arises when considering the modular nature of many models. It is generally the case that a user will have a suite of models that are to be used serially or in parallel. In other cases the user may be unaware that the overall model is accessing various modules. A modular nature to models is attractive in that it

(a) is more transparent, and therefore
(b) is more easily interrogated, and
(c) the modules may be more easily upgraded.

However, a modular development also requires that inter-module connections are developed, evaluated and maintained.

(iii) *Computationally expensive versus inexpensive*

The point to be made here is that some credit must be given to models that are computationally inexpensive. There is little to be gained by having a model which will give the correct answers but is too computationally expensive for the user to use in the necessary manner.

(iv) *'Consensus' model versus 'New' model*

Model development and use is an 'ongoing' activity; however, continued changes in what regulators deem to be acceptable models can lead to inefficient use of resources, particularly in small- to medium-sized companies which will not have internal resources to provide broad model evaluation.

Thus some consensus model should change in discrete steps while state-of-the-art models, by their name, are continuously evolving. Changes in a 'consensus' model should only be considered when serious deficiencies are documented or when major changes in understanding have occurred. Experience suggests that a period of five to ten years is typical for 'consensus' model change.

However, models treating new phenomena or new scenarios that are of interest will be immediately accepted if the demand is there.

The existence of these trade-offs and how they have influenced any particular model development, should be recognised when any protocol for model evaluation is being formulated.

Finally we note one further dimension to model development. It is common for more sophisticated (research grade) models to be used for the development of simple (possibly called state-of-the-art) models and the best of these become, *de facto*, consensus models. A legitimate exercise in the management of the quality of models would be to ensure that this process was recognised and assisted when appropriate. It does, however, presume that the more sophisticated model also has excellence and this need not be so.

4. POSSIBLE ACTIONS TO MANAGE THE IMPROVEMENT OF THE QUALITY OF TECHNICAL MODELS

4.1 Introduction

The following questions relate to the use of technical models within the context of Major Technological Hazards. The responses reflect the experience of the author and discussions held during this study.

(i) How accurate are these models at making predictions? We do not know.

(ii) Are these models using state-of-the-art knowledge in their construction? Technical models undergo an evolutionary development, sometimes continuous, sometimes frozen for several years and then improved or replaced. Thus some are trying to maintain a state-of-the-art while others go through periodic, substantial changes when their deficiencies become embarrassing. There are also many examples of obsolete models which are locked within a larger operational model and are difficult (both scientifically and administratively) to disentangle. This last point reflects the observation that it is disconcerting for users to be required to change their procedures too frequently. In essence, this question requires further consideration of what is desirable in the development of models and what is implied by the term 'state-of-the-art'.

(iii) Have the models been validated? Not in any formal scientific (as distinct from crude) sense.

(iv) Have their limitations been documented? In my experience, never satisfactorily.

In this report it has been demonstrated that there is a need for the consideration of the quality and quality management of technical models that arise in the study of Major Technological Hazards. It also suggested that the CEC was in a position to address these matters or to encourage their consideration.

In this chapter we consider what actions might be undertaken to address the quality and quality management of technical models, indicating areas in which the CEC could provide support.

We first determine some possible common ground among interested parties and some constraints on action that should be recognised and finally list some general and specific recommendations.

4.2 Possible common ground

Before considering actions which may lead to an improvement in the quality and quality management of technical models, a statement of possible common grounds amongst interested parties and one regarding any constraints on possible actions will be useful. This, and the next, section address these points.

Common ground includes:

(i) The use of various different technical models will lead to a distortion of the technical content of 'safety reports'.

(ii) It is uncertain what the magnitude of the distortion may be; the benchmark exercise (Contini et al., 1991) indicates that the magnitude is unlikely to be acceptable.

(iii) Quality management of technical models appears to be based on an 'ad hoc' approach relying on the good intentions of individuals or groups within organisations to ensure that they are able to defend their use of the models; to defend them within and external to their organisation.

(iv) There is a need to encourage an increased degree of quality management of technical models throughout the Community. Note that the final remarks made from the benchmark exercise might be summarised as:

- reduce variability
- provide transparency of assumptions made
- be aware of and indicate the limitations of models
- development and validation of more consistent models.

(v) As some member nations have a longer experience in the submission of 'safety reports' to Competent Authorities etc., there is a need to encourage dissemination of information (subsequent to appropriate quality management) within the Community.

(vi) The interplay between regulators and industry is a significant use of resources but it is also the principal means of ensuring improved model quality and is probably essential.

(vii) Within an industry there is a balance met between concerns about safe operation and responsible commercial activity. This balance must be recognised and not influenced without good reason.

(viii) There is also an interplay between concerns regarding safe operation within an industry and the view that insurers may have of that industry. This interplay must also be respected.

(ix) Continual improvement in the quality of technical models is required but there also needs to be a focusing of attention on those areas where need is greatest. Methods to assist in this focusing of effort should be encouraged.

(x) It is preferable for actions to be voluntary rather than mandatory.

4.3 Standards

In many activities, and in particular those involving safety, a society will accept the setting of various Standards for materials and Standards for operating procedures.

There would seem to be a reasonable case for the requirement of some form of Standards as a technique of quality management, in the development and application of technical models.

Similarly, the setting of Standards for the presentation and reviewing of experimental observations for use in model evaluation might be considered.

Apart from Standards on the quality assurance of computer codes and, in some cases, on the implementation of particular computational techniques, there does not appear to be any Standardisation concerning the evaluation of technical models within the context of Major Technological Hazards.

4.4 Who should assist in improving quality management?

From the previous comments and others it is apparent that a need does exist for guidance on the quality of models and, by implication, what models are and what might be expected of them. It follows also that some objective measure of model quality also clarifies questions concerning the setting of achievable goals and of determining research funding priorities.

It is unclear what guidance is currently available on questions of model quality but it is clear that little support has previously been provided for studies in this area.

Within any country there is likely to be a great difference in the resources available for attention to these matters when comparing a large national or multinational company with a small or medium-sized company. The larger company may be in the process of model development itself whereas the small/medium-sized company will be trying to purchase advice from consultants or to acquire available models with little to guide their choice or the subsequent method of use.

In a similar manner the resources of the regulatory agencies within a country and between countries and their different levels of experience in the field will lead to different views on the need for guidance in these matters.

However, it would seem to be the case that there is an *a priori* need for guidance as to the quality of technical models and to the implementation of measures for improving model quality.

The CEC's Major Technological Hazards and the Industrial Hazards prorammes of DGXII and JRC respectively and other national research programmes are targetted, in part, at increasing knowledge concerning relevant chemical and physical phenomena and at improving the quality of technical models by incorporation of that knowledge.

It might be argued that there is a mismatch between sponsored research on major technological hazards and the development of an effective procedure for assessment of that research and its translation into usable applied results, i.e. technology transfer. This area of formalised model evaluation is one in which the CEC could, if deemed appropriate, have a significant co-ordinating role. It is one of several areas, outside strict technological research, in which much good may come with quite modest investment.

Further to this, the CEC would be an appropriate body to have a role in this activity because of its implied regulatory role in the Seveso Directive and because of its active research programme on Major Technological Hazards. This is particularly so if there is concern (Otway and Amendola, 1989) that "different national approaches to the safety studies...could result in the distortion of industrial competition within the EEC".

There are also many other organisations, national and international, concerned with the quality, and quality management, of technical models and some co-ordination of activities will be required.

4.5 Constraints on action

Before considering action, the following constraints should be borne in mind:

(i) Any encouragement on quality management would want to make use of the existing, very substantial, expertise and experience residing within the various industries and regulators rather than act to disrupt their activities.

(ii) It is important to bear in mind the local situation existing in member nations.

(iii) It is important to note that within any nation the goals and resources of

(a) regulators
(b) large companies
(c) medium- and small-sized companies

will be quite different. Their different views will have to be accommodated.

(iv) Any structured (centralised) model evaluation procedure may lead to an acceptance of favourable assessed models and a reduction in further development. This may be an unsatisfactory outcome.

(v) The comment (iv) may be viewed differently. General acceptance of a favourably evaluated model may be a useful indicator for the timely diversion of resources elsewhere.

(vi) The availability of favourably evaluated models, particularly to small- and medium-sized companies could lead to a reduction in the number of competent personnel employed. As they are likely to be necessary in order to handle local factors, such an outcome would be unsatisfactory.

4.6 Actions that may encourage improvement in the quality of technical models: General

(i) Conducting *research programs* into all aspects that will contribute to a safety report including general risk management procedures and specific technical issues.

(ii) Ensuring that such research programs are *undertaken jointly* by members to assist in information transfer which will lead to common bases for technical model development.

(iii) Ensuring that this and other research is *transformed* into technical models as *soon* as possible.

(iv) Ensuring that the research results and subsequent technical models are *communicated* to possible users, i.e. to facilitate communication regarding models.

(v) Ensuring that the research results and subsequent technical models *and* existing technical models are evaluated in order to provide a degree of quality management. The model evaluation involves both a statistical validation and a scientific review. It may be that in attempting model evaluation it is noted that a validation data base is not available. If it is the case that the model is frequently used and its outcome is of consequence in a 'safety report' (that is, the uncertainty in this model was shown to be an important contribution to the overall uncertainty of the report), then the lack of a suitable evaluation would naturally divert research funds to that area in order to provide the necessary data.

(vi) Facilitate the implementation of models that have undergone some degree of quality management.

(vii) Communicate the need for quality management of models to all users.

(viii) Facilitate intercomparison of techniques including models to allow users to assess their own position regarding quality management.

(ix) Encourage training in the use of operational models together with adequate and appropriate information on the limitations of models and in the correct interpretation of the output.

(x) Address the question as to whether some centralised structure to this activity is required or whether it is more appropriate to facilitate the development of a distributed structure by providing support.

(xi) **Research contracts (1)**

These might be worded to indicate that the intent is for someone to use the information that has been gathered. Consequently the requirements for documentation might be addressed.

(xii) **Research contracts (2)**

Support might be given to research which does address the harmonisation of tools impinging on the development of 'safety reports'.

(xiii) In general to increase awareness that the quality and quality management of technical models are important and substantial issues.

4.7 Actions that may encourage improvement in the quality of technical models: Specific

(i) No action

Reliance on the natural diffusion of correct and appropriate information will lead to an improvement in model quality but over an extended period of time.

(ii) Action leading to mandatory requirements

The use of a common framework for 'safety reports' together with an evaluation of model quality would lead to an improvement in model quality over time.

Such an approach is unlikely to achieve support due to the diverse positions of interested parties and with the difficulties of determining an acceptable level of model quality.

An easier path to follow might be to encourage the harmonisation of models and to facilitate their use. That is, we might, more effectively, address the quality and facility of models, the choice of models remaining free. But how should models be evaluated and is a standard framework necessary for doing so? A standard framework is not essential but there is little reason for not adopting one.

(iii) Non-mandatory actions

(a) Conferences and Workshops

There would seem to be a good case for the sponsoring of both conferences and workshops on these issues.

Conferences are an effective way of increasing awareness of an issue and assisting the transfer of information among interested parties.

Workshops allow for more precise attention to specific goals. These can be successful, particularly if discussion documents are prepared prior to meetings, and the conclusions of the workshop are formally prepared and speedily published.

The workshop may be the appropriate forum for the scientific review of models (see Weil, 1984, for this use in the context of atmospheric dispersion models).

It is unlikely that a statistical model evaluation could be undertaken at a workshop, but the setting up of a protocol and the interpretation of the results would be suited to a workshop format. Perhaps the next step in the consideration of the quality of technical models would be a conference or workshop on the topic. A conference is to be preferred as this allows wide participation.

For a workshop to be useful, some restriction on attendance is desirable.

The following topics are matters that could usefully be given wider discussion at a conference or workshop.

(b) Data base development

There have been many extensive (and expensive) field and laboratory experiments on Major Technological Hazards, several funded by the CEC. It might be argued that there has been less investment in ensuring that this data is used, and used wisely, than the cost of acquiring the data would warrant.

The CEC could encourage the setting up of a data base arising from experiments funded by the CEC and including experimental results available elsewhere.

This would be assisted by ensuring that any contracts contained (and funded) a requirement that the data is made available in an accessible format.

It would not be feasible nor wise to require that all the data should be available; a summary form, preferably on disc, is all that is required, together with an indication as to where access to more detailed data might be obtained. A contractor might sensibly wish to allow the archiving of only the best quality data.

In addition, it would seem necessary to provide some peer review of the experiment, the techniques used and the data archived. An independent assessment of these points would be a helpful guide to any user. CEC would need to provide for communication to interested parties of the existence of the data.

It must also be asked whether this is really worth the effort. Maybe there are only a few organisations interested in the data and they will ask for it if necessary.

The response must be that

(i) in general the experiments have been done principally in order to assist in the eventual development of technical models;

(ii) the experimental results are only of use if they are available, and available in a form that facilitates their use;

(iii) if a contractor is to undertake (i) as part of a contract, it may as well be done with a more general user in mind;

(iv) experience has shown that there can be a difficulty with experimental data used for model comparison in that the data may be released over a period of time and in different forms. Consequently more than one data set may exist and these may differ. This is to be avoided by the release of a unique data set; and

(v) the selection procedure for proposals should have ensured that there would be wide interest in the results.

The point to be made here is that the availability of good quality archived data sets should come about as naturally as possible without the need for any undue administrative effort.

There are also obvious attractions to data bases being produced in a consistent format: a difficult requirement unless some centralised direction is offered.

It may be that the principal requirement here is the acquisition of existing appropriate data bases and subjecting them to a form of review and then reformatting them into a consistent framework.

As an example of data base documentation, Table 2 is taken from the model validation by Hanna et al. (1991). Another example of data base documentation, Table 3, essentially locates and categorises the data.

Modelers' data base for Desert Tortoise ammonia experiments

1. Description				
Test number	1	2	3	4
Date	8/24/83	8/29/11:20	9/1/83	9/6/83
Time	16:37 PDT	PDT	15:37 PDT	18:15 PD
2. Release conditions				
Exit pressure (atm) (avg.)	10.00	11.02	11.23	11.64
Exit temperature (K)	294.7	293.3	295.3	297.3
Nozzle diameter (m)	0.081	0.0945	0.0945	0.0945
Spill rate (kg/s)	79.7	111.5	130	96.7
Spill duration (s)	126	255	166	381
3. Site conditions				
Ambient pressure (atm)	0.897	0.896	0.895	0.891
Rel. humidity (%)	13.2	17.5	14.8	21.3
Air temperature @ 2.5 m (K)	302.4	303.9	306.9	306.0
Soil temperature (K)	304.8	303.8	304.8	304.0
Wind speed @ 2 m (m/s)	7.4	5.8	7.4	4.5
(3-min avg. over 11 sites)				
σ_u (m/s)	1.2	0.7	1	–
σ_θ @ 2 m (deg)	5.7	7.5	8.3	5
Friction velocity, u_*, (m/s)	0.44	0.34	0.45	0.27
Monin–Obukhov length, L, (m)	93	96	571	45
Cloud cover (%)	1	4	70	1
Pasquill stability class	D	D	D	E
T (16 m) – T (2 m) (°C)	0.87	0.46	0.13	0.9
4. Peak concentrations (ppm)				
Averaging time (s)	80	160	120	300
100 m arc	50000	82200	76900	57300
800 m arc	8800	10800	7099	15400
Other arcs	328 (3500m) 101 (5500m)	5000 (1400m)	693 (1400m)	3390 (2800m)

Table 2. An example of data base documentation, taken from Hanna et al. (1991)

Fluid Dynamics Panel Working Group 14 : "Selection of Test Cases for CFD Validation"

PLEASE READ THE INSTRUCTIONS ON THE REVERSE SIDE PRIOR TO COMPLETING THIS QUESTIONNAIRE.

MODEL NAME :

ORGANISATION AND COUNTRY OF ORIGIN :

TUNNELS WHERE MODEL TESTED :

DATES OF TESTING :

1. BRIEF DESCRIPTION OF MODEL GEOMETRY.

component	description
model support details	

2. DESCRIPTION OF FLOWS TESTED.

Mach No range	
Incidence range	
Reynolds number range	

3. MEASUREMENT TECHNIQUES USED.

4. PURPOSE OF TEST.

5. PHYSICAL FLOW FEATURES IDENTIFIED.

6. FLOW REGIONS INVESTIGATED.

7. PROPERTY RIGHTS ON DATA.

ORGANISATION OWNING DATA PROPERTY RIGHTS	
IS DATA FREELY AVAILABLE NOW ?	
IF 'NO', WHAT ACTIONS ARE NECESSARY TO SECURE IT'S AVAILABILITY ?	

8. REFERENCES IN THE OPEN LITERATURE.

PLEASE TURN THE PAGE.

Table 3. An example of data base (location) documentation taken from NATO–AGARD Fluid Dynamics Panel Working Group 14

9. VALIDATIONAL ASPECTS.		10. TUNNEL CHARACTERISTICS.
11. ADDITIONAL DETAILS.		
SIGNATURE :	ORGANISATION :	DATE :

This Summary Sheet is designed to compile a broad range of information on wind tunnel test models where the data already measured may be of value in validating advanced CFD methods. These notes are to assist in completing this summary and the numbers refer to the individual sections of the summary. Please also consult the International Vortex Flow Experiment sheet included as a worked example.

1/. Model geometry can be given by listing its components (eg. wing, canard, pylon/store, body) with a brief description of the component geometry
eg. pylon/store type (eg. 2 drag-free pylons + axisym tank s), 65 deg delta wing with round l.e., pitot intake with variable mass-flow.
Details of the model's support or mounting system should be included.

2/. Details of flow conditions measured can be indicated by
- range of values, sweeps eg. Ma 0.60(0.05)0.95 or Cl=0.4, all Mach

3/. Details of measurement techniques can be indicated by brief description
eg. overall : force and moment
surface : Cp, skin friction, heat transfer, oil flow
flowfield : Schlieren, laser screen, PIV, HWA, LDA, pressure probes.

4/. Details on the original purpose of the tests should be given if possible.

5/. These physical flow features/structures unambiguously identified by the test measurements and/or flow visualisation
eg. vortex flow, shocks, separations, boundary layers, wakes
should be detailed, with an indication of whether mean and/or fluctuating flow quantities are known.

6/. The important flow regions closely investigated by the tests, such as wakes, shocks and boundary layers, should be identified, with an indication of the measurement techniques used in each region.

7/. Availability of the test data can be indicated by tick/YES or cross/NO. If the data is not yet within the public domain, the action that needs to be taken to determine it's possible release as part of this collaborative AGARD venture should be identified, including anticipated dates.

8/. References in the open literature should cover both the testing and its data and the usage of such data in CFD validations to date.

9/. Validational aspects should address the accuracy of tunnel measurements, for example
- are repeat measurements available?
- has the model or a scaled version been tested in another tunnel?
- has the same quantity been measured by different techniques?
- are tunnel and instrumentation calibrations available?

10/. Details on tunnel characteristics should include the tunnel's mode of operation, it's dimensions and the types of walls.

11/. Additional information which may be pertinent to this review can be appended to this sheet eg. General Arrangement diagram of the geometry, photograph, abstract of published report on data.

ON BEHALF OF THE AGARD FLUID DYNAMICS PANEL WG 14, THANK YOU FOR COMPLETING THIS SUMMARY.

Table 3. An example of data base (location) documentation taken from NATO–AGARD Fluid Dynamics Panel Working Group 14

(c) Model evaluation activities

Given the existence of satisfactory data bases from CEC-funded research and elsewhere, the question of the quality management of technical models may be developed. Techniques of model evaluation are not discussed here, only ways by which the quality of the models used might be improved.

4.7.1 Model evaluation working group

The CEC could assist in the establishment and funding of a Model Evaluation working group, within the context of Major Technological Hazards. Such a group would not necessarily undertake evaluation of models but could determine and implement strategies by which evaluation would come about. They might also ensure awareness of, collaboration with and, possibly, amalgamation with other groups undertaking similar tasks.

4.7.2 Model evaluation service

An organisation could be encouraged to provide a service where a submitted model might be tested against known and available data bases and the outcome of the activity documented.

It must be stressed that this is an activity which, to be effective and not misleading, requires careful planning and implementation. If it reduces to an *ad hoc* and arbitrary running of computer codes against data bases without a clear protocol and without the supporting weight of experts (from all interested parties), the effort and resources will be wasted. Experience has shown that this is all too frequently the outcome of such activities.

A possible candidate for this role would be the CEC Joint Research Centre at ISPRA.

The document might, in addition, indicate an independent assessor's view of the physics contained within the model. However this, or similar, judgemental situations seem fraught with difficulty. Maybe this is better left to the open literature; however, it should be noted that many models are never presented to the open literature. It is also not clear that peer review alone, through the open literature, is an adequate technique for the scientific review of operational models.

4.7.3 Model evaluation by user: checklist

A brief checklist might be developed ($\sim$ 2 pages) to include various aspects of any technical model used in reaching conclusions within, for example, a 'safety report', and could include:

name
developer
physical basis
where evidence of any peer review can be found
data bases against which it has been tested
where evidence of that testing can be found
range of applicability of model
where supporting evidence can be found
uncertainty on model prediction.

Such a checklist might be standardised and a regulator might suggest they be supplied with a 'safety report'. It would also be responsible to suggest that a regulator could supply similar documents for their own models.

Such a checklist might become a Standard to be provided with any model, i.e. a model developer might find himself obliged to provide these with any model supplied.

The advantage of such an approach is that it would cause all interested parties to consider an evaluation of the models they were using. It would also be clear to all that such an evaluation had been carried out.

4.7.4 Users groups

The encouragement and support of model users groups can assist in the communication of best available and appropriate models. This currently takes place through conferences and journal articles though exist users groups do exist and are centred around particular model packages, e.g. SAFETI, or are restricted to a limited group of (similar) users. Of course these are closed groups.

5. CONCLUSIONS

(i) Much use is made of technical models within the context of Major Technological Hazards.

(ii) Within the CEC technical models are extensively used as part of a 'safety report'.

(iii) The use of technical models is likely to increase.

(iv) Technical models are usually central to the conclusions of investigations of Major Technological Hazards.

(v) The outcome of investigations may differ sigr.ificantly due to

- (a) the models available for use
- (b) the expertise and training of the user l ·ing inadequate to ensure appropriate model selection, the correct im; ·ementation of a model and the correct interpretation of the model out ıt
- (c) the quality, that is the fitness-for-purpc ·, of the models.

(vi) There is little formalised evaluation of the que y of these models.

(vii) There is a demand by users for a more struct :d evaluation of the quality of models.

(viii) Thus decisions of substantial consequence may based, in part, on investigations for which there is inadequate concern abo ıe quality of elements of that investigation.

(ix) The CEC is an appropriate body to have a role :he study of the quality and the management of quality of technical models.

(x) The evaluation of the quality of technical model tails

- (a) an assurance of correct coding of algorith ; required and this is probably straightforward
- (b) a statistical model validation is required a ıis entails comparison with experimental data sets. Considerable care ·quired in performing such a validation, in particular in the determinat of an appropriate protocol to ensure unambiguous conclusions to be d n from the validation
- (c) a model assessment including a scientific re v and other less objective aspects is also required and should be giver ıal weight to a statistical model validation.

(xi) Actions which may lead to an improvement in m quality are presented in the report.

REFERENCES

Britter, R.E. and McQuaid, J. (1988) *Workbook on the Dispersion of Dense Gases.* UK Health and Safety Executive, Sheffield.

Contini, S., Amendola, A. and Ziomas, I. (1991) *Benchmark exercise on major hazard analysis. Vol.I: Description of the project, discussion of the results and conclusions.* CEC/JRC-ISPRA EUR 13386 EN.

Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans.* CBMMS-NSF-38, Society for Industrial and Applied Mathematics, Philadelphia.

Elderkin, C.E. and Kelly, G.N. (1990) *Proceedings of the CEC/USDOE Workshop on Uncertainty Analysis, Santa Fe, USA.* Prepared for US DOE and CEC/DG XIII-D-3 EUR-13044 EN.

Fox, D.G. (1981) Judging air quality model performance: A summary of the AMS Workshop on Dispersion Model Performance. *Bull. AMS* **62**, 599–609.

Fox, D.G. (1984) Uncertainty in air quality modelling: A summary of the AMS Workshop on Quantifying and Communicating Model Uncertainty. *Bull. AMS* **65**, 27–47.

Hanna, S.R. and Drivas, P.J. (1987) *Guidelines for Vapor Cloud Dispersion Modeling.* CCPS/AIChE.

Hanna, S.R., Strimaitis, D.G. and Chang, J.C. (1991) Evaluation of 14 hazardous gas models with ammonia and hydrogen fluoride field data. *J. Hazardous Mat.* **26**, 127–158.

Hanna, S.R., Chatwin, P., Chikhliwala, E., Londergan, R., Spicer, T. and Weil, J. (1991b) Results from model evaluation panel at International Workshop on Modeling and Mitigating the Consequences of Accidental Releases of Hazardous Materials, May 21, 1991, New Orleans. Submitted to Plant/Operations Progress, A.I.Ch.E.

Hill, M.D. (ed.) (1989) *Verification and Validation of NRPB Models for Calculating Rates of Radionuclide Transfer through the Environment.* NRPB-R223.

Mercer, A. (1988) Methods of validating models of dense gas dispersion: a review. In *Stably Stratified Flow and Dense Gas Dispersion* (ed. Puttock, J.S.), Clarendon Press, Oxford, pp.169–204.

Otway, H. and Amendola, A. (1989) Major hazard information policy in the European Community: Implications for the risk analysis. *Risk Analysis* 4, 9, pp.505–512.

Rouhiainen, V. (1990) *The Quality Assessment of Safety Analysis*. Technical Research Centre of Finland, Publication No.61.

Snyder, W.H. (1981) *Guidelines for fluid modelling of atmospheric diffusion*. Report No. EPA-600/8-81-0009, EPA, Research Triangle Park, NC.

Suokas, J. (1988) Evaluation of the quality of safety and risk analysis in the chemical industry. *Risk Analysis* 8, 4, 581–591.

USEPA (1984) *Interim Procedures for Evaluating Air Quality Models* (revised). Report No. EPA-450/4-84-033, OAQPS, USEPA, Research Triangle Park, NC.

VAMP (1990) *Progress Report No.2*. IAEA, Vienna.

Weil, J.C. (1985) Updating applied diffusion models. *J. Climate Appl. Meteor.* 24, 1111–1130.

EXAMPLES OF MODEL EVALUATION

Two recent examples of model evaluation are presented. The first is an evaluation of hazardous gas models by Hanna and colleagues for the US Air Force and the American Petroleum Institute.

The second is an outline of topics included in an evaluation by Britter.

1. In the study by Hanna and colleagues some fifteen hazardous gas models were evaluated using data from eight field experiments. The paper, 'Hazardous Gas Model Evaluation with Field Observations' by S.R. Hanna, J.C. Chang and D.G. Strimitis (*Atmosphere Environment*, 27A, 15, (1993), pp.2265–2285) is a recent example of model evaluation.

The evaluation is essentially a validation exercise and does not provide any scientific assessment or model verification. Nor are any comments made on the user-friendliness etc. of the models being evaluated.

However, the validation exercise has been thorough and comprehensive. It includes:

(i) a summary of the characteristics of the data sets;

(ii) a summary of the characteristics of the models;

(iii) the creation of a modeller's data archive;

(iv) explicit specification of the parameters to be evaluated;

(v) explicit specification of model evaluation statistics;

(vi) a clear presentation of the evaluation statistics;

(vii) model sensitivity studies by the analysis of residuals.

An important omission from the study is an explicit statement of the limitations of the experimental data used for the evaluation. As a consequence, the results of the evaluation are widely quoted to support the use of some models well outside their region of appicability. For example the evaluation can provide no information on the usefulness of any of the models in realistic accident scenarios involving buildings, complex source structures or topography.

Table 1. Summary of characteristics of the data sets

	Burro	Coyote	Desert Tortoise	Goldfish	Hanford Kr^{85} (continuous)	Maplin Sands	Prairie Grass	Thorney Island (instantaneous)	Thorney Island (continuous)
Number of trials	8	3	4	3	5	4, 8	44	9	2
Material	LNG	LNG	NH_3	HF	K^{85}	LNG, LPG	SO_2	Freon and N_2	Freon and N_2
Type of release	Boiling liquid (dense gas)	Boiling liquid (dense gas)	2-phase jet (dense gas)	2-phase jet (dense gas)	Gas (non-buoyant)	Boiling liquid (dense gas)	Gas jet (non-buoyant)	Gas (dense gas)	Gas (dense gas)
Total mass (kg)	10,700–17,300	6500–12,700	10,000–36,800	3500–3800	11–24*	LNG: 2000–6600 LPG: 1000–3800	23–63	3150–8700	4800
Duration (s)	79–190	65–98	126–381	125–360	598–1191	60–360	600	Instantaneous	460
Surface	Water	Water	Soil	Soil	Soil	Water	Soil	Soil	Soil
Roughness (m)	0.0002	0.0002	0.003	0.003	0.03	0.0003	0.006	0.005–0.018	0.01
Stability class	C–E	C–D	D–E	D	C–E	D	A–F	D–F	E–F
Max. distance (m)	140–800	300–400	800†	3000	800	400–650	800	500–580	472
Min averaging time (s)	1	1	1	66.6–88.3	38.4	3	Dosage	0.06	30
Max. averaging time (s)	40–140	50–90	80–300	66.6–88.3	270–845	3	600	0.06	30
Reference	Koopman *et al.* (1982)	Goldwire *et al.* (1983)	Goldwire *et al.* (1985)	Blewitt *et al.* (1987)	Nickola *et al.* (1970)	Puttock *et al.* (1984)	Barad (1958)	McQuaid and Roebuck (1985)	McQuaid and Roebuck (1985)

* Curies, rather than kg, are used as a measure of the amount of this radioactive tracer released.

† Concentrations are measured beyond 800 m, but there are not well-instrumented measurement arcs.

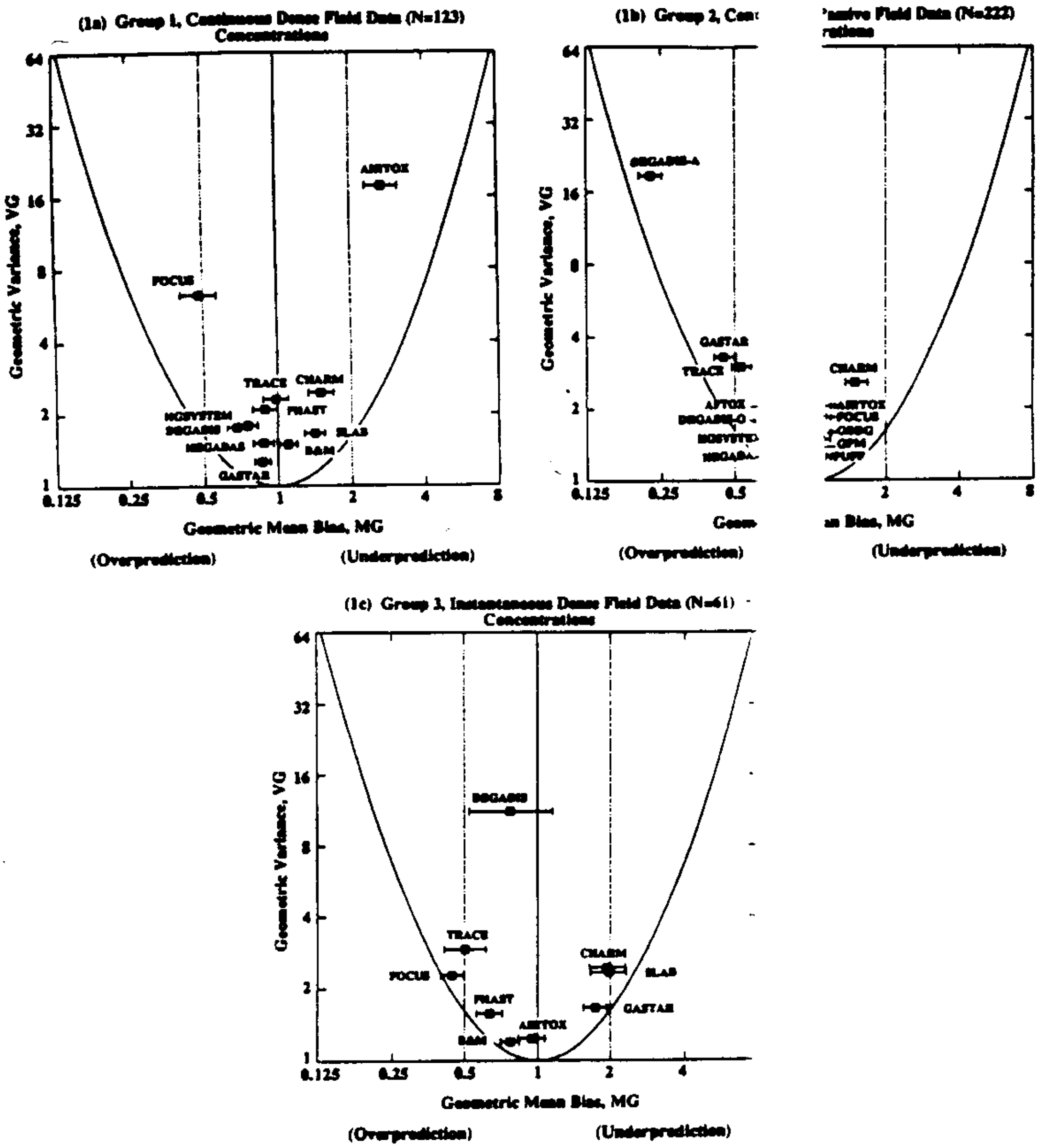


Fig. 1. Model performance measures, geometric mean bias $MG = \exp\ \overline{(\ln C_o - \ln C_p)}$ anc
$[\overline{(\ln C_o - \ln C_p)^2}]$, for maximum plume centerline concentration predictions and observatio
MG are indicated by the horizontal lines. The solid parabola is the "minimum VG" c
represent "factor of two" agreement between mean predictions and observations. (a) Grou
sets (Burro, Coyote, Desert Tortoise, Goldfish, Maplin Sands and Thorney Island), invol
points for the shortest available instrument averaging times. (b) Group 2—continuous pa
and Hanford), involving a total of 49 trials and 222 points. (c) Group 3—the instantanec
Island), involving a total of 9 trials and 61 points.

tric variance $VG = \exp$
confidence intervals on
he vertical dotted lines
ntinuous dense gas data
tal of 32 trials and 123
data sets (Prairie Grass
e gas data set (Thorney

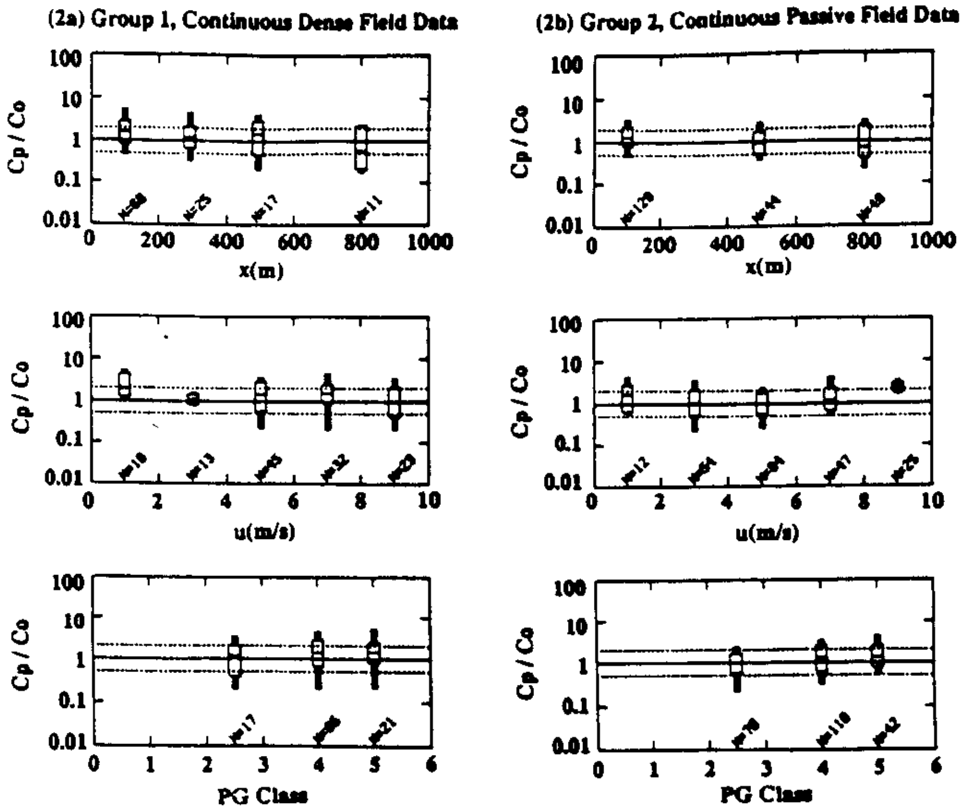


Fig. 2. Distributions of model residuals, C_p/C_o, for HGSYSTEM for maximum concentrations on the plume centerline. The "box plot" format indicates the 2nd, 16th, 50th, 84th and 98th percentiles of the cumulative distribution function of the N points in the box. (a) Group 1—continuous dense gas data sets (Burro, Coyote, Desert Tortoise, Goldfish, Maplin Sands and Thorney Island). (b) Group 2—the Prairie Grass continuous passive gas data set.

2. The contents page of the second evaluation exercise (called assessment) by Britter is presented here to emphasize other aspects of evaluation: the scientific assessment and the model verification. These aspects will be discussed during the presentation.

ASSESSMENT OF
HGSYSTEM/PLUME, PGPLUME

by

Dr R.E. Britter

Cambridge, England

For

November 1993

CONTENTS

FUTURE EVALUATION ACTIVITIES

Commission of the European Communities

MODEL EVALUATION GROUP

MEG

MEG is a European initiative on evaluation of technical models used within the major industrial hazards area and is supported by the CEC DGXII, Directorate-General for Science, Research and Development.

Background

Technical models are used in a number of areas of industrial hazards assessment. It is becomming more and more apparent that most of these models have never been through a procedure of evaluation, but nonetheless are used to assist in making decisions which may directly effect the safety of the public and the environment. As a major funder of European research on Major Industrial Hazards, DGXII is conscious of the importance in ensuring that model development is of a standard which is commensurate with the importance of model use.

A meeting which attracted 35 persons from industry, authorities and research, and with a good European coverage, agreed to establish a Model Evaluation Group and recommended that an initiative be supported by the CEC.

Objectives

The objectives of a Model Evaluation Group are twofold. The primary aim is to improve the culture in which models are developed and used and so ensure that technical models used in all aspects of major hazard evaluation are up-to-date with technical developments and utilised by personnel well-versed in their applicability and functioning.

Secondly, in the course of the work the results are applicable in assisting the CEC in establishing a balanced set of research priorities for its research programmes within Major Industrial Hazards.

Organization

The CEC's advisory body Safety Management and Hazard Assessment Research Cooperation in Europe (SHARE) acts as the steering committee for the initiative.

The CEC has established the MEG to manage the activities. MEG consists of:

K.E. Petersen, Risø, Denmark (chairman)
B. Stork, TNO, The Netherlands
S.J. Jones, SRD, United Kingdom
T. Cartage, SOLVAY, Belgium
R. Britter, Cambridge University, United Kingdom
M. Schatzmann, University of Hamburg, Germany.

The model evaluation activities will be carried out by involvement of European experts within the various scientific fields of major industrial hazards.

Activities

MEG will initiate the following tasks:

1. Classification of available models in major industrial hazards.
2. Development of a general evaluation protocol for evaluation of technical models in major industrial hazards.
3. Development of a guidelines for model developers.
4. Definition of working groups and working group tasks for selected classes of models.
5. Status on model evaluation within the subject area of working groups.
6. Specification of lacking data/model deficiencies within the subject area of working groups.
7. Identify research needs.
8. Open European seminar where results are presented.

Participation

The work in the working groups is open to European researchers and users of technical models used in the major industrial hazards area. The CEC/DGXII is supporting the work covering travel and subsistance for those nominated to perform the tasks.

Information
can be obtained from

Steve Cole
Commission of the European Communities
DGXII/D/1
Rue de la Loi 200
B-1049 Brussels
Belgium

Phone: +32-2-295 0347
FAX: +32-2-296 3024

MODEL EVALUATION GROUP

Steve Cole
CEC
DG XII/D/1 SDM 3/49
Rue de la Loi 200
B-1049 Brussels
Belgium
Tlf. +32-2 295 03 47
Fax. +32-2 296 30 24

Rex Britter
Department of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ
UK
Tlf. + 44 223 332643
Fax. + 44 223 332662

Ben Stork
TNO
Department of Fluid Dynamics
Laan van Westenenk 501
P.O. Box 342
NL-7300 AH Apeldoorn
The Netherlands
Tlf. + 31 55 49 34 93 / 961
Fax. + 31 55 41 98 37

Kurt E. Petersen
System Analysis Department
P.O. Box 49
DK-4000 Roskilde
Danmark
Tlf. +45-42371212/3082
Fax +45-46757101

Steve J. Jones
SRD
Wigshaw Lane
Culcheth Cheshire WA3 4NE
UK
Tlf. + 44-925-254315
Fax. + 44-925-254537

Thierry Cartage
SOLVAY S.A.
Direction Centrale Technique
Rue de Ransbeek 310
B-1120 Brussels
Belgium
Tlf. 32-2 264 29 83
Fax. 32-2 264 32 75

Michael Schatzmann
University of Hamburg
Bundesstrasse 55
D-2000 Hamburg
Germany
Tlf. +49-40 4123-5090
Fax +40-40 4117-3350