DRAFT

July 4, 1994

MODEL VALIDATION FOR PREDICTIVE EXPOSURE ASSESSMENTS

M B Beck *
Lee A. Mulkey **
Thomas O. Barnwell **

* Warnell School of Forest Resources
University of Georgia
Athens, Georgia 30602-2152
and
Department of Civil Engineering
Imperial College
London SW7 2BU, UK

** U.S. Environmental Protection Agency
Environmental Research Laboratory
Athens, Georgia

CONTENTS

MODEL VALIDATION
FOR PREDICTIVE EXPOSURE ASSESSMENTS

1 INTRODUCTION AND BACKGROUND

The construction and use of mathematical models are essential in predicting the possible consequences of releasing chemicals, some of which may be quite novel, into new environments. Substantial costs, and substantial damages to human health and the environment, may attach to the regulatory decisions that are informed and thus guided by the predictions derived from a model. The risk of making a wrong decision will be strongly dependent on the reliability of these predictions, in just the same way as it would be dependent upon the reliability of an observing instrument used for a site survey prior to a construction project. There is a profound concern, therefore, with the need to establish the validity of a given model in performing a specified task, usually of making predictions of future behavior.

The difficulty of meeting this need of model validation tends to increase as the degree of extrapolation from observed conditions in the past increases. And not surprisingly, the greater the is degree of extrapolation so the greater the necessity of relying on a model for the conduct of an assessment.

The difficulty of quantifying a model's validity also increases with the complexity and size of the model. The peer group of analysts capable of scrutinising the composition of a more complex and comprehensive model will tend to be smaller; and the possibility of significant uncertainty (or error) attaching to the model's many constituent parameters (coefficients) will increase. Such uncertainty may well lead to ambiguity in the predictions made by the model about the nature of a system's behavior in the future (Beck and Halfon, 1991). Many combinations of values for the parameters may give equally plausible matches of the model's behavior with the historical observations, but with no clear indication of which of these combinations should be used for making predictions. Yet in the absence of empirical observations of past behavior, which indeed in many instances are by definition not possible, there is a natural tendency to rely on the complexity of a model as a form of insurance against the unknown. For if everything of conceivable relevance has been included in a model, how can its predictions possibly be wrong? Or perhaps conversely, there is a tendency to avoid the use of simpler models, which may make no reference to the internal mechanisms believed to govern the behavior of the system.

In spite of much attention over the years, most notably in the field of geologic repositories for the disposal of high-level nuclear waste (Davis et al, 1990), the problem of model validation seems to remain as intractable as ever. From time to time frustration with this intractability rises to the fore, most recently in the literature on modelling the movement of contaminants through groundwater systems (Konikow and Bredehoeft, 1992; Oreskes et al, 1994). Indeed, it is as though this intractability is reflected in the many labels that have been assigned to what is meant by the process of "model validation", without yet a definitive procedure having emerged for its implementation.

The purpose of this document is to define a set of procedures for model validation in carrying out exposure assessments. These procedures are based both on the concept of peer-group review and evaluation and on quantitative (usually statistical) measures of validity. They complement the EPA Guidelines for Exposure Assessment (EPA, 1991).

1.1 **Requirements for Model Validation Within EPA**

Assessments that estimate, in part, the exposures that motivate and assist regulatory or policy decisions within EPA are often challenged to demonstrate their "scientific validity". Predictive exposure assessment modeling is increasingly common as modeling technology advances and is essential for many Agency risk assessments. Inevitably, and appropriately, model validation is a major concern among both the exposure assessment community and Agency officials that factor such assessments into management or policy decisions. The issue, and problems, of model validation are widely recognized, are active research areas among almost all scientific disciplines, and are specifically targeted as vital in environmental predictions used to assist risk assessments. Perhaps the most recent external expression of needs for model validation within EPA can be found in **"Science and Judgement in Risk Assessment"** (NRC,1994). Interestingly, the authors of this report avoid the term "model validation" per se in favor of the term "model evaluation". In describing the use of air-quality models, for example, "Evaluation of the air-quality models and other components of air-pollutant risk assessment is intended to determine accuracy for providing the details required in a given application and to provide confidence in the results". Absent detailed knowledge of the deliberation leading to the choice of these words, one reasonable interpretation is that "validation" defies a concise definition and further that model accuracy and one's confidence in modeling results are closely related to problem specific situations. The present report largely comports with these interpretations although an operational definition of model validation is a major objective.

Standard E 978 - 84 of the American Society for Testing and Materials (ASTM) sets out standard practice for **"Evaluating Environmental Fate Models of Chemicals"** (ASTM, 1984). Its purpose is to provide "procedures and criteria for development, deployment, and use of mathematical models ... in predictive risk assessments". It is thus highly relevant to the present note and extensive reference to its terminology, in particular, will be made below. The scope of the standard is further qualified thus (ASTM, 1984):

> It does not specify models themselves, but it establishes minimum criteria for distinguishing acceptable models from those which may be incomplete, untested, or inappropriate for intended purposes.

Again, this is of obvious, central relevance to the present discussion.

More recently (in January, 1989) the Environmental Engineering Committee (EEC) of the EPA's Science Advisory Board prepared a **Resolution on Use of Mathematical Models by EPA for Regulatory Assessment and Decision-Making** (US EPA, 1989). Among other items, this Resolution states that: "[t]here is a need for models used in regulatory applications to be confirmed with laboratory and field data" (US EPA, 1989). This is amplified by a more complete statement, a key part of which is the following:

> The stepwise procedure of checking the numerical consistency of a model, followed by field calibration, validation and *a posteriori* evaluation should be an established protocol for environmental quality models in all media ....

The EEC was concerned, in particular, that there should be a consistency of approach across the Agency in demonstrating the validity of a model.

These same themes, of both the ASTM standard and the EEC Resolution, are reiterated in the National Research Council's (NRC) authoritative publication on **"Ground Water Models: Scientific and Regulatory Applications"** (NRC, 1989). Looking towards the future, the document concluded that government agencies and private industry should be aware of the need for, and benefits of, additional research in, inter alia, model validation.

The EPA's current **Guidelines on Exposure Assessment** express the role of models and their validation thus (EPA, 1991):

> Environmental fate models calculate estimated concentrations in media, that in turn are linked to the concentrations at the point of contact. The use of estimated properties or rates adds to the uncertainty in the exposure concentration estimate. When assessors use these methods to estimate exposures, uncertainties attributable to the model and the validation status of the model must be clearly discussed in the uncertainty section ....

The EPA's **Agency Task Force on Environmental Regulatory Modeling Final Report** (EPA, 1993), described, inter alia, model use acceptability

criteria as part of its goal to further the use and value of environmental modeling in EPA's regulatory programs. The Task Force analysis does not place an emphasis on validation.
Clearly, a carefully expressed protocol for model validation in predictive exposure assessments is now timely. Given the general imperative for model validation, the current and varied set of definitions and concepts, and the authors' perspective on predictive exposure assessment modeling within EPA, the problem remains to describe model validation more precisely and operationally in a way that can be implemented as steps in the exposure assessment process. As will be developed and elaborated in later sections, the focus of this effort is sharpened by two important constraints:

(i) The validity of a model cannot be established without specification of the task the model is required to perform. That is, model validation is essentially problem specific and at the present time no model-specific or universally applicable model validation process exists.

(ii) The greater concern for model validation lies with the use of models in the generic screening process (in assessing scenarios for a wide array of situations that could occur), and in other "data poor" situations rather than in site-specific cases where local data are available or can (will) be collected. This is not to say that model validation expectations for intensive, site-specific modeling should be relaxed. Rather, it is to say that there exists a reasonably well-developed history of model validation in the site-specific context and that the greater need lies in cases for which application of the classical approaches is not an option.

1.2 **Organization of the Document**

Model validation is not an easy concept to define in precise terms. Section 2 reviews previous definitions of the problem and proceeds to develop a definition of particular relevance to predictive exposure assessments. Section 3 presents in outline the methods and procedural steps of model validation and defines the role of validation in the overall process of developing a model. It is not intended that this document should serve as a detailed instructional guide for model validation, although Section 3 is accompanied by an Appendix providing some supporting mathematical material. Section 4 discusses the significant role of expert opinion and qualitative judgement in determining the validation status of a model. Finally, Section 5 sets out the forms of evidence that will be necessary in implementing a protocol for judging whether a model can be said to have been validated.

2 **CONCEPTS AND DEFINITIONS**

The primary difficulty in setting out a protocol -- or set of procedures -- for model validation is the great variety of meanings that has been attached to the word "validation" itself (Versar Inc, 1988; Donigian and Rao, 1990). The noticeable use of other terms for the process by which one gains confidence in modeling applications found in recent discussions of modeling in EPA is testimony to this problem.

Within the terms of the ASTM standard E 978 - 84, validation is defined as:

> Comparison of model results with numerical data independently derived from experience or observations of the environment.

This is rather restrictive, since it places a strong emphasis on reference to "observations of the environment", which will not be available for novel chemicals supposed to be moving through previously unencountered environments (at least for the given chemicals under consideration). Nevertheless, the word "experience" in the above, with its undertones of subjective, expert knowledge, opens up the potential for a significantly broader interpretation, as will become apparent.

In essence, however, the above definition is "retrospective" in its outlook, since there can only be observation of the **past** behaviour of the system. It is not radically different in intent from the more contemporary preferences of Konikow and Bredehoeft (1992), who state:

> What is usually done in testing the predictive capability of a model is best characterized as calibration or history matching: it is only a limited demonstration of the reliability of the model. We believe the terms *validation* and *verification* have little or no place in ground-water science; these terms lead to a false impression of model capability. More meaningful descriptors of the process include *model testing, model evaluation, model calibration, sensitivity testing, benchmarking, history matching*, and *parameter estimation*. Use of these terms will help to shift emphasis towards understanding complex hydrogeological systems and away from building false confidence into the model predictions.

The plethora of phrases used collectively to describe the process of validation is immediately apparent. It is as though the inability to demonstrate unequivocally that validation (or invalidation) has been completed has forced upon us a groping for better words to describe what can be done (which is yet not quite what needs to be done). Hassanizadeh and Carrera (1992) have expressed much the same sentiments:

> The difference between the different definitions [of validation] is not a matter of semantics; it is a question of the perception behind validation. The general concern

> about a proper definition of [the] aim and scope of validation is a legitimate one, given the many misconceptions about 'validated models'. Many individuals do not realize that a 'validated model' does not necessarily yield accurate predictions of reality even if it does so once. A theory which has overcome many tests is not ensured of not failing in the next one; theories can be proven wrong, but they cannot be proven right. Thus validation must not be considered as obtaining a label; one should not seek a 'yes or no' answer to model validation.

Further support for the inappropriateness of a simple binary judgement on the validity of a model can be found in the evaluation of models of air quality, notably those for predicting regional acidic depositions (NAPAP, 1990).

Another alternative definition, as follows, can be found in the report by Versar Inc (1988):

> In a general sense, the [validation] of a [model package] refers to the overall process of defining the range of circumstances or situations for which the package's behavior and predictions are satisfactory. However, from the point of view of a user of a potential model package, the diversity of conditions for which the package has been shown to be valid is irrelevant. The potential user is interested only in whether the model will correctly predict system response for the situation of interest to him/her.

This is restated in a glossary of terms provided in the same report (Versar Inc, 1988):

> [Validation is] [t]he process of defining the range of problems or situations for which the behavior and predictions of a [model package] are satisfactory; the iterative expansion of the known applicable range of a [model package] by documenting new site-specific [evaluations] of its performance. A model may be adequately validated for use in a particular situation. However, it is not possible to state whether a model will be valid for all possible situations.

We are, in fact, **attempting the impossible**, for it will be apparent that any statement about validation necessitates extrapolation from past experience. It is highly unlikely that the next purpose for which the model is to be used will be identical in all respects to one or more of the contexts in which it has been applied in the past. Unless one believes that history repeats itself -- exactly -- the task of prediction must almost always embody some element that is not only unknown but also unknowable as seen from the present. That a literal interpretation of the task of validation would be an impossible task has been clearly stated by Oreskes et al (1994), whose contribution to this particular (philosophical) aspect of the debate is likely to stand as the closing statement for some considerable time to come.

A slightly different definition of model validation is therefore required for predictive exposure assessments; and this in turn requires examination of the philosophical foundations of what "validation" is believed (and agreed) to be about.

2.1 **Philosophical and Conceptual Basis**

In a wide-ranging review across several fields of modelling Lewandowski (1982) has made the following salutory observations:

> It is commonly agreed between modeling methodologists that model validation is one of the most important stages in the model building process. ... However, at the present stage of research there are almost no suggestions concerning concrete methods of validation. Practically all authors only discuss definitions of validation - not methods. The number of papers dealing with methods of model validation is also rather limited.
>
> The reason for this gap between methodological consciousness and the practice of model building seems to be obvious - the discussion stays at too high a level of abstraction. In general, all authors consider "model" as a description of reality, and ... it is only possible to generate rather general statements, frequently true but without operational meaning.

The following consideration of the philosophical underpinnings of model validation will seek to avoid the difficulties implied by Lewandowski's observations.

2.1.1 **Caswell's contribution**

Caswell's study of the problem of validation (in 1976) is generally recognized as one of the first, and most important, treatments of the subject (Caswell, 1976). Burns, who has himself made a significant contribution to the subject (in his paper on "Validation of Exposure Models: the Role of Conceptual Verification, Sensitivity Analysis, and Alternative Hypotheses"; Burns, 1983), expressly acknowledges his indebtedness to Caswell.

The key point Caswell makes in his paper is that a judgement about the validity of a model cannot be made in the absence of a specified **purpose** for the model. He identifies two such purposes:

(i) the use of models for the development of insights into, and understanding of, the fundamental mechanisms underlying the behavior of a system;

(ii) their use for the prediction of future behavior, or of behavior under conditions not previously encountered.

Other purposes can of course be specified, in particular, the assessment of compliance with, or violation of, a regulatory standard. Most authors, like Caswell, have attached great importance to the purpose of the model in establishing its validity.

Increasing understanding of the fundamental mechanisms of behavior will **not** be the primary purpose of a model in predictive exposure assessments, but it is by no means entirely irrelevant to this more practically oriented objective. For whatever protocol is specified for the validation of a model to be used in a regulatory context, it would be unacceptable for such a protocol to discourage improvements in basic understanding.

Put simply, a model may be viewed as a complex assembly of several, if not many, constituent hypotheses, and in this respect assessment of the strengths and weaknesses of each **constituent** hypothesis is as important as the more familiar problem of examining the validity of the model as a **whole**. Precisely how one goes about this task of corroborating and refuting the constituent hypotheses of the model is now widely agreed, in philosophical terms, to have been amply and adequately stated in Popper's papers on the scientific method and the evolution of knowledge (Popper, 1959, 1963; as illustrated by Caswell, 1976; Young, 1978; and Reckhow and Chapra, 1983; Konikow and Bredehoeft, 1992; Ababou et al, 1992). This does not imply, however, that application of these philosophical principles to the interpretation of field data is a fully resolved matter; it is not (Beck, 1987). Moreover, the pursuit of improved understanding is self-evidently an unending quest. Yet practical decisions, in sharp contrast, cannot be deferred for ever.

Of special relevance to the more pragmatic and -- for this discussion -- the principal of the two purposes of model validation, is the following quotation from Caswell (1976, p317):

> Models of systems are systems themselves, the interacting components of which are mathematical variables and expressions. They are, moreover, man-made systems .... The construction of such artificial systems is a design problem, and the process of design is in essence a search for agreement between properties of the artificial system and a set of demands placed on it by the designer. It is impossible to evaluate the success or failure of a design attempt without specification of these demands, the task environment in which the artificial system is to operate. Validation of a model is precisely such an evaluation ....

When improved understanding is not the objective, Caswell argues that the intrinsic truth, or realism, of the model is not of interest.

Pragmatically, one can get by with various approximations, which are very probably false or unrealistic, as long as one is aware of the domain of applicability of the model, i.e. the areas in which it is known to work successfully. To illustrate this Caswell quotes a model of the global human population. The model gives a seemingly good prediction of this quantity in the year 2000, yet also predicts a population infinite in number shortly after 2026. The problem becomes one of establishing the boundaries of the domain of the model's applicability. Ultimately, Caswell reduces this latter to tests of the degree to which there is coincidence between the outputs of the model and a set of observations (on which basis, incidentally, the population model quoted above performs very well). Thus, somewhat unsatisfactorily, specification of a practical test of the model's validity has been returned to the rather restrictive definition requiring observations of actual behavior.

The significant point, however, is the view that validation is essentially a problem of **design**.

2.1.2 **Validation as a design task**

For the situation where the analysis of actual in situ observations is not possible, as, for example, in a screening-level analysis in exposure assessment, we are seeking to design an instrument (the model) that will meet certain specifications. Like Caswell's example of the model for forecasting the world's human population, there are situations in predictive exposure assessments where -- out of a constructive pragmatism -- one would accept the use of a model that is known, a priori, **not** to be wholly valid scientifically.

When a novel substance is proposed for release into the environment it is possible to assume that it will experience no "forces" acting on its distribution and presence in the receiving environment other than those of transport, i.e., advection and dispersion. The substance is assumed to be entirely "conservative". If the predicted exposure is found to be acceptable, under this strong assumption with its inherent maximum degree of safety, no further argument about the decision to release the substance should be necessary. If it is conversely unacceptable, then the proponent of the substance's manufacture and release is under the constructive obligation to furnish more detailed information about this substance's chemical or microbial degradability in the given environment.

In this example, the conservative model has proved to be a valid instrument in fulfilling the tasks of:

(i) establishing the risk of exceeding a given level of acceptable exposure;

(ii) identifying the need for more detailed information.

Such tasks are entirely consistent with the roles defined for models in the general field of exposure assessment, as stated by the EPA's Guidelines (EPA, 1991):

> A primary consideration in selecting a model is whether to perform a screening study or to perform a detailed study.
>
> The value of the screening-level analysis is that it is simple to perform and may indicate that no significant contamination problem exists. Screening-level models are frequently used to get a first approximation of the concentrations that may be present. Often these models use very conservative assumptions; that is, they tend to over-predict concentrations or exposures. If the results of a conservative screening procedure indicate that predicted concentrations or exposures are less than some predetermined "no concern" level, then a more detailed analysis is probably not necessary. If the screening estimates are above that level, refinement of the assumptions or a more sophisticated model are necessary in further iterations for a more realistic estimate.
>
> Screening-level models also help the user conceptualize the physical system, identify important processes, and locate available data. The assumptions used in the preliminary analysis should represent conservative conditions, such that the predicted results over-estimate potential conditions limiting false negatives. If the limited field measurements or screening analyses indicate that a contamination problem may exist, then a detailed modeling study may be useful.

Similarly, in **"Science and Judgement in Risk Assessment"** (NRC,1994), a recommendation is made with respect to one particular model that..."The underlying assumption that the calculated exposure estimate is a conservative one should be reaffirmed; if not, alternative models whose performance has been demonstrated to be superior should be used in exposure assessment". Further the report concludes that ..." EPA should particularly ensure that, although exposure estimates are as accurate as possible, the exposure to the surrounding population is not underestimated". These statements expressly identify a model design that not only accommodates the existence of bias but encourages it as a means to accomplish certain tasks. In this case a model valid for the task specification is knowingly wrong in at least some scientific dimension.

The task of the models in the above example and citation was **not** to provide as faithful a prediction as possible of the "true" behavior of the substance if released into the environment. The model's design, as an instrument of prediction, should be capable of successive refinement and adaptation against this task specification.

2.1.3 Composition and performance of the model

The way in which the validity of a model is judged is intuitively based on two features:

(i) The **composition** of the model, i.e. the manner in which its constituent hypotheses are assembled, with then some measure of the consensus (or disagreement) that attaches either to each constituent hypothesis, or to the model as a whole, or both. This can be regarded as an essentially **internal** measure of validity; judgement about the model is being made by reference to its intrinsic mechanisms, which determine how the input (causative) stimuli are related to the output responses. In principle, any such judgement ought to reflect the **generic** properties of the model, irrespective of the current task to which it has been assigned. In reality, however, it must inevitably reflect the accumulating experience of the model (and its earlier successful/unsuccessful performance) up to, but not including, the present task (whatever this may be).

(ii) The **performance** of the model in terms of being a valid instrument for undertaking the current task assigned to it. In contrast, this can be regarded as an essentially **external** measure of validity, in the sense that it will engage some comparison of data derived from the model with data (or conditions) deduced from sources of knowledge utterly independent of the specific model whose validity is to be established. Judgement in this case is being made by reference to a set of required output responses, in which terms the current task will most usually be cast. The independent "sources of knowledge" used to define the task may themselves nevertheless be in the form of alternative (competing, candidate) models. In principle, and in practice, performance validity is a **task-specific** property and will be of general relevance only inasmuch as the specific task in fact encompasses features common to all predictive exposure assessments.

The less familiar concept of compositional validity, or internal validity, requires further elaboration.

Under one label or another, it has in fact been in use for some time (Mihram, 1973; Miller et al, 1976). The report by Versar Inc (1988) defines "face validity" as follows:

> Using subjective opinions regarding the surface, or initial, impression of the model's realism.

Hermann (1967), in an early paper on gaming models for the simulation of international politics, has used the same phrase (face validity) to describe a situation in which the model structure is explained to various experts, and judged by them to be intuitively reasonable. For the present discussion **compositional validity** (internal validity) will be taken to approximate closely this earlier concept of "face validity".

There remains the question of how compositional validity is to be gauged. Given the likelihood that the composition of a model **as a whole** will evolve over the years (and differ from one developer to another), whereas the mathematical expression of the model's many **constituent** hypotheses is rather less likely to change rapidly, the natural preference would be for compositional validity to be measured on the more consistent basis of constituent hypotheses. Such a preference is effectively expressed in the numerical results of the work of EPA's Exposure Modeling Work Group, as reported in Donigian and Rao (1990). The presence, absence, or modification of each constituent hypothesis will also best reflect the evolution of the model over the years. However, it would still be desirable to have a procedure for aggregating these constituent measures into some overall "value" for the compositional validity of the model as a whole.

An obvious quantitative choice for expressing the validity of constituent hypotheses would be the uncertainty attaching to each of the model's parameter (coefficient) estimates. Corroborating evidence from past applications will (in a Bayesian sense) "narrow" the bands of uncertainty attaching to each parameter, and evidence of refutation will effect the reverse (if not occasion the restructuring of the model). In this same spirit, both from a statistical (Reckhow et al, 1990) and a philosophical perspective (Oreskes et al, 1994), compositional validity may be closely associated with the notion of model **confirmation**, qualified as follows (Oreskes et al, 1994):

> The greater the number and diversity of confirming observations, the more probable it is that the conceptualization embodied in the model is not flawed. But confirming observations do not demonstrate the veracity of a model or hypothesis, they only support its probability.

Thus, in Figure 1(a) the compositional validity of Model (1) might be determined at the outset from subjective (expert) introspection and experience; that of Model (2) on the basis of both this initial experience and interpretation of Data Set (1); that of Model (3) on the basis of interpreting both Data Sets (1) and (2) and, now less so, the initial subjective judgement; and so on.

Fundamentally, however, judgement about compositional validity would seem to be well suited to some process of peer group review (as discussed in Section 4).

2.1.4 **Prior and posterior performance validity**

Let us assume that performance validity can be gauged in quantitative terms by the residual errors of mismatch between the model's output responses and the task specification (or observed data set, if available). This is generally the most familiar definition of model validity.

As with the discussion of compositional validity, subjective introspection and experience must be used to make any judgement about the performance validity of the model at the outset, should this be so required (Figure 1(b)). Thereafter, a comparison between the outputs of Model (1) and the observed Data Set (1) may be used for determination of the performance validity. In the event that any properties of Model (1) are adjusted on the basis of this comparison -- such as typically the values of the model coefficients (or parameters) during calibration -- and the errors of mismatch recomputed, we may refer to this more refined concept as **posterior** performance validity. Here "posterior" signifies **after** some preceding interpretation of the reasons for the mismatch of the model's responses with the given data set and based on an adjusted set of residual errors of mismatch. Where no such adjustments are made, as in the cases of Data Sets (2) and (3) in Figure 1(b), we may refer to the complementary concept of **prior** performance validity, signalling thus the qualifying state of being **before** any interpretation and adjustments in the sense intended above.

As will be discussed in Section 3, it is the property of prior performance validity that is especially important in predictive exposure assessments. Indeed, it may be argued that no judgement about a model's validity should be based on a posterior performance statistic or, at the very least, such judgement should necessarily be declared as inferior to what would ideally be required.

2.1.5 **Behavior under novel conditions**

There is a paradox. The greater the degree of extrapolation from past conditions, so the greater must be the reliance on a model as the instrument of prediction; hence, the greater the desirability of being able to quantify the validity (or reliability) of the model, yet the greater is the degree of difficulty in doing just this.

We are strongly accustomed to the idea of performance being specified in terms of a time-series of observations of the model's state (or output) variables. This is a highly restrictive outlook, however, as admirably demonstrated in the seminal work of Hornberger, Spear, and Young (Young et al, 1978; Hornberger and Spear, 1980; Spear and Hornberger, 1980). Behavior, i.e. performance, can be specified in a variety of other ways, including on the basis of expert opinion. In fact, once this notion is cast aside -- of judging a model's performance on the basis of historical observations **alone** -- ways round the difficulties of validating a model for performance under novel conditions become apparent.

The analyst has immense freedom to be creative in defining the task, or

purpose, of the model. For example, this might be as follows:

(i) To fit the historical data as closely as possible (the traditional purpose of calibration);

(ii) To fit a second set of historical data as closely as possible, without altering the model's parameter values (a traditional definition of model validation);

(iii) To locate a sample of randomly generated values for the model's parameters that enable the model outputs to match certain crude constraints on what is **defined** (not actually observed) to be an acceptable statement of past behavior (the task addressed by Young et al, 1978);

(iv) To locate a sample of randomly generated values for the model's parameters that enable the model outputs to match certain crude constraints on what is defined to be radically different behavior of the system in the future (the task addressed in Beck, 1991);

(v) To locate a sample of randomly generated values for the model's parameters that result in an exposure above or below a given level (including "no concern", extreme or "high-end" exposures) or within a given confidence band around a specified probability of occurrence (the primary task of exposure assessments).

Of these, (iii), (iv), and (v) reflect evaluation of the model's performance back from the external definition of the task onto the internal composition of the model. For in essence the Hornberger-Spear-Young (HSY) algorithm (as discussed in Beck (1987), for example) is the identification of those model parameters that are crucial to discriminating a match from a mismatch of the model's outputs with the reference behavior, and (by reflection) those parameters that are redundant to this discriminating function.

The questions of interest may therefore become:

(i) What is it about the model, i.e., which constituent parameter(s) is it, that enables the model to generate behavior that will be radically different in the future?

(ii) What is it about the model, i.e., which constituent parameter(s) is it, that enables the model to perform its task; indeed, what is the balance between key and redundant parameters in performance of the specified task?

(iii) What is it about the model, i.e., which constituent parameter(s) is it, that enables the model to generate "no concern" or "high-end" exposures?

Having thus been liberated from the constraints of developing a test that **must** have access to in situ field observations, and having thus a procedure for distinguishing **key** parameters in the model from those that are **redundant** to the task at hand, one could begin to move towards the notion of a valid model being one that is maximally **relevant** to its task. Here "relevance" could be defined as the ratio of (key/redundant) parameters in model, a property notably independent of the size of the model.

Moreover, if the performance of the task is dependent upon many key parameters that are believed to be relatively well, this gives the analyst greater confidence in judging the model to be valid (albeit a qualitative judgement) than a result in which performance is dependent on just a few key parameters that are believed not to be known very well.

However, these are issues for the future; they are subjects for further research.

2.2 Summary

The validity of a model cannot be established without specification of the task the model is required to perform. In predictive exposure assessments the greater concern for model validation lies with the use of models in the generic screening process (in assessing scenarios for a wide array of situations that could occur), rather than in site-specific cases where local data either are available or can be collected. For the latter, validation can be addressed using "classical" measures of the performance of the model against the familiar definition of desired behavior as a set of time-series observations. The former is essentially a design task. The "scenarios for a wide variety of situations that could occur" will have to be specified a priori, as indeed they are in areas such as assessment of the performance of geologic repositories for the disposal of high-level nuclear waste (Davis et al, 1990). Otherwise the task of design has no meaning. They will almost certainly not be quantifiable in the terms of time-series of observations; they may have to be expressed in less restrictive terms, for example, a numerical encoding of expert opinion, perhaps derived from the manipulation of a belief network (Varis, 1994).

For predictive exposure assessments, then, particular weight will need to be given to determining validation status in terms of (i) compositional validity, and (ii) prior performance validity, in which the task (performance) specification is not necessarily cast in terms of a set of historical observations.

3 PROCEDURE AND METHODS

The problem of validation is simply this: will the given model perform its task reliably, i.e. at minimum risk, where risk is quantified as some function of the probability of an undesirable outcome and the damage resulting from this outcome? Alternatively, we might ask: which -- among several candidate models -- is the most reliable instrument for performing the given task?

Being in a position to answer these questions is the culmination of the entire preceding process of developing the model. The task of validation is served both by this process, which we shall now define, and by any supplementary analysis of uncertainty (with the latter subsuming herein the analysis of sensitivity). Further clarification and definition of the role of some of the more commonly used statistical methods can be found in the Appendix.

3.1 **Procedure of Model Development**

The ASTM standard E 978 - 84 defines a **model** as follows:

> **Model** (ASTM E 978 - 84)
>
> An assembly of concepts in the form of a mathematical equation that portrays understanding of a natural phenomenon.

It further defines **computer code (computer program)** as:

> **Computer code** (ASTM E 978 - 84)
>
> The assembly of numerical techniques, bookkeeping, and control language that represents the model from acceptance of input data and instructions to delivery of output.

The term **algorithm** is then defined as:

> **Algorithm** (ASTM E 978 - 84)
>
> The numerical technique embodied in the computer code.

These three introductory concepts set out merely the common "language" of model-building.

The process of model development proper divides into two parts: that which can be undertaken without reference to any field (or laboratory) data, i.e., that which is a function solely of the knowledge and imagination of the analyst; and that which must be undertaken with reference to some quantitative definition of the behavior of the system to which the model refers. It is clearly important that such a "quantitative definition of the behavior of the system" should be as independent as possible of the "knowledge and imagination of the analyst" that will have gone into the composition of the model.

3.1.1 **Synthesis**

As already noted, a model may be viewed as a complex assembly of several, if not many, constituent hypotheses. The developer of the model may be quite eclectic in the sources of knowledge that are tapped in the expression of this assembly in mathematical form (for example, the results of previous models, microcosm experiments, laboratory toxicity tests, and field studies). The process is essentially one of synthesis, impressively illustrated by Scavia (1980) in his development of a model for the ecology of Lake Ontario. This first of the two stages of developing the model is completed by the act of **(code) verification**, defined thus:

> **(Code) verification** (ASTM E 978 - 84)
>
> Examination of the numerical technique in the computer code to ascertain that it truly represents the conceptual model and that there are no inherent numerical problems with obtaining a solution.

Certain minor variations on this terminology are possible, notably in the use of the phrase **generic** model, defined by its authors as follows (Konikow and Bredehoeft, 1992):

> When a numerical algorithm is implemented in a computer code to solve one or more partial differential equations, the resulting computer code can be considered a *generic* model.

We may consider that the model is therefore constructed and **compositional validity** established. For irrespective of whether any subsequent calibration or analysis of uncertainty is to be undertaken, consensus (or disagreement) on the constituent hypotheses assembled together in the model, and their relative strengths and weaknesses, can be gauged, stated, and possibly quantified as an aggregate measure of the validity of the model as a whole. As already noted, this is a subject for peer group review of the model, and is discussed fully in Section 4.

3.1.2 **Analysis**

It is the purpose of the second phase of development to adjust the settings in this instrument of prediction in order to make it both usable and accurate. It may not be "usable" immediately upon construction simply because some of the coefficients (or parameters) in its many mathematical relationships have not been assigned values; it may not be "accurate" because incorrect values have been assigned to these parameters. There is then a need for **calibration** of the model:

> **Calibration** (ASTM E 978 - 84)

> A test of a model with known input and output information that is used to adjust or estimate factors for which data are not available.

Since the "factors" referred to here are the model's coefficients, or parameters, calibration is frequently referred to as a matter of **parameter estimation**.

The model may prove at this stage to be quite inadequate, if not plain wrong in the expression of some of its relationships, i.e., its constituent hypotheses. Its hypotheses may have been refuted by this test against the external behaviour definition, and no amount of adjustment of the internal parameters appearing in these relationships may compensate for its basic inadequacy. The encoded model may thus suffer from what has variously been called **structural error** (Beck, 1987), **conceptual errors** (Konikow and Bredehoeft, 1992) or **uncertainties in the conceptual model** (Usunoff et al, 1992), or **model error** (Luis and McLaughlin, 1992). A broader meaning and remit can thus be attached to the process of calibration: that it is a search for the source and reason of such error (Beck, 1987), although the simple definition given above will be quite sufficient for present purposes.

At the end of the procedure the model will exist as a specific object. Indeed, since by definition calibration involves at least one test of the model's performance, through a comparison of data derived from the model with data from the prototype system, there will also be evidence of its **posterior performance validity**, i.e., evidence of its reliability as an instrument of prediction. However, the status of this evidence in coming to a view on the crucial issue of the model's **prior** performance validity will be weakened by the degree to which the final values of the measures of agreement (or fit), **after** calibration, are conditioned upon successive adjustments of the model's parameter values. It is the purpose of calibration to seek the best possible match between the behaviors of the model and the system. Providing there is a sufficient number of parameters in the model there are in principle sufficient degrees of freedom to adjust the behavior of the model so that its match with the behavior of the system may become arbitrarily close. This, however, will provide little insight into how the model will perform under conditions not previously encountered.

3.2 **Validation**

After calibration, then, the process of model development will yield a set of relationships, a numerical solution procedure, and a set of values for all the internal parameters of the model. This will be entirely sufficient for application of the model in making the predictions that will comprise the test of the model's **prior performance validity**, providing that no further adjustments of the internal settings of the instrument are incorporated. This, the comparison of the model's results with data (or conditions) deduced from facts and sources of knowledge **utterly independent** of those used in composing the model, is the centerpiece of what the ASTM standard defines as "validation".

In almost all the cases considered hitherto in the literature in situ field observations have been assumed to be available for the purposes of assessing prior performance validity, as clearly reflected in the supplementary discussion of Appendix I. However, for the reasons set out in Section 2, such observations do not **have** to be an indispensable prerequisite for this assessment.

The assessment proceeds as follows, with four distinct components.

(i) **The raw "data"**

Some measure of the correspondence between the performance of the model and the performance embodied in the current task specification is to be computed, ultimately to inform the judgement about the model's validity. Simply, the sequences of model outputs, observed (system) output responses, and the differences between these two sets of sequences, can be considered collectively as the raw "data" available for manipulation in the validation process.

(ii) **Summarizing "properties" of the raw data**

There may well be great benefit in computing certain summarizing "properties" of these raw data. Attributes of their information content can thereby be expressed concisely, with a degree of discrimination against the spurious influences of random errors and events captured in the raw data. Such "properties" include (statistical) distribution functions, the moments of these distributions (e.g., their means), and the sets of coefficients appearing in correlation functions and regression relationships. In statistical terms, the computation of these "properties" from the raw data would be referred to as **estimation** (Reckhow et al, 1990).

(iii) **The "decision"**

Both the raw data and their summarizing properties constitute relevant information for making the central "decision": of whether to accept or reject the model as a valid instrument of prediction (under conditions that will normally be expected **not** to be identical with conditions observed in the past). The fact that the summarizing properties have been (objectively) computed does not deny the relevance of the original raw data, albeit subjectively interpreted, to the making of the "decision". Equally, when several summarizing properties have been objectively computed, which is often the case, their collective use in informing the "decision" will almost inevitably involve some **subjective** balancing of the relative importance of each constituent property (Luis and McLaughlin, 1992; Konikow and Bredehoeft, 1992). There is but a single decision; yet several summarizing properties may be interpreted collectively to inform it. And like all decisions, it would be best made in the light of as much relevant information as possible and will be subject to a particular decision rule, or perspective, of the decision-maker.

(iv) **The decision "statistics"**

Making a choice among several courses of action requires a rule expressing our preferences with regard to the probability of an event occurring and the costs (benefits) associated with the combination of the chosen course of action and the outcome of the (random) event. In the process of validation the choice is binary -- to accept or reject the model as a valid instrument of prediction -- and the outcome of the random event is also binary -- in the event the model may prove to be a "true" or a "false" representation of reality, with differing (monetary) consequences for each of the four possible combinations of the course of action and event outcome. In the light of these consequences consistent, if not automatic, rules determining the course of action deemed most preferable can be adopted. They do not have to be, for the analyst can process subjectively the information relevant to the decision in order to choose the course of action, without expressly declaring any such rules. However, an important form of more detached such rules for the decision are those guided by the computation of certain "statistics" (e.g., chi-squared, Student's t, Kolmogorov-Smirnov, and others), and these rules are what would be familiarly known as **statistical hypothesis testing** (Reckhow et al, 1990; Luis and McLaughlin, 1992). The rule is encapsulated in the degree to which the value of some "statistic" computed from the raw data differs from a reference value, with an element of risk (of making a wrong decision) embodied in the tolerance allowed for in what constitutes an acceptable difference. In practice, the tendency has been to work with a null hypothesis of "no significant difference between the model and the observations", and to err on the side of not rejecting a valid model.

Where further relevant sets of data are available, further tests of the model's performance may be conducted. If the evidence from all such assessments has been mustered, evaluation of the model's validity -- both compositional and performance, and spanning all the tests -- will have been completed. In other words, its reliability will have been determined, as an instrument of prediction up to, but not including, the current task specification. Its reliability vis a vis this current task can only be gauged by the degree to which we believe the features of the current task approximate features encountered in the past; and even this belief (or expectation) may prove, in the event, to be surprisingly false. In this respect, then, what is understood herein as validation differs from the definitions given in Versar Inc (1988) and quoted above. The difference may only be a matter of semantics, but we would argue that "... the diversity of conditions for which the [model] package has been shown to be valid ..." is **not** entirely "irrelevant"; there may be a degree of identity between some features of the current task and those encountered in the past.

There appears to be no formal means of modifying any computed, quantitative measures of compositional and performance validities by this expected "degree of similarity" between current and past task specifications. A qualitative, expert (or peer group) opinion on the subject, however, might be available; and undoubtedly the same subjective approach would be needed for combining the quantitative results of all the objective tests of model validity into a single index of validation status. If such a single index were quantifiable (and desirable), it almost goes without saying that a valid model should score relatively highly according to this index.

3.3 Analysis of Uncertainty

Other forms of analysis, which may be regarded as part of the process of developing the model, also serve the purpose of validation.

It is unreasonable to expect that no uncertainty will attach to a model and the predictions it generates. There are two facets to the analysis of such uncertainty, respectively reflections of the internal and external facets of the model's composition and performance. They are defined as follows:

> **Analysis of uncertainty**
>
> (i) Evaluation of the ranges (or distributions) of values that can be assigned to the model's parameters, where evaluation may be made, inter alia, on the basis of model calibration as a function of the specified sources of uncertainty associated with the data used for this test.
>
> (ii) Evaluation of the ranges (or distributions) of values that are associated with the predictions of the model's output variables, as a function, inter alia, of the uncertainty in the model's parameter values.

The former can be called upon in order to establish the model's **compositional validity**, in terms of each constituent parameter in the model. Such a measure will reflect the accumulating success (or failure) in the performance of the model against any sets of data employed in the development process. Its interpretation can be used to amplify and qualify the expert opinions in a peer group review of the model. The latter permits an alternative test of the model's **prior performance validity**. The reliability of the model can be understood simply as the "inverse", as it were, of the uncertainty of the predictions. The test may be performed under the novel conditions of the current task specification, by providing the scenarios for those situations that could occur in releasing the contaminant, including some quantitative measure of the (subjective) probability of occurrence of a particular scenario.

The analysis of sensitivity is a simpler subset of the analysis of uncertainty. It can be defined as:

> **Sensitivity** (ASTM E 978 - 84)
>
> The degree to which the model result is affected by changes in a selected input parameter.

The possibility that the value assigned to a model parameter is erroneous is thereby acknowledged. However, the magnitude of this error is not evaluated in an analysis of sensitivity. Rather it is assumed, usually to be at a standard level of 1% or 10%, for example, of the best estimated value of the given parameter; alternatively, the model's performance might be tested at the mean, minimum and maximum values of its various parameters. An analysis of the model's sensitivity reveals nothing **directly** of the reliability of the model's predictions. Indirectly, however, a model whose predictions differ greatly as a consequence of minor changes to its parameter values is of suspect reliability, especially if the "offending" parameters are either entirely novel components of the model or known from previous experience to be difficult to estimate accurately. This is similar to the suspect reliability of a model identified as having just a few (among its many) parameters that are crucial to satisfaction of the task definition but not at all well known (as in the earlier discussion of the concept of relevance).

Errors in the predictions of a model may derive from three sources:

(i) the estimated initial state of the system at the start of the forecasting horizon;

(ii) the assumed patterns of future variations in the input disturbances of the system (typically, such as precipitation, solar radiation, and release rates of contaminants);

(iii) the model, by which the actions of the inputs are transcribed into the evolution of the system's output response (typically, the resulting concentration of the contaminant at a receptor point in some sector of the environment).

Where a supplementary analysis of uncertainty is undertaken it implies knowledge sufficient for the quantification of these sources of uncertainty. There are several methods commonly used for such analyses: (i) a first-order error analysis; (ii) Monte Carlo simulation, possibly with a more efficient sampling scheme; and (iii) methods of response surface analysis (Cox and Baybutt, 1981; Beck, 1987; and Iman and Helton, 1988; Zimmerman et al, 1990). It may in addition be of interest to establish which, among the various elements of the above sources of uncertainty, contributes most to the resulting uncertainty of the model's predictions. It would be inappropriate, for example, to condemn a model as "invalid" if its predictions were found to be highly uncertain, but with the major source of this uncertainty deriving from inadequate knowledge of the input disturbances.

In the case of uncertainty in the model, this can simply be assumed to be reflected in the uncertainty associated with the model's parameters, and an initial evaluation of their uncertainty may be sought in the literature. Upper and lower bounds on the feasible ("realistic") values of the model's parameters are normally available, and they may be used to bracket the uncertainty in the relevant parameter. Rather less straightforwardly, quantification of the model's uncertainty may be achieved through more advanced forms of model calibration, although this has rarely been practiced (Beck, 1987). Still less straightforwardly, the model could be said to be uncertain in respect of the mathematical expressions of its hypothetical relationships (in which the parameters appear). This has been referred to as a structural error in the model, and only in the work of van Straten and Keesman (1991) is there evidence of any attempt to quantify it and its consequences for prediction uncertainty.

3.4 **Further Considerations**

3.4.1 **Articulation**

One of the enduring problems of modelling is the question of establishing the correct degree of complexity required of a model for a particular purpose. Balancing the number of parameters in a model (a measure of its "complexity") against the goodness of fit of that model to a set of data (its "accuracy") is an equally enduring problem, as already observed. Intuitively, a "good" model would contain relatively few parameters yet be able to predict behavior accurately over a wide range of conditions.

In the analysis of time-series using polynomial expressions (in the backward shift operator) it is possible to develop criteria, such as Akaike's Information Criterion (AIC; Akaike, 1974), that allow the analyst to determine when, in effect, further improvement in fit is being bought at the expense of over-parameterisation. If a model is over-parameterised it has, in the light of the earlier discussion of calibration, too many degrees of freedom. This freedom can be both a benefit and a liability: a benefit in creating the capacity to predict conditions not previously encountered; a liability in allowing the false impression of a reliable model, able to match closely all the spurious, chance quirks of past observed behavior.

Costanza and Sklar's (1975) contribution to the subject of model validation (interpreted very broadly) is an attempt to find an AIC that is applicable to nonlinear state-space models of the kind commonly used in predictive exposure assessments (time-series models are not used for such purposes). These authors propose a measure of the complexity of the model in terms of its **articulation** (the number of parts, or elements, into which it is divided), along the three dimensions of ecological, temporal, and spatial resolution. Goodness of fit is gauged by conventional measures, such as the coefficient of determination. The "effectiveness" of the model is then defined as the product of these two measures of complexity and goodness of fit.

3.4.2 **Control of user errors**

What is considered a "good" (or valid) model from one perspective may not be considered a "good" model from another perspective. In particular, the position of the user of a model will be different from that of its developer. In this respect Burns et al (1990) have made a significant contribution in arguing a strong case in favor of re-orienting the way in which risks associated with the acceptance or rejection of a model (as an instrument of prediction) are controlled. They have put it this way (Burns et al, 1990, pp 35/6):

> Objective validations can be conducted only when the criteria for validity are objectively specified. Because the social consequences of accepting false models (inadequate chemical safety regulations) are much more serious than the consequences of rejecting true models (continued research and validation studies), model validations should always be phrased to test the null hypothesis that "the model is invalid" ... until proven otherwise.

They have also introduced the relevant means of computing the minimum number of sample observations required to ensure that the validity of the model can be assessed (with confidence) at the chosen levels of developer and user risks of a wrong judgement.

The procedure outlined earlier remains the same. Merely the orientation of the null hypothesis in assessing the model's prior performance validity is thereby changed. Such a change is clearly endorsed by Luis and McLaughlin (1992), who state:

> Decision errors can be classified as either Type I (rejecting the hypothesis when it is true) or Type II (accepting the hypothesis when it is false). If the test is very stringent it will have a small Type II error and a large Type I error (i.e. it will tend incorrectly to reject good models).

and is further supported by the following definition of validation from studies on evaluating the performance of regional acidic deposition models (NAPAP, 1990):

> **Validation** - the determination of the correctness of a model with respect to the user's needs and requirements.

3.4.3 **Capability index**

It is customary for a single value of a state variable, as generated by the model, to be examined for its coincidence with a single observation from the field of the same quantity (at a given time and location). In assessing the validity of the model on this basis the objective is to establish to what extent the model differs from the "truth", **not** the inevitably error-corrupted observation of this (unknown) quantity. The null hypothesis, for which evidence for the rejection or discrediting thereof is sought, is commonly stated as "no significant difference between the model and the truth".

In the extreme case of a perfect observing instrument, the field observation **is** the truth. And since the model, by definition, can never be a wholly truthful representation of reality, Parrish and Smith (1990) have made the point that it would be very unlikely for the model's prediction to be coincident with this observed truth. The model would accordingly be rejected routinely as not valid. They question, therefore, the underlying principle of a test of coincidence that presumes the possibility of "equality" between the model and the truth. They argue that validity should be established as a function of the model's prediction lying within a bounded range of values for the truth, for example, that it lies within a factor of two of the truth. This moves the underlying principle from one of presumed equality to one of presumed "inequality", in the sense of not greater than a certain "distance" from the truth.

For practical purposes Parrish and Smith (1990) construct a test for the validity of the model that rests upon there being an overlap between the ranges of values computed for the two quantities: the model's prediction, where the upper and lower bounds are computed simply as division and multiplication of the nominal model estimate by the chosen factor (e.g., two); and the "truth", as estimated from the mean of the sample of field observations of the quantity, with the addition or subtraction of an appropriate t statistic multiplied by the standard error of the sample of observation. They present the results of their test as a **capability index** (of prediction). This will assume the value of one for any overlap between the two computed ranges, and progressively more than one as the two ranges become distinct and indeed further separated.

Lack of perfection can of course reside on both sides of the test of coincidence: in the model's predictions as much as in the field observations. The arguments of Parrish and Smith clearly spring from considerations of the latter alone. Those of Burns et al (1990) and Reckhow et al (1990) equally clearly acknowledge both sources of uncertainty and their joint role, not merely in undermining, but in perverting, the power of this test. The more uncertain the model, the better able it is to withstand the conventionally directed test of erring on the side of not rejecting a valid model; and it is this problem that has stimulated in Burns et al (1990) the reorientation of perspective in the control of such errors of judgement. In the extreme case, the trivial prediction that all things are equally probable, the truth is bound to be covered somewhere.

In the work of Reckhow et al (1990) there are common elements of concern shared with both Burns et al (1990) and Parrish and Smith (1990). For in discussing the presumed "equality" between the model and the truth, they observed that (Reckhow et al, 1990):

> ... the basic null hypothesis indicative of a good model - H_0, where the underlying distribution of predictions and the underlying distribution of observations are identical - may be accepted because the model provides a good fit to the data or because the model and/or data are quite variable.

The latter is self-evidently misleading.

3.4.4 Adequacy and reliability

The distinctive contribution of Mankin et al (1977) was to introduce a discussion of models, and their performance against observations of the real world, in terms of Venn diagrams. They used these diagrams to define the **adequacy** and **reliability** of a model, i.e.,

Adequacy = {No of agreements between model and "experiments"}/{No of "experiments"}

Reliability = {No of agreements between model and "experiments"}/{No of "model responses"}

Their paper does not include, however, a precise definition of what is meant by an "experiment", or a "model response". For the present purposes a single "experiment" will be taken to be a definition of the required/observed behavior over a **span** of time, not a single observation at a single instant in time (that this was probably the intended usage of the term can be inferred from a closely related paper published subsequently by the same authors; Cale et al, 1983).

Mankin et al (1977) then proceed to argue that the above two measures may be used to discriminate a "better" from a "worse" model, and that they may also be used for the purposes of experimental design (interpreted in its broadest sense, and in a manner closely similar to that of Burns et al, 1990). They adopt a conventional null hypothesis (i.e., a test erring on the side of not rejecting a valid model) and then use Bayes' rule to establish what minimum level of model adequacy would be required for a further "experiment" (set of observations) to yield a smaller risk (posterior probability) of rejecting a valid model. In other words, for Bayes' rule to be employed the following assumptions are made:

The prior:

is the (chosen) probability of rejecting a valid model before the "experiment" (conventionally referred to as probability ").

The accuracy of the observing instrument (as gauged by the probability of "correspondence" between the "model response" and the "experiment"):

is the adequacy of the model.

The posterior:

is the revised probability of rejecting a valid model after the "experiment" (which, to be of interest, should be less than ").

A relationship between model adequacy and the number of "experiments" required to assess the validity of the model (with confidence), at a specified level of risk of rejecting a valid model, can thus be established.

3.4.5 **Relevance**

It is apparent that the definition of reliability, as introduced by Mankin et al (1977), plays no part in their subsequent analysis of experimental design. This is disappointing, because the concept is not spurious and indeed has very strong similarities with the notions of **key** and **redundant** model parameters implied by the Hornberger-Spear-Young algorithm discussed earlier.

It has been said that assessing the prior performance validity of a model need not depend upon the availability of time-series observations of the model's output variables. The design task to be performed by the model can be stated in less narrow terms: as a function of an expert opinion on scenarios for future behavior that could occur (as in Davis et al, 1990) translated, for example, through the use of a belief network (Varis, 1994), into a "corridor" of acceptable ranges of values through which the model's outputs must pass. Different combinations of values for the parameterisation of the model can be generated at random -- albeit guided by reasonable constraints on the degree to which each constituent hypothesis is considered to be uncertain (this being notably the compositional validity of the candidate model). The resulting performance of the model for each of these trial parameterisations may be deemed to have been a success or failure, according to whether the (broadly stated) task is satisfied or not. And the reliability of the model, in the terms introduced by Mankin et al (1977), is then straightforwardly the ratio of the number of such successes divided by the total number of trial experiments with the model.

Furthermore, the search for a model that is maximally **relevant** to its stated task (Section 2.1.5) would emerge from the same form of test as having a high ratio of (key/redundant) parameters, where a key parameter is one that is important in determining whether the model succeeds or fails in its task, and a redundant parameter is one to whose values such success or failure are indifferent.

Both of these measures (i.e., of reliability and relevance) are in principle independent of the size of the model, which is undoubtedly a desirable feature. In fact, the notion of a model being best able to perform its design task has been shifted away from some measure of closeness of its performance to the externally specified design task (in terms of model outputs) towards the properties of its internal composition. This is not to say that such optimal fits are not important. Rather, the benefit of developing these nascent concepts would lie in having some other forms of statistic that depend on the intrinsic composition of the model.

3.5 **Closing Remarks**

It is perhaps a symptom of the difficulty of the problem of model validity that there are so many definitions of it. Unlike calibration, whose definition is widely agreed and has remained consistent over the years, the subject of validation continues to attract attempts at quantitative expression of the natural language terms in which it is discussed (for example, articulation, capability, adequacy, reliability, desirability, and relevance).

In predictive exposure assessments, and in particular for screening-level analyses, the classical definition of validation -- such as the "comparison of model results with observations of the system's behavior other than those used for calibration" -- is not wholly satisfactory. Yet, except for the measures of compositional validity, no quantitative substitute has yet been identified, although promising lines of research are discernible.

If the requisite field data for the classical tests of validation are available, however, the methods developed and illustrated by Burns et al (1990), Reckhow et al (1990) and Parrish and Smith (1990) are collectively those that should be applied as the current best practice.

4 **PEER REVIEW**

4.1 **The goals of peer review**

Peer review is an important element of modeling. Peer review goals for models are no different from any other scientific investigation. In brief, the idea is to provide a sufficiently detailed description of the completed work to enable others to thoroughly evaluate its merits and to provide enough detail to enable others to reproduce, confirm, or challenge the results. Provision of this detail in written form in discipline or problem-oriented journals is the most common, and preferred procedure. For any particular **body of research**, peer review through a series of manuscript submissions, reviews, publications, and subsequent additional research are events in the scientific method. It is this **continuing process** that is most important in improving scientific understanding rather than any individual review (or publication that reflects such a review). This is not to say that seminal and definitive articles do not exist, rather that realization of the importance of the work and its adoption as a basis for further work (and in some cases as a basis for government policy) most often comes after additional work, reviews, and publications.

4.2 **Limitations of traditional peer review approaches as applied to modeling**

The traditional peer review approach of publishing in appropriate scientific journals remains a very desirable if not essential component of model validation but may be inadequate as definitive evidence of scientific rigor as models become more complex, expand the number and extent of their constituent hypotheses, and are refined or redesigned to accomplish new tasks. Mackay (ES&T, 1988) was among the first to articulate this problem as part of a lead editorial note.

> "...environmental science and management rely increasingly on complex models to describe, for example, the complex behavior of chemicals in a multimedia environment, routes to human exposure, spill damages, atmospheric dispersion in complex terrain, or extensive ionic equilibria....How can we ensure that such models are valid, free from mistakes, and thus reliable tools in the hands of scientists and managers?...Complex, computer-based models can play an important role in environmental science, but we cannot expect the existing review system to give them the scrutiny they need and deserve. Those who fund, develop, and ultimately use models must be willing to seek, encourage, and sponsor novel peer reviewing approaches to ensure the scientific rigor of the published word, which is at the core of scientific progress."

Mackay properly identified peer review as a requirement to establish validity but he has also voiced that novel approaches are needed. This view, it could be said, reflects the advances in modeling research. Early in the development of computer-based models, publication of new models was common and deemed appropriate as part of the scientific debate on how to structure and design models, how to solve constituent equations, how to identify and accommodate boundary conditions, or to propose a specific model as an efficient and reliable tool to accomplish certain tasks. Now, peer publication is most often restricted to model application studies and elaborations on model refinements made to improve model performance, usually for a new task. The peer reviews attendant to such publications are an important source of evidence for demonstrating either the compositional validity of the model or its performance in accomplishing a specific task.

Models are most often described in detail only through the "gray" literature. The usual case is publication of reports under the procedures used by the sponsoring organization. The EPA, NRC, USGS, USDA, DoE are but a few government organizations that have and continue to publish reports that describe models. The extent that such documents are carefully and extensively reviewed is variable, largely unknown, and in many cases the agencies add disclaimers that qualify (in some cases essentially disavow) their endorsement of the published material. The variable nature of these publications and the lack of the traditional "discipline-oriented" imprimatur raises questions about the peer review status of the model in question. However, the designation of "gray literature" or the attendant qualified endorsements by the publishing organization does not diminish the importance of such publications. **Because these documents may be the only detailed description of the model, they serve as an absolutely essential reference for the continuing peer review process cited in the previous section.**

Mackay also questions whether the "published word" will ensure scientific rigor in the case of models. This is a problem. Absent a detailed, line-by-line description of the model, all the input data, parameter values, and constituent equations described as part of a publication it is simply not possible to reproduce, confirm, or meaningfully challenge the results presented. (Citations of previously published work may serve to mitigate this problem, at least in part, and this will be elaborated in a later section of this chapter.)

4.3 **Peer review of models within EPA**

The Agency's Task Force on Environmental Regulatory Modeling (EPA, 1993) has recently identified external peer review as an essential phase of model development, modeling applications, and use of models in the decision-making process. The report outlines a developing policy on these matters and suggests specific procedural steps intended to assist Agency program managers in conducting such reviews. Apparently, one motivation for the Task Force effort is the present lack of uniform and consistent practice in this area. The essential point of the report is that an external review is required in order to ensure an independent review. Our perspective is that external peer review as described by the Task Force report is in use to varying degrees within EPA and should be encouraged.

4.4 **Peer review as part of model validation: a three-fold strategy**

Chapter three of this report lays out the procedures and methods for validating models. Peer reviews are important components of the "weight of evidence" to be assembled in that process. The limitations and inherent difficulties presented by traditional peer review have been examined; a particularly articulate critique of peer review of complex models was noted. A three-fold strategy emerges as described below to improve modeling peer review.

(i) Reference code, version documentation, and test data set maintenance. The difficulties in not knowing all the information required to reproduce results completed by others and published in various forms can be alleviated by having an "original" version (or extended versions) of the model resident in an organization that can maintain and distribute unaltered, reference copies to anyone who needs them. Once established, this system would enable citations in published documents that, in principle, permit reviewers to reconstruct the exact model that was used, modified, or discussed in the document. Similarly, a standard test data set enables a convenient way to verify any given code.

(ii) Publication in referred journals. Publication of modeling refinements, performance testing, applications, and solution techniques should continue. As an ongoing process, such peer review adds substantially to the weight of evidence that the model is reliable for the circumstances described. This process is strengthened by the availability of the reference codes as described above.

(iii) Periodic or issue-specific group peer reviews. As models have become increasingly complex and aggregate science across more than one discipline, it is increasingly clear that more than one subject matter expert is required to provide adequate scientific review. Accordingly, group reviews are needed. This comports with recommendations of the Task Force on Regulatory modeling.

5 CONCLUSIONS

It is not reasonable to equate the validity of a model with its ability to predict correctly the future "true" behavior of the system. A judgement about the validity of a model is a judgement on whether the model can perform its designated task reliably, i.e., at minimum risk of an undesirable outcome. It follows that whomsoever requires such a judgement must be in a position to define -- in sufficient detail -- both the **task** and the **undesirable outcome**.

However desirable might be the application of "objective" tests of the correspondence between the behavior of the model and the observed behavior of the system, their results establish the reliability of the model only inasmuch as the "past observations" can be equated with the "current task specification". No-one, to the best of our knowledge, has yet developed a quantitative method of adjusting the resulting test statistics to compensate for the degree to which the "current task specification" is believed to diverge from the "past observations".

This in no way denies, however, the value of these quantitative, objective tests wherever they are applicable, i.e., in what might be called "data-rich" problem situations. Indeed, there is the prospect that in due course comparable, quantitative measures of performance validity can be developed for the substantially more difficult (and arguably more critical) "data-poor" situations, in which predictions of behavior under quite novel conditions are required by the task specification.

In this concluding section, the purpose of the protocol for model validation set out below is to provide a **consistent** basis on which to conduct the debate, where necessary, on the validity of the model in performing its designated task reliably. It seeks **not** to define what will constitute a valid model in any given situation, but to establish the framework within which the process of arriving at such a judgement can be conducted. It acknowledges that no evidence in such matters is above dispute, not even the evidence of "objective" measures of performance validity, which themselves must depend on some subjectively chosen level of an acceptable (unacceptable) difference between a pair of numbers.

5.1 **The Protocol**

There are three aspects to forming a judgement on the validity, or otherwise, of a model for predictive exposure assessments:

(i) the nature of the predictive **task** to be performed;

(ii) the properties of the **model**; and

(iii) the magnitude of the **risk** of making a wrong decision.

For example, if the task is identical to one already studied with the same model as proposed for the present task and the risk of making a wrong decision is low, the process of coming to a judgement on the validity of the model ought to be relatively straightforward and brief. Ideally, it would be facilitated by readily available, quantitative evidence of model performance validity. At the other extreme, if the task is an entirely novel one, for which a novel form of model has been proposed, and the risk of making a wrong decision is high, it would be much more difficult to come to a judgement on the validity of the model. Evidence on which to base this judgement would tend to be primarily that of an expert opinion, and therefore largely of a qualitative nature.

While the depth of the enquiry and length of the process in coming to a judgement would differ in these two examples, much the same forms of evidence would need to be gathered and presented. It is important, however, to establish responsibilities for the gathering of such evidence, for only a part of it rests with the agency charged with the development of a model. In the following it has been assumed that a second, independent agency would be responsible for specification of the task and evaluation of the risk of making a wrong decision. The focus of the protocol will accordingly be on the forms of evidence required for evaluation of the model.

5.1.1 Examination of the model's composition

The composition of a model embraces several attributes on which evidence will need to be presented. These are as follows:

(i) **Structure**. The structure of the model is expressed by the assembly of constituent process mechanisms (or hypotheses) incorporated in the model. A constituent mechanism might be defined as "dispersion", for example, or as "predation of one species of organism by another". The need is to know the extent to which each such constituent mechanism has been used before in any previous (other) model or previous version of the given model. There might also be a need to know the relative distribution of physical, chemical and biological mechanisms so incorporated; many scientists would attach the greatest probability of universal applicability to a physical mechanism, and the smallest such probability to a biological mechanism.

(ii) **Mathematical expression of constituent hypotheses**. This is a more refined aspect of model structure. The mechanism of "bacterial degradation of a pollutant" can be represented mathematically in a variety of ways: as a first-order chemical kinetic expression, in which the rate of degradation is proportional to the concentration of the pollutant; or as, for instance, a function of the metabolism of bacteria growing according to a Monod kinetic expression.

(iii) **Number of state variables**. In most models of predictive exposure assessments the state variables will be defined as the concentrations of contaminants or biomass of organisms at various locations across the system of interest. The greater the number of state variables included in the model the less will be the degree of aggregation and approximation in simulating both the spatial and microbial (ecological) variability in the system's behavior. In the preceding example of "bacterial degradation of a pollutant", only a single state variable would be needed to characterize the approximation of first-order chemical kinetics; two -- one each for the concentrations of both the pollutant and the (assumed) single biomass of bacteria -- would be required for the constituent hypothesis of Monod kinetics. Similarly, a lake characterized as a single, homogeneous volume of water will require just one state variable for the description of pollutant concentration within such a system. Were the lake to be characterized as two sub-volumes (a hypolimnion and an epilimnion), however, two state variables would be needed to represent the resulting spatial variability of pollutant concentration.

(iv) **Number of parameters**. The model's parameters are the coefficients that appear in the mathematical expressions representing the constituent mechanisms as a function of the values of the state variables (and/or input variables). They are quantities such as a dispersion coefficient, a first-order decay-rate constant, or a maximum specific growth-rate constant. In an ideal world all the model's parameters could be assumed to be invariant with space and time. Yet they are in truth aggregate approximations of quantities that will vary at some finer scale of resolution than catered for by the given model. For instance, the first-order decay-rate constant of pollutant degradation subsumes the behavior of a population of bacteria; a Monod half-saturation concentration may subsume the more refined mechanism of substrate inhibition of metabolism, and so on. In problems of groundwater contamination the volumes (areas) over which the parameters of the soil properties are assumed to be uniform are intertwined with this same problem of aggregation versus refinement. There is immense difficulty, however (as already noted in discussion of the concept of **articulation**), in establishing whether a model has the correct degree of complexity for its intended task.

(v) **Values of parameters.** Again, in an ideal world the values to be assigned to the model's parameters would be invariant and universally applicable whatever the specific sector of the environment for which a predictive exposure assessment is required. In practice there will merely be successively less good approximations to this ideal, roughly in the following descending order:

(a) The parameter is associated with an (essentially) immutable law of physics and can accordingly be assigned a single, equally immutable, value;

(b) The parameter has been determined from a laboratory experiment designed to assess a single constituent mechanism, such as pollutant biodegradation, under the assumption that no other mechanisms are acting upon the destruction, transformation, or redistribution of the pollutant within the experiment;

(c) The parameter has been determined by calibration of the model with a set of observations of the field system;

(d) A value has been assigned to the parameter on the basis of values quoted in the literature from the application of models incorporating the same mathematical expression of the same constituent process mechanism.

It is misleading to suppose that the result of (b) will be independent of an assumed model of the behavior observed in the laboratory experiment. The coefficient itself is not observed. Instead, for example, the concentration of pollutant remaining undegraded in the laboratory beaker or chemostat is observed. Once a mathematical description of the mechanism assumed to be operative in the experiment is postulated, then the value of the parameter can be inferred from matching the performance of this model with the observations (which in effect is the same procedure as that of (c)).

(vi) **Parameter uncertainty**. Evidence should be presented on the range of values assigned to a particular parameter in past studies and/or on the magnitude and (where available) statistical properties of the estimation errors associated with these values. In many cases it might be sufficient to assume that such ranges of values and distributions of errors are statistically independent of each other, but this can be misleading. Supplementary evidence of the absence/presence of correlation among the parameter estimates and errors could be both desirable and material to the judgement on model validity. For example, unless determined strictly independently -- and it is not easy to see how that might be achieved -- the values quoted for a bacterial growth-rate constant and death-rate constant are likely to be correlated. A pair of low values for both parameters can give the same net rate of growth as a pair of high values, and knowledge of such correlation can influence both the computation of, and assessment of, the uncertainty attaching to a prediction of future behavior.

(vii) **Analysis of parameter sensitivity**. The extent to which the predictions of the model will change as a result of alternative assumptions about the values of the constituent parameters can be established from an analysis of parameter sensitivity. On its own such information provides only a weak index of model validity. It may be used, nevertheless, to supplement a judgement on the model's compositional validity based on the foregoing categories of evidence. In the absence of any knowledge of parameter uncertainty an analysis of sensitivity may yield insight into the validity of the model's composition through the identification, in extreme cases, of those "infeasible" values of the parameters that lead to unstable or absurd predictions. It could be used thus to establish in crude terms the domain of applicability of the model, i.e., ranges of values for the model's parameters for which "sensible" behavior of the model is guaranteed. In the presence of information on parameter uncertainty an analysis of sensitivity may enable rather more refined conclusions about the validity of the model. In particular, a highly sensitive, but highly uncertain, parameter is suggestive of an ill-composed model.

It is clearly impossible to divorce an assessment of the evidence on the model's compositional validity -- its intrinsic properties and attributes -- from the current task specification. In particular, the less immutable the hypothesis (law) incorporating a given parameter is believed to be, the more relevant will become a judgement about the degree to which the current task specification deviates from those under which the values previously quoted for this parameter were derived. Such judgement will be especially difficult to make in the case of quantifying the correspondence (or divergence) between the laboratory conditions used to determine a rate constant and the field conditions for which a predictive exposure assessment is required. The judgement, nevertherless, is directed at the internal composition of the model, albeit conditioned upon the degree of similarity between the current and previous task definitions.

5.1.2 Examination of the model's performance

Evidence must also be assembled from the results of tests of a model's performance against an external reference definition of the prototype (field) system's behavior. This will have various levels of refinement, approximately in the following ascending order.

(i) **Unpaired tests**. In these the coincidence between values for the model's state variables and values observed for corresponding variables of the prototype system at identical points in time and space is of no consequence. It is sufficient merely for certain aggregate measures of the collection of model predictions and the collection of field data to be judged to be coincident. For example, it might be required that the mean of the computed concentrations of a contaminant in a representative (model) pond over an annual cycle is the same as the mean of a set of observed values sampled on a casual, irregular basis from several ponds in a geologically homogeneous region. Within such unpaired tests, there are further, subsidiary levels of refinement. A match of mean values alone is less reassuring than a match of both the means and variances, which is itself a less incisive test than establishing the similarity between the two entire distributions.

(ii) **Paired tests.** For these it is of central concern that the predictions from the model match the observed values at the same points in time and space. Again, as with the unpaired tests, subsidiary levels of refinement are possible, in providing an increasingly comprehensive collection of statistical properties for the errors of mismatch so determined.

(iii) **Sequence of errors.** A paired-sample test, as defined above, makes no reference to the pattern of the errors of mismatch as they occur in sequence from one point in time (or space) to the next. When sufficient observations are available a test of the temporal (or spatial) correlations in the error sequences may yield strong evidence with which to establish the performance validity of the model. In this case a "sufficiency" of data implies observations of the contaminant concentration at frequent, regular intervals over relatively long, unbroken periods.

In much the same way as it is not possible to divorce an assessment of the compositional validity of a model from its current and past task specifications, so it is not possible to divorce an assessment of performance validity from the composition of the model. Thus a further two categories of evidence are relevant.

(iv) **Calibration.** The task of model calibration necessarily involves adjustment and adaptation of the model's composition. The extent to which the values of the model's parameters have thereby been altered in order for the model to fit the calibration data set may render inadmissible the use of any associated error statistics for the purposes of judging model validity. It is therefore especially relevant for evidence of this form to be declared.

(v) **Prediction uncertainty.** All models may be subjected to an analysis of the uncertainty attaching to their predictions. Such an analysis will depend on the composition of the model -- through the quantification of parameter uncertainty; and it will depend upon the task specification, through a statement of the scenarios for the input disturbances and initial state of the system, i.e., the boundary and initial conditions for the solution of the model equations. The fact that the ambient concentration of the contaminant cannot be predicted with sufficient confidence does not necessarily signify an invalid model, however. For there are three sources of uncertainty in the predictions, two of which (the initial and boundary conditions) are independent of the model. Good practice in the analysis of prediction uncertainty (if a judgement on model validity is the objective) should therefore include some form of ranking of the contributions each source of uncertainty makes to the overall uncertainty of the prediction. Where Monte Carlo simulation is used to compute the distributions of the uncertain predictions, some -- perhaps many -- runs of the model may fail to be completed because of combinations of the model's parameter values leading to unstable or absurd output responses. As with an analysis of sensitivity, this provides useful information about the robustness of the model and restrictions on its domain of applicability. The less the model is found to be restricted, so the greater is the belief in its validity. In some cases, it may be feasible and desirable to state the output responses expected of the model in order for the task specification to be met, thus enabling a more refined assessment of the domain of applicability of the model (as in discussion of the concept of **relevance**). The use of combinations of parameter values leading to unacceptable deviations from the behavior of the task specification can be placed under restrictions.

5.1.3 Task specification

Judgements on both the compositional and performance validity of the model are inextricably linked with an assessment of the extent to which the current task specification diverges from the task specifications of previous applications of the model. Categories of evidence relating to the fundamental properties of the task specification must therefore be defined, in a manner similar to those assembled in order to conduct an assessment of the model.

For example, a model used previously for prediction of a chronic exposure at a single site with homogeneous environmental properties may well not be valid -- in terms of performing its task reliably -- for the prediction of an acute exposure at several sites with highly heterogeneous properties. It is not that the model is inherently incapable of making such predictions, but that there is an element of extrapolation into novel conditions implied by the different task specification. It is not the purpose of this document, however, to provide anything other than a very preliminary indication of the categories of evidence required to assess the degree of difference between current and past task specifications, as follows.

(i) **The contaminants**. The class(es) of chemicals into which the contaminant would most probably fall, such as chlorinated hydrocarbon, or aromatic compound, for example, must be specified. The number of such chemicals to be released, and their interactions (synergism, antagonism, and so on) vis a vis the state variables of interest in the environment, must also be specified.

(ii) **The environment**. Several attributes can be employed to characterize the similarities and differences among the environments into which the contaminant is to be released. These include, inter alia, the geological, hydrological, and ecological properties of the sites of interest, together with statements of the homogeneity, or heterogeneity, of the site according to these attributes.

(iii) **Target organism, or organ**.

(iv) **Nature of exposure**. The obvious distinction to be made in this case is between acute and chronic exposures of the target organism to the contaminant.

APPENDIX

ROLE OF STATISTICAL MEASURES AND METHODS

In order for the presentation of the applicable methods and statistical measures of validation to be as clear as possible it is necessary to introduce here some mathematical statements of the model and the terms and variables associated with it.

1 **Preliminaries**

Let us assume therefore that the model quantifies the unsteady-state, i.e., dynamic, behavior of the environment into which contaminants might be released as the following set of state-space, ordinary differential equations

$$\dot{x}(t) = f\{x, \alpha, u; t\} + \xi(t)$$

(1)

associated with which (error-corrupted) observations at discrete instants t_k of the outputs **y** may be available

$$y(t_k) = h\{x, \alpha; t_k\} + \eta(t_k)$$

(2)

Here **x** is the vector of state variables (typically, the concentrations of the contaminant at various points in space), **u** is a vector of input disturbances of the system (typically, such as precipitation, solar radiation, or the rate of release of the contaminant, for example, from a landfill site), **"** is the vector of model parameters (constants or coefficients, such as a growth-rate parameter, or volatilization rate constant), **y** the set of observed values of the state variables, **>** a vector of unknown, random disturbances of the states of the system, and **Θ** is a vector of measurement errors associated with the observation **y**. The dot notation in **xΘ** equation (1) denotes differentiation with respect to time t, and t_k in equation (2) signifies the pragmatic restriction of the observations to discrete instants (k) in time. It is highly unlikely that strictly continuous records of the movement of a contaminant through an environment would be available in practice.

The behavior of the system is observed through its input perturbations (**u**) and its output responses (**y**). These "observational facts", or data [**u**,**y**], may be referred to as the **external description** of the system's behavior. The prior constituent hypotheses about the mechanisms believed to govern this behavior are composed as the model in terms of its state variables (**x**) and the parameters (**"**) that appear in the relationships among **u**, **x**, and **y**. Collectively [**x**,**"**] may be referred to as the **internal description** of the system's behavior; they reflect the nature of the (hypothetical) internal workings of the system.

The observed facts [**u**,**y**] are both subject to errors, although equations (1) and (2) above acknowledge formally and explicitly the errors associated only with the observations of the outputs, i.e., **Θ**. Any errors attaching to **u** may therefore be assumed to be implicit, in effect, in the definition of **>**. Given [**u**,**y**] over a period of time (t_0 # t # t_N) for calibration purposes, the data for **u** are inserted in sequence into the model such that an estimate **í** of the observations **y** can be generated at each measurement instant t_k, where **í** is given by

$$\hat{y}(t_k) = h\{\hat{x}, \hat{\alpha}; t_k\} \tag{3}$$

in which **x**(t_k) is derived from the integration of equation (1) as

$$d\hat{x}(t)/dt = f\{\hat{x}, \hat{\alpha}, u; t\} \tag{4}$$

Equation (4) will require an evaluation of the initial state of the system $\mathbf{x}_0$ at time t = t_0, and both equations (3) and (4) will require knowledge of the parameter values, i.e., the estimates **"**. Since neither **>** nor **Θ** can be known, they do not appear in equations (3) and (4). (Their presence in equations (1) and (2) signals merely that they are present in reality, and must therefore enter into a judgement about the validity of the model).

Strictly speaking, the measure of agreement required for assessing the performance validity of the model can only be computed in terms of the external description of the system, i.e., in the context of that which can be observed. Invariably this implies agreement in terms of the output, i.e., **y** and **í** (from equations (2) and (3) respectively), so that the mismatch between the behavior of the model and behavior of the system is gauged by the error **e**, where

$$e(t_k)=y(t_k)-\hat{y}(t_k)$$

(5)

Frequently, the relationship between the state variables **x** and the outputs **y**, as characterized by the function **h**{@} in equation (2), will be straightforwardly such that the outputs are simply error-corrupted observations of the states, i.e.,

$$y(t_k)=x(t_k)+\eta(t_k)$$

(6)

in which case **í** is identical with **x**.

The relevant methods of assessment can now be introduced in the order of the procedure set out in Section 3.1 above, i.e., **model development**, **prior performance validity**, and **analysis of uncertainty**, where this last includes the analysis of sensitivity and subsumes the quantitative evaluation of compositional validity.

2 Model development: posterior performance validity

For all practical purposes conventional tests on the performance validity of the model (both prior and posterior) are conducted on the basis of the numbers available for the quantities **y**(t_k), **í**(t_k), and **e**(t_k). However incomplete such tests may be, as we have argued above in Chapter 2, they are by and large the only quantitative measures upon which to base a decision about the model's validity.

The details of these tests will be discussed fully under the principal heading of **prior performance validity** below. The point requiring emphasis here is elaboration of the way in which a test result deriving from the process of calibration during model development, denoting thus **posterior** performance validity, is a more or less weak approximation of the all-important property of **prior** performance validity.

Let us suppose that when the model is first confronted with the observed data [**y**] -- for calibration purposes -- a set of values **"**$_0$ are substituted for the parameters. Under this substitution the model generates estimates of the outputs, i.e., **í**, which we may denote more precisely as **í**{**"**$_0$;t_k}, and from which in turn values of the errors **e**{**"**$_0$;t_k} can be obtained. According to the nature of these errors, adjustments are made to the values of the parameters used in the model, for example, a change from **"**$_0$ to **"**$_1$, with **"**$_1$ then being substituted in a second trial attempt at matching the model with the observations. The process of adjustment may be repeated many times, leading ultimately to a set of parameter values **"**$_M$ that reflects the conclusion of M successive attempts at matching **í** with **y**.

Progress in moving from **"**$_0$ to **"**$_M$, and by association, success in calibration, are gauged by the extent to which the errors **e**{**"**$_M$;t$_k$} are in some quantitative sense "smaller" than **e**{**"**$_0$;t$_k$}. At the extreme, if **e**{**"**$_0$;t$_k$} is judged sufficiently small at the first (and only) attempt, then the statistical properties of this set of errors can be used as a measure, in effect, of prior performance validity. The qualification of "prior" signifies the performance validity of the model, for the specific, **given** data set, prior to any adjustments of the model's parameter values. In the vast majority of cases such good fortune is unlikely to obtain, and the statistics of the errors will essentially refer to a measure of the **posterior** performance validity of the model, **after** changes to the parameters values have been made on the basis of the test observation set. The extent to which in subsequent trials changes to **"**$_i$, as a result of mismatches between the model and observations, cause **e**{**"**$_i$;t$_k$} to differ from **e**{**"**$_0$;t$_k$}, will leave the concluding error characteristics **e**{**"**$_M$;t$_k$} progressively weaker approximations of the desired prior performance validity.

3 **Prior performance validity**

The raw "data" for the test of prior performance validity consist of the sequences of [**y**(t$_k$)], [**í**{**"**$_0$;t$_k$}], and [**e**{**"**$_0$;t$_k$}], where the argument **"**$_0$ signals as before that **no** adjustments have been made to the values of the model's parameters on the basis of interpretation of the test observation set, and the notation [@] signifies a sequence covering (N+1) instants in time from t_0 to t_N. It might be the case that simple visual inspection of these sequences would be sufficient for an expert judgement on the prior performance validity of the model; but this is not the subject of the present discussion (see Chapter 4).

The summarizing "properties" of these data, which are to be computed to assist in the assessment, are expressed, inter alia, as:

(i) a distribution function, i.e., the frequency with which certain values of a variable, e.g., y, are found to occur;

(ii) the moments of this distribution function, almost always, its mean and variance statistics;

(iii) the degree to which a pair of quantities are correlated, as measured by the coefficients appearing in a linear regression relationship between y(t$_k$) and **í**(t$_k$);

(iv) the degree to which the value of a variable at one instant in time, e(t_l), is correlated with its value at another instant, e(t_m), as measured by the auto-covariances, or auto-correlation coefficients; and

(v) the degree to which a pair of quantities -- typically y(t_l) and **í**(t_m) -- are correlated, as measured correspondingly by the cross-covariances, or cross-correlation coefficients.

Their purpose is compression of the properties of [**y**(t_k)], [**í**{**"**$_0$;t_k}], and [**e**{**"**$_0$;t_k}] into a single numerical value, in the same way that visual identification by the expert of a (single) pattern in the sequence e(t_k) will allow that expert to come to a judgement on the validity of the model. The crucial difference, however, is that these summary properties are arrived at in an objective, expert-independent fashion. Their computation and use are governed in part by the volume and quality of the available data; increasingly more and better quality data will in general be required as one progresses from property (i) through to property (v).

Such summary properties have two distinctive attributes: they may be based on **paired** or **unpaired** sets of data; and they may, or may not, be based also on treatment of the data in **sequence**. These attributes capture increasingly more detailed aspects of the match between the outputs of the model and the corresponding observed quantities. Thus:

(i) Properties based on **unpaired** sets of data, such as the mean and variance of the cumulative distribution functions for [**y**(t_k)] and [**í**{**"**$_0$;t_k}], overlook entirely the contemporaneous variations in these quantities. It matters only that the observed and simulated behavior occupy -- collectively -- the same portion of the output "space"; it matters not that they may, or may not, be in the same "locality" at the same time. For instance, [**y**(t_k)] and [**í**{**"**$_0$;t_k}] could have in summary essentially identical mean values and variances, yet the maximum value of y(t_k) could be observed to occur at the same time as the minimum value of **í**(t_k).

(ii) Properties based on treatment of the data in **sequence**, such as the auto-correlation function of [**e**{**"**$_0$;t_k}] and the cross-correlation function for [**y**(t_k)] and [**í**{**"**$_0$;t_k}], must also be based on paired sets of data. They encapsulate the tendency for the mismatch between observed and computed behavior to be persistently one of under- or over-estimation (at one instant in time after another) or, conversely, consistently random in time with no detectable pattern.

Both of these properties are quantitative measures of patterns relatively easily detectable by eye.

There are also minor variations on the basic theme of some of these properties, for example, in the computation of a normalized (or relative) mean error defined as

$$\bar{e}_i/\bar{y}_i \tag{7}$$

in which $\bar{e}_i$ is the mean of the errors between y_i and **í**$_i$ and $\bar{y}_i$ is the mean value of the observations y_i. As with any procedure of normalization, computing relative error properties permits intercomparison of the performance validity of the model across individual output variables y_i that may have quite different characteristic numerical scales.

Taken together, the raw "data" and summarizing "properties" listed above provide information with which a "decision" can be made. This "decision" is simply the judgement on whether to accept or reject the model as a valid instrument of prediction. And like all decisions, it would be best made in the light of as much relevant information as possible, and will be subject to a particular decision rule, or perspective of the decision-maker.

These generic properties of decision analysis devolve down to the procedures of statistical hypothesis testing (as admirably demonstrated by Reckhow et al, 1990):

(i) The decision is tantamount to answering the question of whether **y** is the same as **í** (or alternatively whether **e** is zero);

(ii) The perspective of the decision-maker is determined by his or her stance as a model developer vis a vis a model user (as already discussed in respect of the work of Burns et al, 1990);

(iii) A risk must be specified for the making of a wrong decision -- from either of the two perspectives of (ii) above, and where (presumably) the decision-maker could be risk-averse, risk-neutral, or risk-prone in his or her preferences for the weighting of the costs and benefits of the outcome of the decision.

There can be minor variations on the theme of what constitutes "similarity" (or "sameness") in the question of (i), and an array of "statistics" can be computed for comparison with standard values of these statistics in order to establish whether similarity, or conversely dissimilarity, can be said to hold -- at a chosen level of risk (iii) from a chosen perspective (ii).

REFERENCES

Ababou R, Sagar B and Wittmeyer G (1992), "Testing procedures for spatially distributed flow models", Advances in Water Resources, 15(3), 181-198.

Akaike H (1974), "A new look at statistical model identification", IEEE Trans on Automatic Control, AC-19, pp 716-722.

American Society for Testing and Materials (1984), "Standard practice for evaluating environmental fate models of chemicals", Standard E 978-84, American Society for Testing and Materials, Philadelphia, Pennsylvania.

Beck M B (1987),"Water quality modeling: a review of the analysis of uncertainty", Water Resources Research, 23(8), pp 1393-1442.

Beck M B (1991), "Forecasting environmental change", J Forecasting, 10(1/2), pp 3-19.

Beck M B and Halfon E (1991), "Uncertainty, identifiability and the propagation of prediction errors: a case study of Lake Ontario", J Forecasting, 10(1/2), pp 135-161.

Burns L A (1983), "Validation of exposure models: the role of conceptual verification, sensitivity analysis, and alternative hypotheses", in Proceedings 6th Symposium, Aquatic Toxicology and Hazard Assessment (W E Bishop, R D Cardwell, and B B Heidolph eds), Philadelphia: American Society for Testing and Materials, ASTM Special Technical Publication 802, pp 255-281.

Burns L A, Barber M C, Bird S L, Mayer F L, and Suárez A (1990), "PIRANHA: Pesticide and Industrial Chemical Risk Analysis and Hazard Assessment", Report, Environmental Research Laboratory, United States Environmental Protection Agency, Athens, Georgia.

Cale W G Jr, O'Neill R V and Shugart H H (1983), "Development and application of desirable ecological models", Ecological Modelling, 18, pp 171-186.

Caswell H (1976), "The validation problem", in Systems Analysis and Simulation in Ecology", (B C Patten ed), New York: Academic, Vol IV, pp 313-325.

Costanza R and Sklar F H (1975), "Mathematical models of freshwater wetlands and shallow water ecosystems: an articulated review", in Proceedings SCOPE International Conference on Freshwater Wetlands and Shallow Water Bodies.

Cox D C and Baybutt P (1981), "Methods of uncertainty analysis: a comparative survey", Risk Analysis, 1(4), pp 251-258.

Davis P A, Price L L, Wahi K K, Goodrich M T, Gallegos D P, Bonano E J and Guzowksi R V (1990), "Components of an overall performance assessment methodology", Report NUREG/CR-5256, SAND88-3020, Sandia National Laboratories, Albuquerque, New Mexico.

Donigian A S and Rao P S C (1990), "Selection, application, and validation of environmental models", in Proceedings International Symposium on Water Quality Modeling of Agricultural Non-Point Sources, (D G DeCoursey, ed), Report ARS-81, Agricultural Research Service, United States Department of Agriculture, pp 577-600

Environmental Protection Agency (1991), "Guidelines for Exposure Assessment", SAB Draft Final, August.

Hassanizadeh S M and Carrera J (1992), "Editorial: special issue on validation of geo-hydrological models", Advances in Water Resources, 15(1), 1-3.

Hermann C F (1967), "Validation problems in games and simulations with special reference to models of international politics", Behavioral Science, 12, pp 216-231.

Hornberger G M and Spear R C (1980), "Eutrophication in Peel Inlet, I. Problem-defining behavior and a mathematical model for the phosphorus scenario", Water Research, 14, pp 29-42.

Iman R L and Helton J C (1988), "An investigation of uncertainty and sensitivity analysis techniques for computer models", Risk Analysis, 8(1), pp 71-90.

Konikow L F and Bredehoeft J D (1992), "Ground-water models cannot be validated", Advances in Water Resources, 15(1), 75-83.

Lewandowski A (1982), "Issues in model validation", Angewandte Systemanalyse, 3(1), pp 2-11.

Luis S J and McLaughlin D B (1992), "A stochastic approach to model validation", Advances in Water Resources, 15(1), 15-32.

Mackay, D (1988), "Editorial", Environmental Science and Technology, 22(2).

Mankin J B, O'Neill R V, Shugart H H and Rust B W (1977), "The importance of validation in ecosystem analysis", in New Directions in the Analysis of Ecological Systems, (G S Innis ed), La Jolla, California: Simulation Council, Proceedings Series, 5(1), pp 63-72.

Mihram G A (1973), "Some practical aspects of the verification and validation of simulation models", Operations Research Quarterly, 23, pp 17-29.

Miller D R, Butler G and Bramall C (1976), "Validation of ecological system models", J Environmental Management, 4, pp 383-401.

National Acid Precipitation Assessment Program (NAPAP) (1990), "Evaluation of regional acidic deposition models and selected applications of RADM", Acidic Deposition: State of Science and Technology, Vol I, Report 5, The National Acid Precipitation Assessment Program, Washington, DC.

National Research Council (1990), "Ground Water Models: Scientific and Regulatory Applications", Water Science and Technology Board, United States National Research Council, Washington DC: National Academy Press.

National Research Council (1994) "Science and Judgement in Risk Assessment" United States National Research Council, Washington DC: National Academy Press.

Oreskes N, Shrader-Frechette K and Belitz K (1994), "Verification, validation, and confirmation of numerical models in the earth sciences", Science, 263, 4 February, 641-646.

Parrish R S and Smith C N (1990), "A method for testing whether model predictions fall within a prescribed factor of true values, with an application to pesticide leaching", Ecological Modelling, 51, pp 59-72.

Popper K R (1959), "The Logic of Scientific Discovery", New York: Harper.

Popper K R (1963), "Conjectures and Refutations: The Growth of Scientific Knowledge", New York: Harper.

Reckhow K H and Chapra S C (1983), "Confirmation of water quality models", Ecological Modelling, 20, pp 113-133.

Reckhow K H, Clements J T and Dodd R C (1990), "Statistical evaluation of mechanistic water-quality models", J Environmental Engineering, Proc American Society of Civil Engineers, 116(2), pp 250-268.

Scavia D (1980), "An ecological model of Lake Ontario", Ecological Modelling, 8, pp 49-78.

Spear R C and Hornberger G M (1980), "Eutrophication in Peel Inlet, II. Identification of critical uncertainties via generalised sensitivity analysis", Water Research, 14, pp 43-49.

United States Environmental Protection Agency (1989), "Resolution on Use of Mathematical Models by EPA for Regulatory Assessment and Decision-Making", Environmental Engineering Committee, Science Advisory Board, United States Environmental Protection Agency, Washington DC.

United States Environmental Protection Agency (1993), "Final report from the Agency Task Force on Environmental Regulatory Modeling", United States Environmental Protection Agency, Washington DC.

Usunoff E, Carrera J and Mousavi S F (1992), "An approach to the design of experiments for discriminating among alternative conceptual models", Advances in Water Resources, 15(3), 199-214.

Van Straten G and Keesman K J (1991), "Uncertainty propagation and speculation in projective forecasts of environmental change: a lake-eutrophication example", J Forecasting, 10(1/2), pp 163-190.

Varis O (1994), "A belief network methodology for modelling environmental change", (in preparation).

Versar Inc (1988), "Current and suggested practices in the validation of exposure assessment models", Office of Health and Environmental Assessment, United States Environmental Protection Agency, Washington DC.

Young P C (1978), "General theory of modelling badly defined systems", in Modelling, Identification and Control in Environmental Systems, (G C Vansteenkiste ed), Amsterdam: North-Holland, pp 103-135.

Young P C, Hornberger G M and Spear R C (1978), "Modelling badly defined systems - Some further thoughts", Proceedings SIMSIG Simulation Conference, Canberra: Australian National University, pp 24-32.

Zimmerman D A, Wahi K K, Gutjahr A L and Davis P A (1990), "A review of techniques for propagating data and parameter uncertainties in high-level radioactive waste repository performance assessment models", Report NUREG/CR-5393, SAND89-1432, Sandia National Laboratories, Albuquerque, New Mexico.