Quality Assurance of Environmental Models (Draft - Under Review)

Alice Shelly David Ford Bruce Beck

NRCSE

Technical Report Series

NRCSE-TRS No. 042

April 19, 2000

The **NRCSE** was established in 1996 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

EPA

Report of a Workshop on

Quality Assurance of Environmental Models

Seattle, September 7 - 10 1999

Prepared by

Alice Shelly, David Ford and Bruce Beck

National Research Center for Statistics and the Environment

Environmental Protection Agency

Table of Contents

1 Introduction 1
2 Research Presentations 4
2.1 Defining the Problems of Model Assessment and Quality Assurance 4
2.1.1 The need for assessment techniques 4
2.1.2 How can we assess models? 7
2.1.3 The special problems of complex models 10
2.2 Developments of Methodological and Quantitative Techniques 11
2.3 Assurance of Models in Environmental Regulation 14
3 Discussion Groups 17
3.1 Quality Assurance of Models and the Life Cycle of Models 17
3.2 Peer Review of Environmental Models 19
3.3 Very High Order Models 20
3.4 Tool Chest and Methodological Vision for Model Assessment 22
3.5 In Retrospect 23
4 Applicability to Model Use Acceptability Guidance 26
5 Research Needs 29
6 References 31
Appendix A : Glossary of terms and acronyms 32
Appendix B : Abstracts 35
Appendix C : Summary of Discussion Group Sessions 55

1 Introduction

This report on the Quality Assessment of Environmental Models (QAEM) Workshop is part of a larger effort by scientists, modelers, and regulatory officials to develop an exemplary process for model assessment. Mathematical models are increasingly relied upon for environmental decision making – it is essential to ensure their quality. However, uncertainties occur in ecological and environmental theories that models are based on, in choices made about how to represent a theory when constructing a model, in the data used in calibration and testing, and in how models should be applied. Furthermore, models used to describe environmental and ecological processes are becoming increasingly complex, with some including multiple media, multiple pathways, and multiple endpoints. Uncertainties that occur in construction and use of even the simplest models may be even more likely for these larger models.

Techniques are available for assessing some components of uncertainty in models, but overall assessment procedures, including how these various techniques might best be used, have not been clearly defined. As the importance and difficulties of quality assessment of models are growing, the public is becoming more aware that environmental regulations rely on environmental models. Consequently, it will become increasingly necessary to develop methodology to convey critical uncertainties in an accessible fashion.

The United States Environmental Protection Agency (EPA), through its Science Advisory Board (SAB) and various committees and working groups has been participating in an ongoing effort to ensure that environmental regulatory models (ERMs) are developed, used, and implemented appropriately. In 1994, the Agency Task Force on Environmental Regulatory Modeling (ATFERM) developed the *Agency Guidance for Conducting External Peer Review of Environmental Regulatory Models* and recommended the creation of a Committee on Regulatory Environmental Modeling (CREM) to implement a model evaluation process. However, this recommendation was specific to fate and transport models and exposure assessment models. In 1998, the Science Policy Council (SPC) issued a Peer Review Handbook (USEPA, 1998),

recommending the mechanisms, criteria, and documentation necessary for external peer review. This handbook is not specific to modeling, so many issues of particular importance to modeling are not discussed. In 1999, a follow-up "white paper" report was issued (USEPA, 1999) by the SPC outlining options for implementing current ATFERM recommendations on model acceptance criteria (MAC) and peer review. This document also leaves many questions unanswered.

The particular issues that concern model assessment, raised in the context of ERMs, are pertinent to a wider group of environmental models being developed and used by scientists and environmental managers. The EPA and the National Research Center for Statistics and the Environment (NRCSE) organized the QAEM workshop for two reasons. First, EPA wished feedback on recommendations for model assessment being developed within the agency. Second, it is important for the academic/scientific community to work with the EPA in the development of a recognized process for model acceptability.

The QAEM workshop brought together an international group of experts on model assessment: modelers, regulators, statisticians, and scientists with interests in the assessment of process-based deterministic models, to discuss the large and diverse topic of assessing the quality of environmental models. The conference was organized in three presentation sessions and three discussion sessions. The presentation sessions were:

Day One: Defining the Problems of Model Assessment and Quality Assurance

Day Two: Development of Methodological and Quantitative Techniques

Day Three: Assurance of Models Used in Environmental Regulation

Discussion topics were determined on the first afternoon of the conference, and were discussed by two concurrent discussion groups on Wednesday and Thursday. These were:

Tool Chest and Methodological Vision for Model Assessment

Peer Review of Environmental Models

Quality Assurance of Models and the Life Cycle of Models

Very High Order Models

The final morning of the conference was used to re-cap and expand upon the discussion group findings.

This report includes summary descriptions of presentations and discussions from the QAEM workshop in sections 2 and 3. In section 4, we tie the workshop findings to the EPA regulatory process. In section 5 we highlight research needs that were apparent from workshop discussions. Appendices include a glossary of acronyms and terms, workshop abstracts, and detailed summaries of the discussion group topics.

2 Research Presentations

The types of models discussed at the QAEM workshop are called *process*, or sometimes *deterministic*, models. Their essential feature is that they attempt to simulate the operation of ecological and environmental systems based, at least in part, on theory about how the system functions. Because such models are developed from theories of system function, it is not surprising that the scientific literature on assessment of such models is divided among the different disciplines producing and using them. Consequently, issues of general importance are frequently presented in the light of specific examples or environmental problems and may not receive wider discussion. For example, the idea that repeated tests of a model against data sets does not provide *validation* was discussed in a Special Issue of Water Resources Research (e.g., Hassanizadeh and Carrera 1992). Understanding of the wide range of processes involved in uncertainty, certainly extending beyond tests of a model against specific data sets, is an integral part of risk assessment in policy analysis (Morgan and Henrion, 1990). The idea that there is no absolute standard for model assessment but that a standard must be set relative to the objectives of a particular regulatory requirement has been discussed by scientists involved with modeling aspects of exposure assessments (e.g., Beck et al. 1997)

2.1 Defining the Problems of Model Assessment and Quality Assurance

2.1.1 The need for assessment techniques

A standard method for model assessment, involving first calibrating a model using one data set and then testing that parameterization against a second data set, has proved to be wholly inadequate, even for simple models with ample data. Dr. Naomi Oreskes and others in the first session of the workshop described the many reasons why this is so. For example, a test of calibrated models that is based on model predictions is a measure of how a model may accommodate to data. Where there is intrinsic variability

in data, as there is in data from environmental systems, then deficiencies in model structure can be overcome by flexibility in parameter estimates – the process of accommodation. As the number of parameters in a model is increased, accommodation can become easier to achieve. Also, models in a regulatory setting are most often used for forecasting future conditions, but there is no guarantee that a model that has successfully predicted an independent data set will succeed in predicting future conditions. Ecological and environmental data are realizations of stochastic processes – the environment from which we collect data is ever changing. Properties of ecological or environmental systems may subsequently be found that were not detected when a model was first constructed. Changes in the forcing functions of the system, e.g., due to climate variation, may result in new variables having an important influence. These factors combine to ensure that the implicit assumption of stasis necessary for predictive models can never be completely realized.

Testing the predictive abilities of a model also falls short in terms of testing the assumptions and theory that were used to arrive at the predictions. There are many assumptions involved with every modeling exercise, not all of which may be apparent when the model is constructed. Auxiliary and simplifying assumptions may be blind choices when evidence on either side does not convince, or they may be choices with strong evidence in their support. Some choices about model structure may be borne of necessity, e.g. there is not enough computing power to develop a model with comprehensive representation of a particular process. In still other cases, some assumptions may not be explicitly considered in the modeling process at all – the most expedient technique or model form known to the modeler may be selected. Thus, the theory and assumptions underlying the model must be a part of model assessment.

With the above factors in mind, Dr. Oreskes and others urged caution in terminology. No matter what type of assessment is used, terms like *validation* or *assurance* imply affirmative confirmation that is not warranted. We can evaluate, examine, assess, audit, review, and scrutinize models – and detail their strengths and weaknesses – but we can never assure they are "*valid*" in the strict sense of the word. In fact, if such assurances are claimed to the public and the model fails, credibility is lost.

Several examples presented at the workshop gave clear indications of the limitations of current model assessment methods, particularly for complex models. Dr. Robin Dennis provided an example from air quality modeling in which interaction between two parameters changed the model outcome and interpretation in profound ways. A model predicting stratospheric ozone concentration over a region of the eastern United States appeared to function well. However, when the geographical region of interest was extended model performance declined. In the initial case the concentrations of nitrous oxides had been important driving variables. When the model was extended in its application it was found essential to model effects of volatile hydrocarbons on ozone concentrations. This detail can have huge ramifications on science regulatory policy, but would not be discerned using standard model assessment processes applied to the initial model.

To prevent these types of problems, Dr. Dennis proposes that diagnostic evaluation and sensitivity analysis are two of the critical components of adequate model evaluation. The goal of diagnostic evaluation is to develop tests of the model's realism (outputs in relation to data) relative to its intended application. Sensitivity analysis assesses the behavior of the full model as a system. The sensitivities of interest must be the decision made from the model application, rather than simply the model-predicted outcome variable.

Dr. Ray Whittemore presented an example of a complex modeling system that has not undergone thorough assessment prior to being released to, and employed by, the public. The BASINS (Better Assessment Science Integrating Point and Nonpoint Sources) system integrates geographic information system (GIS), national watershed data, and all relevant watershed scale models into one software package. In its current form, problems with BASINs include a lack of adequate data Quality Assurance (QA), incompatible data types, and inadequate training for users (both EPA scientists and the public). Automatic application of BASINS, with no warnings about the inaccuracies that can occur, can lead to misuse and misunderstanding. As BASINs continues to grow in scope and complexity, the potential for inappropriate usage and errors will dramatically increase. To prevent this, Dr. Whittemore proposes that QA of model components,

data sets, and model coding and a full assessment of individual component models be undertaken.

2.1.2 How can we assess models?

It is clear both from practical experience and on theoretical grounds that accurate predictions do not mean that a model is correct. Tests using prediction may be considered necessary but they are certainly not sufficient for model assessment. A more comprehensive approach must be taken that should focus on evaluating the explanatory capacity of the model – what it explains, and how well, versus what it does not explain.

Dr. Iris Goodman used examples from her work with landscape models to highlight different levels of modeling efforts that may require differing intensity of quality assurance. For example, models have different ecological domains, different spatial scales, and different time scales. Purposes for models include: ranking vulnerabilities, comparing and quantifying risks, predicting current conditions, and guiding environmental restoration. In addition to being adaptable to these different objectives and types of models, model assessment procedures will have to include answers to issues that have not been fully examined such as:

- Appropriate matching of modeling approach to scientific or management objective;
- Model parameterization using remotely sensed data;
- Assessment of models for applications in which supporting empirical data is not available; and
- Determining the importance, rather than the statistical significance, of model results, relative to uncertainty in model simulations.

Dr. Helen Dawson described how an initial review suggested that little formal Quality Assurance was being applied to models used in her EPA program, or by contractors associated with that program. She also described problems of evaluating performance

and reliability of intermedia transfer models used in probabilistic human health risk assessment.

A wider approach to model assessment is required than simply tests against data. Dr. David Ford suggested the approach should examine the scientific inference that can be made about models relative to their intended purpose. There are two requirements: to define *explanatory capacity* of a model – what it tries to answer and what theory it uses to do that; and to determine the *explanatory coherence* – how well the theory used in a model is able to answer the question. The explanatory capacity of a model is defined by the following:

(1) The strength of the theory on which the model is based. While there are no formal procedures for assessing a theory represented by a model, the following aspects should be examined:

(a) Note where the theory, and model constructed from it, is *incompletely specified*. In many theories, if not most, there is something not understood, i.e., there is uncertainty about model structure.

(b) Note where relationships are *underdetermined*, i.e., there may be reasonable alternatives that have not been ruled out by specific data based investigations. Competing sub-theories and imprecise definitions of ecological concepts can be missed or ignored.

(c) Note the *domain* of theory being used. The inductive use of models, that we fit them under one set of conditions and then attempt to use them in another, is the fundamental reason why validation is inadequate. Careful specification of the domain structure of a theory and model is essential.

(2) The quality and quantity of the empirical database. Problems associated with data are those most frequently considered in model assessment projects. However, these operate within a framework set by the problems listed in (1).

The second requirement is to define and evaluate the *explanatory coherence* of the model as follows:

(1) The scientific explanation required must be specified. This details the purpose for the model, specifies the topic to be modeled, defines contrasting alternatives, and defines the relevance of the model. Specifying a scientific explanation requires more focus than specifying a theory – it places bounds on the knowledge required according to purpose.

(2) The explanatory coherence of the theory and the model should be evaluated by answering the following questions:

(a) How acceptable are the individual propositions when tested against data?

(b) Are concept definitions consistent throughout the theory and model network? Theories frequently use concepts that can be difficult to quantify. Sometimes these problems are well known and are the subject of detailed investigations, such as *forest health* or the *biological integrity* of a stream. However, there can be concepts in the working detail of models that require careful translation, and the particular assumptions required should be specified.

(c) Are part and kind relationships consistent throughout the theory network and model implementation? Typically we construct process models by quantitative representation of different *parts*, and how they interact, e.g., the different plant or animal species in an ecosystem response model. If these species respond in the same way they are the same *kind.* Model construction requires decisions about what should be aggregated and what should not, and these decisions should be made using the same part and kind relationships as the theory is based on.

(d) Are there any ad-hoc propositions? These are statements that are added, or excluded, without a strict background in the theory of the process under investigation, in order to explain some particular feature or

to tie pieces of the model together. Ad hoc propositions can also be used to excuse a model from having to explain something.

2.1.3 The special problems of complex models

Dr. Robin Dennis noted, "The leading edge of environmental modeling has been pushed towards more complex process-based multimedia science models as our understanding of the network of interrelations in the physical environment grows and computational barriers fall". There is some need for caution as this trend continues. Highly complex models are neither unique nor transparent in their construction, but they can become irrefutable since small modifications will often fix "failures" without necessarily improving the quality of the model.

Several workshop participants presented examples of very high-order models (VHOMs). Dr. McDonnell presented the results of an effort to model human responses to ozone exposure as a function of ozone concentration, duration of exposure, minute ventilation, and age. Dr. Richardson presented a modeling quality assurance plan which could be viewed as a QA prototype, particularly for projects that involve holistic, multi-media modeling approaches for large systems like the Great Lakes. He noted that current guidance for model QA includes guidance for computer code, data validation and calibration, peer review, and differences between models for research and models for regulation, but it lacks guidance for complex multi-level linked modeling projects. Also, there seems to be a lack of scientific perspective.

Dr. Meiring discussed the challenges involved in evaluating the space-time structure of environmental model results with that of observations. For example, there may be differences in the spatial and temporal scales of the model versus the data, due to the expense of both running models and environmental sampling. In addition, the environmental processes may be non-stationary in space and time due to topographical influences, pollutant emissions, chemical reactions, and transport processes.

Conference participants discussed the special problems of VHOMs in an afternoon discussion group. See Section X for details on the outcome of that discussion.

2.2 Developments of Methodological and Quantitative Techniques

In general, current techniques for model assessment are techniques for quantifying uncertainty, and they are classified under two broad headings. Uncertainty Analysis (UA) can be defined as the process by which parameter uncertainty (lack of knowledge of true values of parameters) in a model is described and quantified. Sensitivity Analysis (SA) can be defined as the process by which the consequences of uncertainty are explored. Using SA, we ascertain the impact (on selected model outputs) of different types of uncertainty. Some methods discussed at the QAEM workshop were easy to classify as either UA or SA. Others were more general and could not be clearly assigned one of these classes. Both frameworks are important in model assessment, but a larger unified approach is necessary to adequately address the problems discussed in the previous section.

There are many types of uncertainty in the modeling process. Dr. Tony O'Hagan provided this list, which he states is not exhaustive:

- Parameter uncertainty: Lack of knowledge of the true values of parameters.
- Model inadequacy: The uncertainty of the underlying theories and assumptions of the model.[1]
- Intrinsic Variation: The variability in the underlying real-world process. The lack of stasis in the natural world.
- Observation Error: The inadequacy and imperfection of measurements.
- Parametric Variation: The parameter itself is something that is not constant, but varies within a distribution.
- Code Uncertainty: An artifact of inaccurate parameter calibration. For simple models with ample data, this can be reduced to negligible amounts. For highly complex models, or for situations where data are very rare, this uncertainty can be very high.

[1] Epistemology is the study of how we find things out.

Dr. O'Hagan's presentation focused on model calibration, and he suggested that residuals from the best fitting calibration provide a measure of model inadequacy (this was also discussed by Dr. Saltelli). He also raised the point in later discussions that some uncertainties are difficult to enumerate and require expert judgment. In his experience, expert judgment is extremely difficult to elicit in the tails of parameter distribution (e.g. deciding how wrong it could be, or beliefs about rare events), and in regards to correlation between uncertainties (how does uncertainty about one parameter effect knowledge of another.) The correlation between parameter values in the model is often not taken into account.

Dr. Adrian Raftery presented an example of a technique for uncertainty analysis known as Bayesian Melding. Using this method, deterministic simulation models can be put on solid statistical footing, if the determinism is viewed as a simplifying assumption (like linearity). Bayesian melding is an expansion of the common Bayesian parameter fitting approach, in which the modeler uses professional knowledge about inputs to form the prior distribution, data to form the likelihood, and prior information (based on data) about the outputs to update the input distributions. The sampling importance resampling (SIR) algorithm (Rubin, 1987) can be used to obtain the posterior distribution, by weighting the input distribution based on model outputs. This approach removes some of the reliance on particular data sets in the calibration process.

In addition to the problem of data accommodation, the choice of output measures and application determines the path of theoretical emphasis of model results. Dr. Joel Reynolds discussed a technique to judge a model based on its performance on multiple output criteria. The Pareto Optimal Model Assessment Cycle (POMAC) is a multiple criteria model assessment methodology for exploring uncertainty in the relationships between process conception, model structure, and assessment data. POMAC is used to assess the conceptualized model structure based on multiple outputs, rather than the usual procedure of optimizing performance with respect to a single criterion. The first step in applying POMAC is to develop the key criteria or characteristics by which model performance can be evaluated. Next, the criteria are applied to all parameter settings to test if any of the resulting model outputs meet all of the criteria, and if not, why not? POMAC can then be used to uncover the source of model deficiencies.

Dr. Andrea Saltelli presented a Bayesian framework for model calibration and correction integrating sensitivity and uncertainty analyses. He suggests that the framework of modern sensitivity analysis can be expanded to include most aspects of model building, specifically:

- Question formulation - Is the model capable of delivering the answer to the question?
- Model conceptualization - What factors to include/exclude based on consistency between the model and the system being modeled.
- Model relevance - Do the answers vary in response to model inputs?
- Is the model consistent with observations within the realm of their uncertainty?
- Discrimination between models.
- Selection of most important factors – an aide in planning calibration
- Distinguishing between different types of uncertainties.
- Identifying critical areas of model outputs.
- Prioritizing research and investment to decrease uncertainties
- Quality assurance – the application of SA can lead to discovery of coding and other types of errors.

Dr. Saltelli emphasizes that uncertainties arise in all aspects of model formation, but these uncertainties may vary in their level of acceptance and in their impacts to the final result. Sensitivity analysis evaluates the quality of models by identifying unimportant variables, providing a metric for model comparisons, and by allowing the analyst to test the sensitivities of the models to see if they match the biological framework of the problem.

There was an undercurrent at this workshop dealing with the perception of the public when it comes to mathematical models and their uncertainties. Dr. Jan Rotmans notes that one problem with standard presentations of model results is that there is often no prioritization of critical uncertainties, the types or sources of uncertainties are not well-

explained, and/or the model is not comprehensible. We must acknowledge that science is not truly objective, and that knowledge does not equal truth or certainty. There are many degrees of variability in every system, which lead to ignorance and indeterminacy, which in turn leads to incomplete or lack of knowledge. In addition, parameter values are not likely to be wholly independent. For example, a person who looks at the world a particular way expects a certain bias on parameters, which is applied across the board. In a less sociological sense, climate change is likely to affect all parameters in a certain, unified way. Thus, we need to assess parametric uncertainty in a unified way in order to uncover true uncertainty in model results.

Dr. Rotmans' multi-perspective evaluation of uncertainties can be useful to identify the full range of possible outcomes from a model. After evaluating the technical aspects of a model, the critical uncertainties and their priorities must be determined. Next, the uncertainties are categorized based on sources and types. Then, alternative models are devised, based on extreme cultural theory perspectives (i.e. egalitarian, hierarchist, and individualist). These alternative models are calibrated and evaluated in standard ways, then compared based on uncertainties to the original model UA/SA. This integrated assessment of models, scenarios, and uncertainty analyses can be used to translate scientific insights into decision-making.

None of these approaches has provided a unified approach to quantifying all sources of uncertainty, although all speakers hinted at the potential and movement in this direction. This is the next challenge for researchers in model assessment.

2.3 Assurance of Models in Environmental Regulation

All speakers in this session appeared to agree that model quality assurance should refer to a process that is adaptable to different situations. Dr. Barnwell discussed the USEPA draft protocol for model validation guidance, indicating that there are three aspects that shape the type of process needed for model assessment:

- The nature of the predictive task to be performed;
- The properties of the model; and

- The magnitude of the risk of making a wrong decision.

Dr. Kirkland provided a look at the history of the model quality assessment process within the USEPA. She emphasized that the agency is taking a broad view of model evaluation, which includes addressing the following five areas of model uncertainty:

- The theory upon which the model is based (i.e. is the theory appropriate, relevant, and correctly applied);
- The translation of theory into mathematical representation (i.e. are the functions and parameter derivations correct, have alternatives and the propagation of errors been examined);
- The transcription of the mathematical representation into computer code (i.e. software quality assurance);
- The assignment of parameter values and calibration (i.e. data quality assurance); and
- Model test choices and implementation.

The process of model assurance should be adaptable for use on models with varying objectives and possibly at different stages of development or complexity. It may be necessary to use models of different status in regard to what is known of underlying theory and/or available data. The standard of assessment should be specified, but not to the point of restricting tailoring of the process to specific uses.

We can expect the demands and standards for model assessment to change with the questions being asked. Dr. David Stanners provided a perspective from the European Environment Agency, and shared their experience in preparing a report on status and projections in environmental indicators for the European Union. This summary effort relied heavily on model outputs, but there has been no consistent effort of quality assurance applied to the models. Working with existing data and information to draw immediate conclusions was, and is often necessary for policy makers. Although the assessment of uncertainty and sensitivity of these models is important, in the end the quality has to fit the purpose. A very practical approach is therefore needed to remain relevant and useful to decision makers and to ensure that the best available information

is put to use and not ignored. Dr. Stanners presented a general QA checklist for efforts linking results from multiple models:

- Are the data correct and sound? Are there gaps in the data?
- Are the assumptions relevant and reasonable?
- Has each independent model been validated and verified?
- Is the suite of models reasonable and are independent models compatible?
- Is the analysis consistent?
- Are the results and findings reasonable?
- Do the results cover the right issues and areas (relevance)?
- What are the problems framing the participatory process?
- Are the results interpreted and used correctly?

This general checklist can be implemented with different levels of detail, depending upon the importance of the model purpose and the time frame allowable for quality assessment.

3 Discussion Groups

Within the structure of the Workshop as a whole, the first opportunity for extensive discussion came in the afternoon of the second day. It followed, therefore, the broad-ranging presentations of the first day, with then the second day's focus on questions of quantitative (statistical) methods, in which latter – as evident in the foregoing – the discussion had gravitated towards issues of model calibration, the interpretation of data, and the reconciling of models with data.

3.1 Quality Assurance of Models and the Life Cycle of Models

Given the absence of the decision-making context from the preceding presentations, the group opened its discussion by re-affirming the purpose of the Workshop as being that of QA in the service of better environmental decision-making (including regulation), in which, in particular, computational models play a central role. The issue, so the group considered, was therefore that of assuring quality for the Process (with a capital "P") of developing and applying computational models trained on a specific decision-making task. Improving the scientific content encoded in the computational models is a part of this Process, an inner loop, perhaps, which does not always have to make specific contact with the public discourse concerning the decisions to be made. In the future, this process will have to be much more transparent, open and responsive to the public, with the public even being involved in it much earlier – at the stage of developing the model. What is done technically and scientifically in developing and applying the model will have to be defended in public. Many of the things done by scientists are not simple, however. These complexities may well be germane to the public discourse and should not be rendered overly simplistic merely for the purposes of serving the needs of being listened to by the public. The "Right to Know", it was thought, would be uppermost in the public's mind.

Development of the model (and therefore any procedure of model QA) is embedded within the broader process of mobilizing science for the purposes of providing input/guidance to the making of decisions about protection, preservation, and management of the environment. The system of QA we are seeking must cover all the elements of this entire process. The state of the science available, however, may be

far from free of dispute (or, for that matter, free from the subjective values and opinions of scientists). Different elements of the Process may be relevant to different problem-solving contexts; and this may mean that the elements of a "complete" QA may be context-dependent. That is to say, judgments about QA in a context where we have few empirical observations, speculative theories (possibly in conflict) and therefore highly uncertain models, and potentially massive consequences of wrong decisions, may have to differ – in the way they are reached – from contexts in which data are plentiful, there are many parallel problem situations, and low negative consequences of poor decisions. What matters, then, is this: (a) what are the elements of the Process of developing and applying a model for a given task of decision-making (regulation); (b) what are the current procedures of QA already in place and are they adequate; and (c) what new methods/procedures of QA must we develop and put in place?

In answering the first of these questions, the group came up with a prototypical Process. It had self-similar nested loops/steps. Starting from the "inside" (the tactical) and working "outwards" (towards the strategic), it contained – within the phase of "Model Development" – the following: (i) conceptualization; (ii) system requirements (specification of equations); (iii) design phase (code flow diagrams); (iv) writing of source code; and (v) implementation, i.e., internal, but independent, checking of the source code. Other than for the first of these steps, which might well relate to Dr David Ford's concerns regarding the testing of theory, the group took the view that appropriate procedures of QA were already in place, although they were not always adhered to. "Model Development" itself forms but one step within a somewhat wider process (set out by Linda Kirkland): (i) development; (ii) collection and use of data in support of model development; (iii) acceptance of the model – as legitimate for its intended purpose – by EPA's Program Managers, with consequent release into the public domain for use; and (iv) application. Effective QA procedures are not yet in place for items (iii) and (iv). These largely scientific and technical steps in the Process overlap with some of the components of the wider definition of the Process set out by Dr. David Stanners (as discussed above in Section 2.3), yet they are also subsumed within this wider scheme.

The group's discussion brought thus the deliberations of the Workshop to the point at which it could be concluded that a shell of the constituent steps in the Process of mobilizing Science (through a model) for the purposes of environmental decision-making was in place, with some impression of the state-of-the-art of QA procedures for these steps, but no specification of, or speculation on, how to improve these procedures. In this respect, there might be much to be learned from existing analogs of the Process required for model QA, e.g., the legal process, QAs for tools, and chemical laboratory analysis. QA/QC for the procedures and methods of analytical chemistry, for example, are devoted to "consistency" and "error analysis". What is defensible is judged in several dimensions, defined as: scientific, technical, and legal. Yet a one-to-one mapping of ideas from these other (analog) fields is not entirely satisfactory. Some aspects of the problem are unique and will demand of model builders and model users that we think through for ourselves what must be done. Many actors may participate in "touching" the model, as it passes through its life in the service of better environmental decision-making. For instance, there are: (i) model developers; (ii) peer model developers; (iii) CREM; (iv) program managers; (v) technically literate users; and (vi) other stakeholders. How then are the experiences of the various actors to be orchestrated, i.e., fed into the Process of QA, with the goal (again) of making better environmental decisions?

3.2 Peer Review of Environmental Models

With so much of the Workshop having been focused on the discussion of technical and philosophical issues, it was salutary to be reminded of the importance of the process of peer review. Some things are extremely difficult to deal with in quantitative terms, most notably the quality of the constituent hypotheses and the "pedigree" of the process mechanisms incorporated into the structure of the model. Whether these many components of the model have been endorsed by overwhelming consensus in their choice, or are strongly disputed, or considered highly speculative, are factors material to QA and ones ideally suited to the process of peer review. Indeed, they are especially important the more difficult it becomes to evaluate the model's overall performance against field data, and field data (in turn) will be particularly hard to acquire for models of a multi-media character.

This second discussion group listed all the customary aspects of the ideal role of peer review but, in addition, drew attention to its role in improving communication among the various disciplines contributing to the development of the project, which will usually be several in the multi-media case. They noted that peer review was a matter entirely separate from soliciting public comment on a model – and we have already seen from the foregoing discussion group the importance of the public-science connection. The group was quite specific in setting out the steps of the peer review: (i) allocate budget for review and ensure review is integral to the work plan for developing the model; (ii) select reviewers; (iii) express modeling team's expectations of peer review; (iv) schedule receipt of documents and associated meetings; (v) provide institutional support to reviewers; and (vi) publish both the review and the modeling team's response in an open and easily accessible medium (nowadays, the internet, for example). Among several recommendations the EPA's Science Advisory Board (SAB) was cited as the kind of centralized institution required to provide proper guidance and oversight for the process. For the past two years the EPA has in fact maintained a Subcommittee on Environmental Modeling, most of whose work is devoted to issues of QA. The Subcommittee has itself been employed in a prototypical process of acting as a sounding board from the very early stages in the development of the Total Risk Integrated Model (TRIM) framework.

3.3 Very High Order Models

From the perspective of QA, the form of VHOMs developed and presented by Dr Robin Dennis (Section 2.1.1), for example, can readily be appreciated as being uniquely complex, highly dependent upon multi-disciplinary knowledge bases, extremely difficult to scrutinize, and doubtless strongly immune to empirical refutation. Furthermore, they almost defy the application of any adequate peer-review process, within time and cost constraints and with a sufficient number of peers having no "conflict of interest". Their uniqueness and completeness make it highly likely that the entire group of qualified peers will be drawn into the development of the model, in one way or another. In terms of the conventional measures of history-matching and peer review VHOMs therefore present enormous challenges to the task of developing appropriate protocols for QA. In the light of these distinctive problems the group acknowledged that the traditional,

partisan positions of being philosophically “for large models” or “against small models”, should be shed in favor of fashioning a common cause, not least by reaching out to quite different fields (of Integrated Assessment Modeling, and QA procedures for the computational schemes employed by NASA or the Boeing Corporation). Ever larger models will continue to be built. It will not suffice to be thrown into a state of mental paralysis on the issue of their QA. But neither will it be sufficient to argue that the quality of a VHOM has been assured simply by virtue of every conceivable constituent hypothesis having been a priori incorporated.

The group came to the view that there was a critical need for a rational procedure for modular assembly of VHOMs, such that the power of the QA protocol could be exercised at this modular, possibly mono-disciplinary, level. The exercise of peer review might require a very long “institutional memory”, however, with the consequent danger of reviewers being invited in due course to criticize the validity of their own earlier suggestions for modifications of the VHOM under scrutiny. In addition, given the very few scientists qualified and sufficiently experienced with a particular VHOM to have acquired a strategic “bird’s-eye” view of the entire model, it is possible that some elements of the QA process could become very sensitive to the credibility and personal integrity of these scientists. In any case, it was felt the imperative would be for a high-level, conceptual description to be developed for each VHOM, such that this succinct, easily comprehensible image of the quintessential workings of the model could then be used to guide its disassembly and interrogation. For example, using the image of a conveyor belt in order to understand the circulation of heat within the North Atlantic can be vital to mounting an evaluation of the underlying sets of partial differential equations and assumptions employed to simulate the associated physical oceanography. There is therefore also a need for a rational procedure of disassembly, scrutiny and interrogation of the underpinning details. The latent mental models – or high-level conceptual descriptions, or again the “auxiliary assumptions” mentioned in Dr. Naomi Oreskes’s presentation – must themselves then be expressed and subjected to challenge and scrutiny.

Field data for evaluation of VHOMs would be of a non-routine nature, necessitating a much better dialog between those involved in developing the model and those at the

forefront of monitoring technologies. Some of the data to be collected would clearly have to be specified by the high-level conceptual descriptions used to structure interrogation and evaluation of the VHOM. The methods of sensitivity analysis and uncertainty analysis would also somehow have to be made computationally routine at this large scale.

3.4 Tool Chest and Methodological Vision for Model Assessment

The need for cost-effective tools and procedures for the practice of model QA can be demonstrated by example. The Clean Water Act requires the setting of Total Maximum Daily Loads (TMDLs), for as many as 17,000 bodies of water across the nation. Before building consensus on the TMDLs, assessment studies of all these segments of the aquatic environment must be undertaken, and will typically involve the application of some form of model. The National Council for Atmospheric and Stream Improvement (NCASI) has calculated that implementing the assessments alone will cost upwards of $4 billion (as a conservative estimate), with very much more money being required for the restorative and protective measures themselves. The costs of making wrong decisions, based on the outcome of a model, even in this relatively mundane context, are impressive and beg the question (as Dr Ray Whittemore would put it): when is a model calibration good enough?

This question reveals much about how we have usually viewed the process of model QA, as a matter of matching history, admittedly with a particular purpose in mind for the model when found to match past observations sufficiently well. In addition, we can today point to textbooks on the subject of Independent Verification and Validation (IVV) of software, indicating the very considerable previous experience consolidated as conventional, taught material. The third discussion group, however, also had the charge of seeking to go beyond summarizing the techniques previously in widespread use for model validation, such as the embedded techniques of model parameter estimation and local sensitivity analyses (SA), for instance. It was noted that significant advances have been made in the methods of global SA in the past five or so years, creating thus the prospect of a substantially enriched tool chest for the conduct of QA, provided suitable computational realizations of the associated algorithms are available

(which indeed is now the case). Communication of the results of model QA to a wider audience, beyond just the professional developers of models, was identified as an issue in no way to be under-estimated. In short, one might observe that the tool chest is rather replete with methods of analysis. The challenge may be shifting away from developing more of these forms of tools towards that of channeling all the outcomes of their analyses into a summary judgment on the trustworthiness of the model. How does one weave together evidence from an analysis of uncertainty, the matching of history, the statistical properties of sets of residual fitting errors, and a peer review, in order to move forward in using the model for the task at hand?

The group generated some healthy tension in its discussion, between placing the entire issue of model QA within the framework of a Bayesian decision analysis, from which subjective (or cultural) perspectives are largely excluded, and casting it alternatively in the context of Post Normal Science, wherein a variety of value-laden perspectives is tolerated. Here the term "subjective" has a much wider connotation than merely that attaching to the choice of subjective probabilities and the recognition of risk-averse, risk-neutral, and risk-taking decision-makers. There were strong echoes in this discussion, then, of the earlier presentation of Dr. Jan Rotmans (section 2.2).

3.5 In Retrospect

Validation, evaluation, scrutiny – whatever the Process is to be labeled – is a profoundly difficult issue. It is a tangle of philosophical, cultural, ethical, legal, technical, and mathematical (statistical) considerations; and has long been the object of intensive scrutiny itself. It is not at all easy, therefore, to bring a fresh perspective to bear on the subject.

Much of the previous discussion of the subject has been tied to interpretations of history-matching and peer review, from which judgments on the worthiness of the model – relative to some task, of providing guidance in the making of a decision – have been absent. This "disconnection" – of the model not being embedded in a process trained ultimately on the making of a decision – was largely banished from the Workshop's discussion. That is progress.

Over the past decade the debate has also been shifted in its philosophical basis. Where once the common modeler (practitioner) sought to validate his/her model, we have come to appreciate the consensus of science moving forward through refutation – or invalidation – rather than validation. Rather swiftly thereafter it became obvious that this, invalidation that is, is a principle governing growth in our accepted, legitimate knowledge base. It is not a principle to be applied in assuring that the quality of environmental decision-making, where this is founded on the use of a model, is continually improved. In the past decade too, the notion of a model as a truth machine has given way to the image of a model as a tool of policy forecasting. Thus, we can begin to conceive of judging the quality of a model on the basis of whether it is well designed, i.e., well- or ill-suited to fulfillment of its intended purpose (of assisting in the making of a decision). Similarly, coming to a judgment on the trustworthiness of a model has shifted from the strictly statistical domain of hypothesis testing to something more akin to the process of legal debate, in which approval or disapproval of the suitability of a model for a particular purpose is governed by the weight of evidence on one or the other side of approval.

Technically, the association of uncertainty with the development and application of a model has become widely accepted, from the cradle to the grave: we construct models from uncertain theory; we must reconcile these models with uncertain observations; our predictions, as generated from the model, are therefore uncertain; we must make decisions in spite of this uncertainty; or, if the uncertainty is such as to render us impotent in this sense, we seek to rank the sources of contributing uncertainties, thence to attack and reduce the most critical of them. But all of this analysis of uncertainty took place conventionally in the tidily quantitative domain of structural error, parametric uncertainty, random variables, stochastic processes, and so on. Within the past few years we have come to acknowledge the many other forms of uncertainty entering into the process of employing models in providing support for the making of decisions, in particular, those uncertainties having to do with differing personal views on the man-environment relationship (the perspectives of Cultural Theory, which were apparent in the presentation of Dr. Jan Rotmans). We have too (at last) the luxury of professional philosophical enquiry being brought to bear on the development and use of models in Environmental Science – witness the presentations of Drs Naomi Oreskes and David

Ford in this particular Workshop. In Environmental Science the problems we face are different from those of Physics, traditionally the subject of choice for philosophical enquiry into the business of the scientific enterprise.

Last, but not least, given that the provision of data has always been subject to various forms of QA/QC (typically those applied in the procedures of the analytical laboratory), and prompted now simply to recall that computer models are prolific generators of data, we have come to the realization that models too – or rather the procedures of bringing science to bear on the making of environmental decisions (in which both measured and computed data are pitched into the discourse) – should be subject to similar procedures.

In short, we are armed now with much more – in particular, a wider palette of metaphors and analogs (the legal process, the design of tools, the control of procedures in an analytical laboratory) – with which we ought to be able to fashion a more adequate sense of the QA Process. And, at a time when the public is to be much more involved in this Process, we have been made aware of the role of cultural perspectives in the science of Environmental Modeling. We have been sensitized to a much wider definition of uncertainty.

Ten years ago, had one been asked, validation might have been defined as the assessment of a model's predictive performance against a second set of (independent) data given parameter values identified from a first set of data. It is perhaps only when one looks back to this that one can illuminate how much matters have now changed, not least through this Workshop.

4 Applicability to Model Use Acceptability Guidance

The "white paper" report issued by the SPC (USEPA, 1999) was designed to address six pertinent questions regarding model quality assessment. General answers to these questions from a survey of USEPA processes were provided in the report. This section will provide feedback from the QAEM workshop to the EPA, addressing each of the questions from the white paper.

How do the issues of peer review and QA/QC relate to acceptability determination?

Response in White Paper: Models require both peer review and QA. Although EPA has issued general guidance on these procedures, it is not being consistently followed within the agency. The white paper adds recommendations for further implementation of these processes, and requests clarification for requirements of the Peer Review Handbook.

Response from QAEM Workshop: Peer review is very important to the process of model assessment, but it is not in itself sufficient. Efforts from one discussion group outlined important elements and requirements of a useful peer review process (see section x). QA/QC is a well-developed methodology most applicable to software development and data testing, both of which are key elements in model development. However, the strict rule-based structure of current QA/QC methods does not lend themselves to the broad range of topics that must be examined when assessing complex environmental models. Discussions at the workshop indicate that a flexible process, which is somewhat transparent, is needed.

What is the consensus definition of model use acceptability criteria? (I.e., what is being done right now within the agency?)

Response in White Paper: Techniques are not consistent throughout the agency, with differing levels of sophistication in different divisions. Also, the types of models used vary from the most basic to large integrated model systems for which no guidance on quality exists. Thus, the white paper recommends a model evaluation *strategy*, which could be adapted to specific uses, rather than a set of inflexible criteria.

Response from QAEM Workshop: Not Applicable.

Does acceptability correspond to a particular model, or specific applications of a model?

Response in White Paper: Both approaches can be accommodated by the model evaluation strategy.

Response from QAEM Workshop: The consensus of the workshop is that the quality of a model can only be viewed in relation to its intended application. That is, a particular model can not be deemed acceptable for any and all applications.

Does acceptability cover only models developed by the EPA, or can it cover externally developed models?

Response in White Paper: All models used in agency regulatory decision making are covered.

Response from QAEM Workshop: The discussions at the QAEM conference centered on environmental regulatory models, although many topics were applicable to other modeling applications.

Does acceptability mean the agency will develop a clearinghouse of models that meet EPA's definition of acceptable?

Response in White Paper: Not directly, but there will be efforts to share information to avoid duplicating efforts.

Response from QAEM Workshop: Some workshop participants expressed interest in some sort of inventory of regulatory models and the methods used to assess their quality. Initially, this would serve mainly as a set of examples of QA processes. In the long run, it could help avoid duplication of efforts.

Would each program/region develop their own system for evaluating acceptability?

Response in White Paper: Yes, in terms of program specifications.

Response from QAEM Workshop: A general system of model assessment, which would cross boundaries of programs and regions, is desirable. If this is framed as a process with different emphasis required on different aspects of the process according to program requirements, the process should be acceptable to all parties. The idea is to build consensus of users and developers across programs on appropriate and acceptable model assessment methods.

Should EPA apply a generic set of criteria across the board to all categories of ERMs or should acceptability criteria differ depending on the complexity and use (e.g., screening vs. detailed assessment) of a model?

Response in White Paper: The CREM should be established to provide updated general guidelines on MAC, which should be consistently applied throughout EPA. However, the specific application of the MAC will be adaptable to each situation as determined by the program manager.

Response from QAEM Workshop: This "generic set of criteria" should be process-based, and provide great detail on methods for assessment. Then, the "acceptability criteria" can be applied with varying levels of importance depending on the needs of the program.

5 Research Needs

The papers and discussions at the workshop illustrated the state of the art and ideas under development in research, and illustrated current practice in EPA. Some detailed research requirements regarding VHOMs and peer review are mentioned in the reports of the discussion groups. Two additional themes were mentioned:

1. ***An integrated approach to model assessment***

   There was recognition by a substantial number of research scientists that *validation*, in the sense of tests against data that allow a model to be accepted as true, is inadequate. A current research trend is to develop methods to perform a more comprehensive analysis of sources of uncertainty, for example by extending the traditional application of sensitivity analysis, or examining multiple inputs and outputs. There are also attempts to develop an overall framework within which the multiple sources of uncertainty – and their consequences – can be defined, such as the Bayesian framework. However, these attempts are few, their application is, as yet, to simple models, and there is no consistency between researchers that would enable an agreed method for assessment of environmental models to be defined. Given the importance of models for regulatory and environmental management this topic is a research priority. Some specific research questions are:

   - Can techniques be developed to assess uncertainty in model structure as well as uncertainty in model inputs and outputs?
   - Can software be developed that can automate uncertainty analysis for a defined model?
   - Can we develop an understanding of how multiple causes of uncertainty interact, and produce methods for defining principal sources of uncertainty?

2. ***Advancing current practice***

   A concern of many participants was the lack of an accepted current practice for model assessment. Although there is discussion and debate about how to

develop a comprehensive approach there are, nevertheless, sets of techniques illustrated at the workshop, that could be applied to environmental models to improve our understanding of the uncertainty. One reason why these techniques might not be being used is the lack of rigorously worked examples. Illustration of best practice for different types of environmental models could help to define a current standard.

6 References

Beck, M. B., J. R. Ravetz, L. A. Mulkey and T. O. Barnwell. 1997. On the problem of model validation for predictive exposure assessments. Stochastic Hydrology and Hydraulics, **11:229-254**.

Morgan, M. G. and M. Henrion. 1990. Uncertainty. Cambridge: Cambridge University Press.

Hassanizadeh, S. M. and J. Carrera. 1992. Editorial. Advances in Water Resources, **15:1-3**.

Rubin. 1987.

USEPA. 1999. White Paper on the Nature and Scope of Issues on Adoption of Model Use Acceptability Guidance. Prepared by The Science Policy Council Model Acceptance Criteria and Peer Review Working Group. Draft version May 4, 1999.

USEPA. 1998. Science Policy Council Handbook: Peer Review. EPA 100-B-98-001. Office of Science Policy, Office of Research and Development. December, 1998.

Appendix A : Glossary of terms and acronyms

Ad-hoc propositions –

ATFERM – EPA Agency Task Force on Environmental Regulatory Modeling

BASINS – EPA watershed model system named *Better Assessment Science Integrating Point and Nonpoint Sources*.

CREM – EPA Committee on Regulatory Environmental Modeling

Deterministic model -

EPA – United States Environmental Protection Agency

Epistemology - The division of philosophy that investigates the nature and origin of knowledge, the theory of the nature of knowledge.

ERM – Environmental Regulatory Model

Explanatory capacity –

Explanatory coherence –

Explanatory relevance -

Forcing functions –

GIS – Geographic Information System

MAC – Model Acceptance Criteria

Model calibration –

Model instrumentation -

Model parameterization -

NRCSE – National Research Center for Statistics and the Environment

Part and kind relationships –

POMAC – Pareto optimal model assessment cycle

Proposition - A plan or scheme, a statement containing only logical constants and having a fixed-truth value.

QAEM – Quality Assurance of Environmental Models

Quality Assurance (QA) -

SAB – EPA Science Advisory Board

Sensitivity Analysis - Explores the consequences of uncertainty (if I make a change in x, what happens to y?). Ascertains how a model depends upon the information fed into it.

SPC – EPA Science Policy Council

Stasis - A condition of balance among forces; motionlessness.

Stochastic process -

Transparent Propositions -

Uncertainty Analysis - Describes the parameter uncertainty. Quantifies the uncertainty in what comes out of a model.

Underdetermined Theory - A scientific principle that is generally accepted as fact, but has alternatives that have not completely been ruled out.

Upward Inference - Theories are built on observation, and are revised as observations fail to completely validate the theory.

Valid - Supportable, efficacious, legally sound and effective, containing premises from which the conclusion may logically be derived, correctly inferred from a premise.

White paper - An essay detailing currently recognized problems.

Appendix B : Abstracts

Model Assessment: Where Do We Go From Here?

By Naomi Oreskes
Department of History and Program in Science Studies
University of California, San Diego

In recent years, there has been growing recognition that complex models of natural systems cannot be validated, and that the term “validation” is misleading from both scientific and regulatory standpoints. But if we cannot validate models, what do we do when the regulatory context requires some form of model evaluation? This papers presents three areas for consideration in model assessment: terminology, prediction vs. explanation, and representation vs. refutability.

- Terminology remains a problem in model assessment. The term validation is misleading because it implies an affirmative result, and from this perspective ‘model assurance’ is little better (what are we claiming to assure)? Alternative terms are available: assessment, evaluation, diagnosis.

- Prediction or explanation? Many factors encourage modelers to focus on improving predictive capability. While there may be good reasons to run a model in a predictive mode, predictive capacity is not an adequate basis for model assessment. An alternative is to focus on explanatory capacity, which relies on the strength of the model’s theoretical underpinnings, the quality and quantity of the empirical data base, and the independence of the evidence supporting the model conceptualization. Model evaluation should focus primarily on these model inputs rather than on predictive output.

- Representation vs. refutability. The desire for accurate representation of natural systems typically leads to increasing complexity in models. Social and political pressure for integrative models adds further incentive for complexification. But the more complex a model, the more difficult it is to evaluate it. Highly complex models can become effectively immune from refutation. From a scientific standpoint, bigger is not necessarily better. More discussion of the trade-off between representation and refutability is warranted.

Defining Similarities and Differences in Quality Assurance Requirements for Classes of Environmental Models

By E. David Ford
College of Forest Resources, University of Washington and NRCSE

The term environmental model can refer to a wide variety of model types that may be constructed for different purposes. The quality assurance that is required can depend on the purpose for the model. How that assurance is obtained, and how the degree of assurance is defined, depends upon model type. Similarities and differences between environmental models constructed for different purposes will be reviewed with regard to both their types and degrees of quality assurance.

It is now generally agreed that validation, when model outputs fit data not used in model construction, provide insufficient criteria for quality assurance of many models, and can not be achieved, technically, for others. A wider approach is required that examines the scientific inference that can be made about models relative to their intended purpose. However, an underlying difficulty arises. Environmental scientists, whether working as field or laboratory investigators, statisticians, or process modelers, each have pragmatic standards but no consistent definition of what constitutes scientific inference.

The components of making scientific inference about environmental models will be defined and the potential for developing a unified approach to quality assurance for environmental models will be discussed. This will focus on:

- The components of what constitute an explanation for a scientific question.
- How we make scientific inference according to the status of an explanation.
- How models of different type complicate the process of making a scientific inference due to their characteristics of construction.

EPA's Basins Model - Is It Good Science Or Serendipitous Modeling?

Ray C. Whittemore, Ph.D.
National Council of the Paper Industry for Air and Stream Improvement, Inc.
Tufts University

An EPA geographic-based watershed tool has been developed to better integrate point and nonpoint source water quality assessments for the Nation's 2100+ watersheds. The tool builds upon federal databases of water quality conditions and point source loadings for numerous parameters to allow comprehensive assessments and modeling in typical total maximum daily load (TMDL) computations. While the TMDL utility is the primary reason BASINS was developed, other longer-range water quality assessments will become possible as the Agency expands the suite of assessment models and databases in future releases. The simplistic approach to modeling and user-friendly tools gives rise, however, to technical concerns that are elaborated in this paper. Similar quality assurance issues will also be elucidated.

Facing Prediction and Multimedia Modeling, Model Evaluation is a Science and Knowledge Task: Recommendations from Air Quality Modeling

By Robin L. Dennis
Atmospheric Modeling Division
U.S. EPA/NOAA, Research Triangle Park, NC

The task of modeling in a regulatory setting is forecasting future conditions. The leading edge of environmental modeling has been pushed forward towards more complex process-based multimedia science models as our understanding of the network of interrelations in the physical environment grows and computational barriers fall. However, "validation" of models such as these is strictly impossible for reasons that Oreskes et al. have shown. Even so, the models must still be evaluated so that we can be assured of their utility for their given tasks, and so that we can improve our understanding. Opening up complex, black-box-like models to an evaluation that gets at their component process interactions requires such an enormous effort and specialized measurements, however, that substantive progress has been very slow to come. But the knowledge we gain in making such an investment will advance us considerably toward our goal of understanding the realism of these models relative to their intended applications.

Standard comparisons between measured and predicted variables of concern, while important to clients, have proven to be wholly inadequate. We must use all of the approaches at our disposal and move well beyond regression-model style comparisons. These include diagnostic evaluation, sensitivity analysis, instrumentation of the models, system comparisons, model structure analysis, laboratory studies, and specialized data sets. Experience with advanced air quality modeling indicates the first three are critical. This talk will concentrate on these three, with illustrations from our investigations of nonlinear photochemical models.

Diagnostic evaluation is aimed at testing the process descriptions in models and at helping us to explore for ignorance. Our approach to diagnostic evaluation involves an tripartite framework: a high-level conceptual model, process instrumentation of the model, and process oriented test measures. The goal is to develop tests of the model's realism relative to its intended application. Because evaluation tests will not be definitive, the approach also calls for a procedure to develop a weight of evidence judgment of the model's ability to address the required predictive tasks. The tripartite framework is illustrated with an example showing how in situ diagnostic measures related to an ozone response surface were developed from the tenets of the conceptual model and realized in process terms of the instrumented model. The diagnostic measures probe the model's differential sensitivity to emissions changes on either of ozone's two main chemical precursors, volatile organic compounds and oxides of nitrogen. A second illustration shows how an important area of ignorance in the mechanism of photochemical processing can be probed through diagnostic measures. Instrumenting the model to reveal the behavior of its component processes for direct evaluation is shown to be the crucial element of the diagnostic model evaluation. Difficulty in making the necessary ambient observations to populate the measurement

suite needed for a process-oriented diagnostic evaluation is shown to be a significant limitation to evaluating complex process models.

Sensitivity analysis, whereby the behavior of the full model as a system is assessed, is a significant complement to diagnostic evaluation, but must be carried-out with care. is shown with an example using ozone. The ozone response of the model to input perturbations is not the same as the response of system to the same perturbations when examined with respect to the forecast for emissions reductions. The relevant sensitivity is the latter one since the policy role for the model is forecasting the effects of such emissions changes. It can be seen that sensitivity analysis with these complex process models is not trivial crank-turning, and that the models' forecasted response must be carefully unpacked to provide understanding.

Also, insights gained regarding dynamic cycling will briefly be noted, and the possibility of some cross-over to larger ecosystem-scale models will be discussed. Finally, insights will be presented regarding systems comparisons, where the model's virtual reality is compared against the physical reality of the environment, and the need for additional technique development noted.

Ecological modeling to assess the effect of land cover on water resources: A summary of approaches and modeling issues

By Iris Goodman
US EPA Landscape Ecology Branch
National Exposure Research Laboratory
Las Vegas, NV

Ecological models are increasingly being used by the US Environmental Protection Agency to assess the effect of land cover on the condition of water resources within the United States. Regional and sub-regional scale studies are underway investigating land cover effects on water quality, the timing and magnitude of streamflows, and the biological integrity of freshwater and coastal aquatic systems. A variety of modeling methods are being used, including statistical models, lumped parameter or fully distributed physical models, and knowledge-based systems. Remotely sensed land cover data, or derivative landscape indicators, are used in all the studies. This paper summarizes the variety of modeling methods being used and contrasts the objectives of the studies with respect to the rigor and accuracy required of the models for their intended application. This suite of studies also highlights a number of modeling issues requiring further exploration by the scientific community, including appropriate matching of modeling approach to objective; model parameterization using remotely sensed data; "validation" of models for applications in which supporting empirical data is not available; and determining the importance (v. statistical significance) of model results, relative to uncertainty in model simulations.

Exposure-Response Modeling of Ozone-Induced FEV_1 Changes in Humans: Effects of Concentration, Duration, Minute Ventilation, and Age

By William F. McDonnell[1], Paul W. Stewart[2], and Marjo V. Smith[3].

[1]Human Studies Division, U.S. EPA, RTP, NC.
[2]Dept. Of Biostatistics, University of North Carolina, Chapel Hill, NC.
[3]MVS Biomathematics, Raleigh, NC.

Short-term exposure of humans to ozone results in acute decrements in lung function manifested by changes in FEV_1. It is known that individuals vary considerably, but reproducibly, in magnitude of response to ozone and that response is a function of ozone concentration (C), duration of exposure (T), minute ventilation during exposure (V_E), and age, although a model which accurately and simultaneously describes the relationship between FEV_1 change and these independent variables has not been identified. The purposes of this analysis were 1) to identify a model consistent with known ozone E-R characteristics; 2) to estimate model parameters using a large existing data set; and 3) to test the predictive abilities of the model using cross-validation techniques. 485 healthy, young adult male volunteers were exposed once for two hours in an environmental chamber to one of six ozone concentrations while exercising at one of three nominal exercise levels. FEV_1 response was measured after one and two hours of exposure. Two models with different variance structures were fit to portions of the data, and observed and predicted values from the other portions of the data were compared. Both models were found to predict mean response equally well, with a substantial amount of individual variability in response not accounted for by the model. In summary, we have identified a biologically plausible, predictive model that accurately quantifies the relationship between the ozone-induced mean FEV_1 change and C, V_E, T, and age.

Sensitivity analysis and the quality assessment of environmental models

By Andrea Saltelli

Institute for Systems, Informatics and Safety,
The European Commission, Joint Research Centre, Ispra (I)

Sensitivity analysis has a role to play in model building and model assessment. SA provides a tool to monitor the transparency, the relevance, the robustness and the parsimony of a given model. The model X-raying offered by SA is a necessary (though by no mean sufficient) prerequisite for the use of a given model in a policy context. Models are flawed by the unavoidable presence of uncertainties, which arise at different stages, both in the construction / corroboration of the model itself, in its quality assurance, and especially in its use. An understanding of the influence of the above on the course of action suggested by the model is crucial as:

- Different level of acceptance (by the general public, by the decision-makers, by stakeholders) may be attached to different types of uncertainty.
- Different uncertainties impact differently on the reliability, the robustness and the efficiency of the construct.
- The relevance of the construct (its appropriateness to the task) strongly depends on the impact of (which) uncertainty on (what) outcome of the analysis. Examples of settings where the above applies are infinite, and include:

  ⇒ Construction of environmental indicators

  ⇒ Analysis and forecast of risk

  ⇒ Model optimisation and calibration

Sensitivity and uncertainty analysis can be helpful to tackle these issues. Sensitivity Analysis (SA) aims to ascertain how a given model (numerical or otherwise) depends upon the information fed into it. Uncertainty Analysis (UA) aims to quantify the uncertainty in what comes out of the model. SA "evaluates" models (with respect to quality) in several respects:

- By allowing unimportant factors to be fixed or eliminated, thus testing the relevance of models (a model is of poor relevance for a given task if it contains many input factors whose value does not drive variation in the output being sought for the task)
- By falsifying or corroborating a model, or by allowing the selection among different models (which model has the preferred sensitivity pattern? Does the given model serve the purpose of -say - an envisaged calibration given the uncertainties...)
- By allowing a test of the sensitivity pattern against the analyst's understanding of the problem (why is the given factor so important?)

This is illustrated with three worked examples:

- A chemical kinetics model used in atmospheric chemistry studies of the sulphur cycle;
- A problem of trend extraction from time series
- A case of use of environmental indices in a policy context.

Statistical Inference for Deterministic Simulation Models: The Bayesian Melding Approach

By Adrian E. Raftery
Department of Statistics, University of Washington, and NRCSE

Deterministic simulation models are used in many areas, including the making of environmental and other policy decisions, atmospheric science, engineering, pharmaceutical research, and demography. They tend to be complex, and to require the specification of many inputs. This is often done in an ad-hoc manner, and little attention has been given to taking proper account of uncertainty and evidence about the inputs and outputs to the model. Statisticians have only started to be involved in the analysis of such models, although their skills have the potential to contribute a great deal.

I got involved in this problem through my work for the International Whaling Commission on determining if bowhead whales could safely be subjected to aboriginal subsistence hunting by the Inuit people of Alaska, and on setting the quota. This has traditionally done using deterministic population dynamics models. Our first effort to take proper account of the uncertainties involved was the Bayesian synthesis method of Raftery, Givens and Zeh (1995, JASA). However, this suffers from the Borel paradox, according to which the results may not be invariant to reparameterizations of the model. I will describe the Bayesian melding method, which overcomes this difficulty by bringing together ideas from modeling, measure theory and the pooling of expert opinions. An application to environmental risk assessment will be briefly described.

This is joint work with David Poole and Samantha Bates.

Bayesian Calibration and Model Correction

By Tony O'Hagan

University of Sheffield, UK

There are many kinds of uncertainty in the use of mechanistic models. I will offer one way of categorising uncertainty sources, and discuss the nature of each.

Users of models are familiar with the idea that they typically need calibrating in order to provide predictions in a given application. This entails trying to identify values of some of the unknown parameters of the model that provide the best fit to observational data on that application. Calibration addresses and tries to reduce one of the sources of uncertainty, which I call parameter uncertainty. The observational data allow us to reduce our uncertainty about the "true" values of the unknown parameters. An important point, and one that is too often neglected in practice, is that calibration does not eliminate parameter uncertainty. After calibration it is still important to quantify remaining uncertainty about the unknown parameters, and to assess the effect of that uncertainty on predictions.

We also know that all models are wrong, and that even with the most well-chosen values for its parameters a model will never predict perfectly. Model inadequacy is another important source of uncertainty, and is notoriously difficult to quantify. A way to assess, and even to reduce, model inadequacy is offered by the calibration process. The residuals remaining between the observations and the best fitting calibration allow us to assess the extent of model inadequacy in the particular application. Furthermore, suitable smoothing of these residuals has the potential to reduce model inadequacy - a process I call model correction.

I will present a Bayesian framework which addresses these points, with the following key benefits.

- All the sources of uncertainty are acknowledged and quantified.
- The calibration process is tackled as a statistical estimation problem in which the errors in the estimates can be quantified.
- Model correction is applied.
- The approach is integrated with uncertainty and sensitivity analysis.
- The methodology will be illustrated with some examples.

Open Questions in Applying the Pareto Optimal Model Assessment Cycle

By Joel H. Reynolds, Ph.D.
Gene Conservation Laboratory
Alaska Dept. of Fish and Game, Commercial Fisheries Division
Affiliate Asst. Professor, Department of Statistics, University of Washington
Member NRCSE

The Pareto Optimal Model Assessment Cycle (POMAC) is a multiple criteria model assessment methodology for exploring uncertainty in the relationships between process conception, model structure, and assessment data. Model performance is optimized to satisfy, simultaneously, each component of a vector of assessment criteria (model outputs), rather than the usual procedure of optimizing performance with respect to a single criterion. Pareto Optimality is used to define the vector optimization. The Pareto Optimal Frontier reveals which combinations of assessment criteria the process model can satisfy simultaneously. This is more stringent and informative than traditional model assessment procedures as it uses multiple criteria without weighting and aggregating them. The inability to satisfy all criteria simultaneously highlights deficiencies in the selection of component process hypotheses underlying the model, the mathematical representations of these hypotheses in the model structure, and/or the selection and formulation of the assessment criteria.

POMAC has been shown to improve a researcher's ability to detect model deficiencies and locate their sources (Reynolds and Ford, 1999). Software has been created to solve the multiple criteria optimization problem underlying the generation of a model's Pareto Optimal Frontier for a given set of assessment criteria. However, using POMAC and the software requires a number of technical decisions in two main areas: (1) selection of assessment criteria and their error measures, and (2) specification of key characteristics in the evolutionary computation optimization algorithm. This presentation will start with a brief overview of POMAC and then focus on defining and discussing these open technical questions. Possible solutions and, where available, results from empirical investigations will be shared. The issues will be illustrated by the assessments of process models of blood cell production in cats, three dimensional crown development and canopy competition in a dense forest stand, and annual plant competition.

Space-time model evaluation

By Wendy Meiring, Ph.D.

Comparison of the space-time structure of environmental model results with that of observations is an important component of model evaluation. Challenges may include accounting for differences in the spatial and temporal scales of the observations compared with those of the environmental model results. In addition, the environmental processes may be non-stationary in space and time due to topographical influences, pollutant emissions, chemical reactions, and transport processes. Model results may be available only for short time periods due to the cost of running the models. Observations also may be sparse in space and time. I discuss space-time model evaluation techniques and challenging directions for further development, primarily in air quality studies.

"Best Available Information" to support European policy making (what is good enough and sufficient, and how do we get there)

By David Stanners
Programme Manager for Integrated Assessment and Reporting European Environment Agency, Copenhagen

The European Environment Agency published in June 1999 its latest report on the state and prospects of the EU environment up to the end of the first decade of the next century[1]. Despite the successes with environmental policies over the past 20 years, the prospects are not good when you take into account the expected socio-economic developments on an agreed baseline of EU-wide developments projected onto environmental impacts. The issue at stake is that without integrating environmental (or more widely sustainability) considerations into sectoral policies, environmental policies will be overcome by the sheer scale of increasing economic activities. Absolute de-linking between socio-economic development and environmental impacts are required to reverse this trend, something which is beginning to be seen in some parts of a few selected areas (eg, energy) and for a few pollutants (eg, SO_2), but not yet at all in a consistent fashion across the board.

The modelling exercise which has yielded these results is the state-of-the-art in Europe linking modellers and policy makers to the European Environment Information Observation Network of existing capacities of information gathering across the EU Member States and beyond to the EU Accession countries of Central and Eastern Europe. Linking these data in such a broad assessment is an important part of the quality assurance, but a thorough uncertainty and sensitivity analysis has not yet been possible. As far as the results go, and the political process is ready, it could be said that the results are good enough. But the scope for improving and refining the results to indicate the scope for change and opportunities for alternative courses of action still remain huge and largely untapped.

For our current needs the quality may be sufficient; to go further will require more effort. In the end the quality has to fit the purpose. Defining the use to which results will be put is an essential prerequisite of any quality assurance exercise when working with complex problems and such highly disharmonised systems of data and information as found across Europe. Working with existing data and information and drawing conclusions now are therefore necessary and constant challenges of information providers to policy makers. A very practical approach is therefore needed to remain relevant and useful to decision makers and to ensure that the "Best Available Information" is put to use and not ignored.

[1] Environment in the European Union at the Turn of the Century. European Environment Agency, Copenhagen. Office for Official Publications of the European Communities, Luxembourg, 1999, pp 446.

Modeling Quality Assurance Plan for the Lake Michigan Mass Balance Project

By William L. Richardson, Douglas D. Endicott and Kenneth R. Rygwelski
USEPA, ORD, NHEERL, MED-Duluth, Community-Based Science Support Staff
Large Lakes Research Station

With the ever increasing complexity and costs of ecosystem protection and remediation, the USEPA is placing more emphasis on ensuring the quality and credibility of scientific tools, such a models, that are used to help guide decision-makers who are faced with difficult management choices in these areas. The Agency has issued several documents covering broad requirements of the development and use of mathematical models and these are used in the formulation of the Modeling Quality Assurance Plan (MQAP) for the Lake Michigan Mass Balance Project (LMMBP). This is a stand-alone document, separate from, but related to the LMMBP Quality Assurance Project Plan (QAPP). Because guidance for modeling QAPs is new and somewhat limited, the LMMBP MQAP could be viewed as a prototype particularly for projects that involve holistic, multi-media modeling approaches for large systems like the Great Lakes.

The LMMBP modeling design includes a number of computational components: hydrodynamics, sediment transport, eutrophication, chemical transport and fate, and food web bioaccumulation. In addition, the MQAP includes the quality assurance (QA) process for the development of atmospheric models used to describe the emission of the current-use herbicide, atrazine, from the agricultural lands in the U.S. and its transport and deposition to the lake. It also includes the quality assurance process for the estimation of tributary and atmospheric depositional loads for atrazine, as well as persistent, bioaccumulative chemicals including polychlorinated biphenyls (PCBs), trans-nonachlor (TNC), and mercury.

An extensive sampling program was designed based primarily on modeling requirements. These requirements included information for mass inputs from tributaries and the atmosphere, boundary and initial concentrations in air, water, sediment, and biota, and concentrations of state variables in space and time for comparison to model computations. In the final analysis the model will be judged by how well it simulates real-world conditions. The sampling program was executed in 1994 and 1995 and data are currently becoming.

This estimation and evaluation of uncertainty in model predictions is of great importance for decision-making. For the LMMBP, uncertainty in the predicted relationship between controllable mass loadings, and contaminant exposure and fish tissue concentrations over time, is of highest concern. Estimating these uncertainties involves quantifying the uncertainty due to parameter identification within each modeling component, as well as the propagation of uncertainty from one component to another. The procedures to carry out this process have been developed in previous studies but never applied at such a large scale or complexity. Results of uncertainty analysis must also be properly communicated to the clients of the LMMBP, as well as interest groups and the public.

We will explain our approach to modeling quality assurance, present the methods by which models will be compared to data and show how uncertainty of model predictions will be determined.

Model Use Acceptability Guidance: Part 1)
Model Validation for Predictive Exposure Assessments: A Draft Protocol

By Tom Barnwell, Bruce Beck and Lee Mulkey

The purpose of the protocol for model validation is to provide a consistent basis on which to evaluate 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 a process for arriving at such a judgment. There are three aspects to forming a judgment 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 judgment on the validity of the model should 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 entirely new 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 judge the validity of the model. Evidence on which to base this judgment would tend to be primarily that of an expert opinion, and therefore largely of a qualitative nature.

While the depth of the inquiry and length of the process in coming to a judgment 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 model developer. In the proposed protocol it has been assumed that a second, independent entity 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.

Model Use Acceptability Guidance: Part 1)
Updating the Protocol for General Agency Use: Stakeholder Input

By Linda Kirkland
Environmental Protection Agency, Washington D.C.

An Environmental Protection Agency-wide work group convened by the Agency's Science Policy Council recommended updating the 1994 draft Model Validation Protocol for Predictive Exposure Models as an approach to providing general guidance on model acceptance criteria and peer review. An overview of current practices suggested expanding the coverage to all models used by the Agency and outlining a broader model evaluation process addressing five areas of model uncertainty:

- theory upon which the model is based;
- translation of the theory into mathematical representation;
- transcription into computer code (software quality assurance);
- assignment of parameter values and calibration (data quality assurance); and
- model test choices and implementation.

Work group discussions of the qualitative and quantitative aspects of the uncertainty areas identified a need to clarify the roles of the Agency's programs for peer review and quality assurance, respectively. For example, what record from quantitative quality assurance evaluations are needed for answering peer review charter questions?

An third phase, an overall assessment, was added to the protocol using the first phase, task definition as a basis for programmatic evaluations of regulatory applications subsequent to selection of the model for Agency use. Workshop participants are encouraged to comment on the completeness of the proposed regulatory environmental model evaluation process and roles for quality assurance and peer review.

Evaluating Performance and Reliability of Intermedia Transfer Models Used in Probabilistic Human Health Risk Assessment

By Helen E. Dawson, Ph.D
Hydrogeologist, Superfund Program Support
U.S. Environmental Protection Agency, Region 8

A methodology for estimating the performance and reliability of intermedia transfer models used in probabilistic human health risk assessment is presented in this paper. Results of intermediate-scale VOC leaching and volatilization experiments, computer modeling, deterministic and probabilistic human health risk calculations, and statistical data analysis are used to illustrate application of the methodology. The aim of the proposed methodology is to provide needed information to risk assessors and decision-makers who must establish "acceptable" modeling error and uncertainty tolerances for risk assessments in accordance with the U.S. EPA data quality objectives process.

A novel multimedia lysimeter apparatus was developed for experimentally measuring soil VOC fluxes during simultaneous leaching and volatilization under controlled but representative laboratory conditions. The multimedia lysimeter included a soil block underlain with a porous plate and hanging water column, a dynamic volatilization flux chamber that overlies the soil, an artificial rainmaker, and an array of sensors that monitor soil and atmosphere environment variables. A model validation exercise was performed using 24 leaching and volatilization data sets obtained from the lysimeter apparatus for experiments with four VOCs (1,1,1-trichloroethane, tetrachloroethene, trans-1,2-dichloroethene, and toluene) and three soil matrixes (a medium sand, a silty sand, and a heterogeneous soil matrix). The experiment results were compared with time-average predictions made independently by three chemical fate and transport models (EMSOFT, VLEACH, and SOILMOD).

Using a new adaptation of the data quality objectives process, the three models' reliability was evaluated by applying risk-based decision performance goal diagrams. For four typical human exposure scenarios that were evaluated based on model-predicted vapor and leachate (uncertain) concentrations, upperbound probabilistic cancer risk estimates were within a factor of three to seven times the upperbound risk estimates calculated based on hypothetical perfect (exact and correct) concentration information. For commonly observed risk management conventions, the performance of all three evaluated models was well within the risk uncertainty tolerance limits that defined acceptable model reliability. Application of the methodology reported in this paper to reach conclusions for a particular site mitigation scenario must consider the representativeness of the existing model validation database and the applicability of the assumptions used in the decision analysis.

Appendix C : Summary of Discussion Group Sessions

Discussion Group 1: QA of Models/Life Cycle of Models

(Wednesday, September 8)

This discussion was based on the broad topic of the status of quality assurance (QA) for models. Several key points emerging from the free-ranging discussion are detailed below:

1. The group agreed that the purpose of model QA is to facilitate better environmental decision-making and regulations. Therefore, all aspects of model conceptualization, development, and communication must be included in the discussion of model quality.
2. In the future, this modeling process will be under increased public scrutiny. As the public exercises their "right to know", scientists will have to defend all aspects of model development. Although it is unreasonable to expect that "the public" can understand all technical and scientific aspects of complex models, there must be some aspect of transparency.
3. QA for the technical development of the model is embedded in QA of the modeling process. These technical aspects are not necessarily simple to check and approve. All models are based on theories and assumptions which may be open to dispute. Different elements of the process may be relevant to different contexts, so QA may be context-dependent. Issues which need to be addressed and the level of QA required depend upon the degree of uncertainty in the science and the stakes in the decision. QA in a context with few empirical observations, speculative theories, high uncertainty, and large consequences of wrong decisions (i.e. "post-normal" science) differs from QA in a context with plentiful data, parallel problem situations, and small consequences from wrong decisions (i.e. "true science"). Most problems are likely to fall somewhere in the middle, in situations requiring consultation between differing models and theories.
4. Decision analysis is an existing conceptual framework for decision-making which incorporates different types of uncertainties. More consistent application of clearly understandable decision processes is likely to aid in the public discourse.
5. We already have several different checklists on the elements of the modeling process, and they are self-similar. Within this list, and others like it, some items have QA procedures that are good but are not consistently followed, while others have no procedures in place. An example list of elements:
    - 5.1. Model development (including conceptualization, specification of equations and system requirements, designing and writing computer codes, and implementation).
    - 5.2. Collection and use of data in support of model development,
    - 5.3. Model acceptance, and

5.4. Model application.

Additional comments from final discussion session September 10:

1. Project plans should include an integrated peer review plan.
2. It is necessary that the component functions of the model (including evaluation and use) are performed by those in the correct profession. Interdisciplinary teams are often needed, but are difficult to manage and maintain.
3. The QA process must include a feedback loop, so that quality is continually evaluated as a model is updated, corrected, and adapted to different uses.
4. It may be necessary to prioritize certain aspects of the modeling process which are in critical need of quality inspection.

Discussion Group 2: Peer Review

(Wednesday, September 8)

Five topics which need clarity for EPA's peer review process in modeling projects were discussed in this session.

1. **Definitions**

The scientists whose project is to be reviewed are the *modelers*. The scientists to review the project are the *reviewers*. A reviewer should be a person independent of the modeling project. They should have expert knowledge on the subject of the project. Interdisciplinary reviewers may be needed to cover the full range of activities/issues being reviewed. The number of reviewers must be appropriate to the complexity of the model. Public comment is separate from peer review. The purpose of the peer review process is to:

1.1. Provide scrutiny of a modeling project in the interest of challenging the modelers in a constructive sense. The reviewers should give recommendations for model improvement.

1.2. Provide the scrutiny to assess:

1.2.1. The adequacy of the modeling system,

1.2.2. The qualifications of the model application users,

1.2.3. The adequacy of the model for its intended purpose, and

1.2.4. The qualifications and records of the modelers.

2. **The Ideal Role of Peer Review in Quality Assurance**

Some modelers see peer review as a hostile process, especially when it is done by outside reviewers. EPA must open up the modeling project process to peer review. Modelers should submit their modeling project to a peer review for the following reasons:

2.1. To improve the science associated with the modeling project.

2.2. To improve the model project's application and effectiveness.

2.3. To gain credibility of a model's project, which may lead to more credibility for EPA as a whole, more project funding, etc.

2.4. To improve communication among disciplines involved in the model project.

2.5. To avoid disasters and costly mistakes.

2.6. To promote openness and honesty in the whole model project and peer review process.

3. **Difficulties, Limitations, and Timing of Peer Review**

There are difficulties to conducting peer reviews at the institutional, managerial, and project levels for financial, administrative, temporal, and social reasons:

3.1. Peer review requires the time of the people involved (modelers and reviewers) and money. These needs should be considered at the beginning, but are often not estimated and budgeted into the model project. Financial needs for the review process are: travel costs, review dialogue time, and writing time.

3.2. The reward structure for participation in the review process does not include peer review for:

3.3. Those who recommend that reviews should occur,

3.4. The modelers who undertake to have their project reviewed, and

3.5. The reviewers.

3.6. There can be sociological problems, such as reviewers from the same or very different groups as the modelers in terms of discipline, institution or geographic location. Other difficulties such as time and money aggravate these problems.

3.7. Timing of the peer review process is very important. It should begin with work plan of model project and continue or occur at logical steps through the model project process.

4. **Process, Outcome, and Timing**

The peer review process may be summarized by these steps:

4.1. Allocate and budget resources for a peer review. The peer review process should be integral to model project work plan.

4.2. Select the reviewers. Decide whether the collection of reviewers, known as the peer review panel, will be on-going or one time. This choice will affect step 1.

4.3. Lay out modelers' expectations of review panel.

4.4. Schedule the receipt of materials for review and ensure they are received in a timely fashion.

4.5. Provide institutional support to reviewers, e.g., photocopying, phone calls, faxes.

4.6. Review panel needs to give critique of model project and recommendations.

4.7. Modelers may give a project response to the reviewers' critique.

4.8. The review and the modelers' response (steps f-g) should be published on the internet or in a journal, not just in an agency publication (which is often hard to access). This step would allow the review process to be more open, creative, integral to the scientific process, and accessible to the scientific community.

5. Recommendations and Conclusions

The peer review process has its difficulties, but it is the best tool available for model project evaluation. It is in the USEPA's best interest to make the peer review process an integral part of the model project process for reasons of time, money, integrity, and the agency's reputation. The discussion group wanted to stress the importance of the peer review process for several points in particular:

5.1. It is expensive but extremely valuable when done effectively.

5.2. It requires institutional support at all levels.

5.3. It needs guidance and central institutional oversight, e.g., Science Advisory Board, to be effective.

5.4. The level of effort needs to be prioritized according to levels of importance and complexity.

Additional comments from final discussion session September 10:

1. The peer review process is not anonymous, not detached. There must be interaction and feedback between reviewer and modeler.
2. This process could include participatory public review to include different cultural perspectives in the assessment of quality.
3. There is a problem of terminology if we shift from strict technical validation of a model to review of the modeling process. Both the model and its application must be reviewed, so a greater variety of review mechanisms are required.

Discussion Group 3: Very High-Order Models

(Thursday, September 9)

The group discussed several issues about very high-order models (VHOMs), focusing on VHOMs in relation to quality assurance (QA) and use.

1. **Rational Procedure for Model Assembly -** While there seems to be no single protocol for model assembly, the group identified these important steps:
    1.1. Explicitly define the problem, the client or clients, funding sources, and the disciplines related to the problem.
    1.2. Gather an interdisciplinary team for the model's construction.
    1.3. Create a conceptual model using a modular construction so that model may be examined as pieces, e.g., cartoons, diagram, flow charts, and nuggets. Before one begins building the model, the conceptual model illustrates all the pieces one needs in order to develop an answer to the problem. A *nugget* is defined as a multimedia, but singular problem that will include many disciplines for its construction. Features of the conceptual model such as assumptions, scale (global, national, regional), and media (soil, water, air) must be specified explicitly.
    1.4. Gather another interdisciplinary team for the peer review. VHOMs require peer review panels to be involved from the beginning for at least the following reasons:
        1.4.1. The complexity of VHOMs requires that the panel have much expertise about the model and the science involved in order to understand and critique the model. There may be problems finding competent reviewers who are not competitors or do not have other conflicts of interest.
        1.4.2. VHOMs should be reviewed at many different stages during their development. It would be time and resource consuming to educate a new panel at each review stage.
    1.5. Resolve difficulties over integration of VHOM modular components in terms of data, theory, scale, and code.
2. **Rational Procedure for Disassembly, Scrutiny, and Interrogation -** As there are problems with constructing VHOMs, so are there problems with breaking them down for review. The group identified several in particular:
    2.1. Many VHOMs may not be reduced to individual components to permit scrutiny by individual disciplinary experts. Instead, interdisciplinary teams of reviewers within the review panel must do the review.
    2.2. VHOMs require evaluation externally and internally:

2.2.1. Externally by such techniques as data comparison. That is, history matching between observed data and model output.

2.2.2. Internally by diagnostic tests. That is, measures on processes within the VHOM. Some processes will be theoretical, some structural (mathematical equation parameters or computational code). Observed data are also necessary for diagnostic tests.

2.3. Sensitivity analysis (SA) and uncertainty analysis (UA) should be performed on the model as part of the model's evaluation.

2.3.1. These analyses may be complicated by connections between modular components and due to scale differences that result from disconnects in data, knowledge, and theory.

2.3.2. For methods and discussion on SA, see meta-modeling as found in operations research literature, and variance-based SA as discussed by A. Saltelli.

2.3.3. For methods and discussion on UA, see abstract by A. O'Hagan.

2.3.4. SA and UA must be made routine steps in VHOM development. The execution of these steps is constrained by the availability of resources.

2.3.5. Features in the VHOMs should facilitate the diagnostic tests, e.g., reporting of internal process rates, meaning for error messages during runtime.

3. **Data -** One of the critical issues associated with VHOMs relates to data. Part of the value of VHOMs is that they allow us to envision a greater understanding of complex ideas than we could possibly achieve through empirical reason alone. Unfortunately, theories must be grounded by some tangible proof, proof that we often reliably achieve from data. The group recognized these difficulties relating to data for VHOMs:

3.1. Measurable variables for building and evaluating VHOMs may not exist, e.g., data for evaluating process rates as are necessary in diagnostic tests.

3.2. There are problems with data collection across multiple media or large geographical scales, especially when data collection must occur simultaneously.

3.3. Modelers should be more proactive about getting funding and resources for data collection. Hindrances included marketing, politics, and industry.

3.4. Communication between modelers and the monitoring community must be improved:

3.4.1. Often monitoring data are not of use in VHOMs.

3.4.2. Modelers must direct data collection through better communication with the data collectors if they want to get data that are applicable to VHOMs.

3.4.3. The modelers should be more specific about their data needs in terms of quality, frequency, intensity, etc.

3.4.4. More research is needed to allow for data collection of process rates instead of just endpoint variables.

4. **How do we refute VHOMs?** The particular nature of VHOMs complicates the ability to refute them. They should be evaluated at multiple stages. Often the modelers have to do the technical evaluation themselves, e.g., sensitivity analysis and uncertainty analysis. They present the results to the peer review panel. The reviewers may have the opportunity to run evaluations on the models directly through beta test sites on the web. The interaction and communication between the modelers and reviewers must be open to allow for a competent evaluation.

Additional comments from final discussion session September 10:

1. Two characteristics of VHOMs make them difficult to evaluate; they take a long time to run and they have lots of inputs. Methods of sensitivity and/or uncertainty analysis that employ monte carlo methods (e.g., A.Saltelli's SA, K. Beven's GLUE, or A. Raftery's Bayesian melding) are impractical when individual run times are long. Model-based methods (e.g., A. O'Hagan's Bayesian method) have not yet been developed for models with many inputs. (However, see recent work by Michael Goldstein with Bayes linear approach to problems with many inputs.) New assessment tools to resolve the problems simultaneously must be developed. This issue is an area that needs research.
2. The need for data is crucial, and must be more vigilantly pursued by modelers and their managers.

Discussion Group 4: Tool Chest for Model Assessment

(Thursday, September 9)

This discussion was segregated into two large topic categories: existing tools and their usefulness, and tools which are needed.

Existing Tools and Their Properties (e.g., parsimony, efficiency, limitations, and appropriate use)

1. Sensitivity Analysis and Uncertainty Analysis:
    - 1.1. The potential uses are many, throughout modeling, model assessment, and policy analysis. However, it is not often applied or properly applied.
    - 1.2. Examples are required for its use in model validation.
    - 1.3. There is a "Handbook for Sensitivity Analysis" (A. Saltelli) to be published next year which will address some of these issues.
2. Elicitation of Expert Knowledge:
    - 2.1. There is a prevailing existence of uncertainty in environmental decision making, and there are many sources for this uncertainty. Some sources, such as those underlying the physical processes of the model are difficult to quantify and may require expert knowledge.
    - 2.2. Problems:
        - 2.2.1. Difficulties in eliciting information about extremes
        - 2.2.2. Difficulties in eliciting information and beliefs about even low order associations.
        - 2.2.3. Methodology for the selection of experts
        - 2.2.4. Reproducibility
    - 2.3. Based on the choice of experts, the choice of relevant data/models may vary.
3. Multi-level Codes
    - 3.1. Making use of reduced (simplified, nested) codes as an aid to efficiency without losing information in the more complex model.
    - 3.2. This may aid in the perception of the adequacy of simpler models to address complex problems.
4. Software Independent Verification and Validation (IVV)
    - 4.1. There are books on the subject (including documentation on methods and user's guides)

4.2. Reproducibility of code on different platforms is a problem.

4.3. Very detailed analysis, with some automated tools.

4.4. Limitations: Expensive

4.5. Advantages: Eliminate errors

4.6. Should be applied according to the importance of the problem.

Requests for New Tools

1. Uncertainty Analysis for Functions:
   1.1. Statistical science does have many tools, but which ones to choose and how to choose them is not always obvious.
2. Best practice for Analysis of Conceptualization
   2.1. Peer review
   2.2. David Ford's list of questions (see White Paper on the Nature and Scope of Issuses on Adoption of Model Use Acceptability Guidance, May 1999)
   2.3. Pareto analysis is a possibility.
3. Tools for Decision Making
   3.1. Techniques are needed for producing and selecting endpoints and data summaries that reflect the reality of the situation in very simplified terms (example: Green GDP).
   3.2. How robust is the decision based on the quality of the model or models being used.
4. Tools for propagation of errors
   4.1. Generalized likelihood uncertainty estimation (GLUE), probabilistic risk assessment, and uncertainty analysis research.
   4.2. How does one combine qualitative and quantitative uncertainties?
   4.3. May not be able to quantify uncertainty, but you may be able to discuss it by conditioning on an auxiliary variable (by using a variance partition formula).
5. Tools for communication
   5.1. Need tools for communicating conceptualization of model.
   5.2. Also for communicating model application and results to the "non-technical risk manager" (a.k.a. the non-modeler).
   5.3. Problem: Those who do not understand the model may not trust the model.
   5.4. Graphical models/methods for displaying results should be encouraged.

Additional comments from final discussion session September 10:

1. Maintenance of computer code through multiple hardware and software generations is an important topic. When does the accumulation of changes trigger the need for a new assessment of quality? Current state of version control is inadequate.
2. Tools are needed for combining and correctly expressing disparate sources of information funneled into a decision or summary judgment. Procedures are needed to distinguish between rhetoric and candid technical integrity.
3. Science communication should be treated as a specific component of the QA process. This is the responsibility of those who need to understand as well as those who are explaining the process.
4. Standardized task data sets and a “clearing house” for these should be considered.
5. Better methods for addressing space and time scale disparities between models and data are needed.