EPA/100/K-09/003 | March 2009
www.epa.gov/crem

United States
Environmental Protection
Agency

Guidance on the Development, Evaluation, and Application of Environmental Models

Office of the Science Advisor
Council for Regulatory Environmental Modeling

EPA/100/K-09/003
March 2009

Office of the Science Advisor

Guidance on the Development, Evaluation, and Application of Environmental Models

Council for Regulatory Environmental Modeling
U.S. Environmental Protection Agency
Washington, DC 20460

PBN1505

Preface

This *Guidance on the Development, Evaluation, and Application of Environmental Models* was prepared in response to a request by the U.S. Environmental Protection Agency (EPA) Administrator that EPA's Council for Regulatory Environmental Modeling (CREM) help continue to strengthen the Agency's development, evaluation, and use of models (http://www.epa.gov/osp/crem/library/whitman.PDF).

A draft version of this document (http://cfpub.epa.gov/crem/crem_sab.cfm) was reviewed by an independent panel of experts established by EPA's Science Advisory Board and revised by CREM in response to the panel's comments.

This final document is available in printed and electronic form. The electronic version provides direct links to the references identified in the document.

Disclaimer

This document provides guidance to those who develop, evaluate, and apply environmental models. It does not impose legally binding requirements; depending on the circumstances, it may not apply to a particular situation. The U.S. Environmental Protection Agency (EPA) retains the discretion to adopt, on a case-by-case basis, approaches that differ from this guidance.

Authors, Contributors, and Reviewers

This document was developed under the leadership of EPA's Council for Regulatory Environmental Modeling. A number of people representing EPA's core, program, and regional offices helped write and review it.

PRINCIPAL AUTHORS:

Council for Regulatory Environmental Modeling Staff:
Noha Gaber, Gary Foley, Pasky Pascual, Neil Stiber, Elsie Sunderland

EPA Region 10:
Ben Cope

Office of Environmental Information:
Annett Nold (deceased)

Office of Solid Waste and Emergency Response:
Zubair Saleem

CONTRIBUTORS AND INTERNAL REVIEWERS:

EPA Core Offices:
Office of Research and Development:
Justin Babendreier, Thomas Barnwell (retired), Ed Bender, Lawrence Burns (retired), Gary Foley, Kathryn Gallagher, Kenneth Galluppi, Gerry Laniak, Haluk Ozkaynak, Kenneth Schere, Subhas Sikdar, Eric Weber, Joe Williams

Office of Environmental Information:
Ming Chang, Reggie Cheatham, Evangeline Cummings, Linda Kirkland, Nancy Wentworth

Office of General Counsel:
James Nelson (retired), Barbara Pace, Quoc Nguyen, Manisha Patel, Carol Ann Sicilano

Science Advisory Board:
Jack Kooyoomjian

EPA Program Offices:
Office of Air and Radiation:
Tyler Fox, John Irwin (retired), Joe Tikvart, Richard (Chet) Wayland, Jason West

Office of Prevention, Pesticides and Toxic Substances:
Lynn Delpire, Alan Dixon, Wen-Hsiung Lee, David Miller, Vince Nabholz, Steve Nako, Neil Patel, Randolph Perfetti (retired), Scott Prothero, Donald Rodier

Office of Solid Waste and Emergency Response:
Peter Grevatt, Lee Hofmann, Stephen Kroner (retired), Larry Zaragoza

Office of Water:
Jim Carleton, Sharon E. Hayes, Marjorie Wellman, Denise Keehner, Lauren Wisniewski, Lisa McGuire, Mike Messner, James F. Pendergast

EPA Regional Offices:
Region 1:
Brian Hennessey, Michael Kenyon

Region 2:
Kevin Bricke, Rosella O'Connor, Richard Winfield

Region 3:
Alan Cimorelli

Region 4:
Nancy Bethune, Brenda Johnson, Tim Wool

Region 5:
Bertram Frey, Arthur Lubin, Randy Robinson, Stephen Roy, Mary White

Region 6:
James Yarborough

Region 7:
Bret Anderson

Region 10:
David Frank (retired), John Yearsley (retired)

Contents

Executive Summary

In pursuing its mission to protect human health and to safeguard the natural environment, the U.S. Environmental Protection Agency often relies on environmental models. In this guidance, a model is defined as a "*simplification of reality that is constructed to gain insights into select attributes of a particular physical, biological, economic, or social system."*

This guidance provides recommendations for the effective development, evaluation, and use of models in environmental decision making once an environmental issue has been identified. These recommendations are drawn from Agency white papers, EPA Science Advisory Board reports, the National Research Council's *Models in Environmental Regulatory Decision Making*, and peer-reviewed literature. For organizational simplicity, the recommendations are categorized into three sections: *model development*, *model evaluation*, and *model application*.

Model development can be viewed as a process with three main steps: (a) specify the environmental problem (or set of issues) the model is intended to address and develop the conceptual model, (b) evaluate or develop the model framework (develop the mathematical model), and (c) parameterize the model to develop the application tool.

Model evaluation is the process for generating information over the life cycle of the project that helps determine whether a model and its analytical results are of sufficient quality to serve as the basis for a decision. Model quality is an attribute that is meaningful only within the context of a specific model application. In simple terms, model evaluation provides information to help answer the following questions: (a) How have the principles of sound science been addressed during model development? (b) How is the choice of model supported by the quantity and quality of available data? (c) How closely does the model approximate the real system of interest? (d) How well does the model perform the specified task while meeting the objectives set by quality assurance project planning?

Model application (i.e., model-based decision making) is strengthened when the science underlying the model is transparent. The elements of transparency emphasized in this guidance are (a) comprehensive documentation of all aspects of a modeling project (suggested as a list of elements relevant to any modeling project) and (b) effective communication between modelers, analysts, and decision makers. This approach ensures that there is a clear rationale for using a model for a specific regulatory application.

This guidance recommends best practices to help determine when a model, despite its uncertainties, can be appropriately used to inform a decision. Specifically, it recommends that model developers and users: (a) subject their model to credible, objective peer review; (b) assess the quality of the data they use; (c) corroborate their model by evaluating the degree to which it corresponds to the system being modeled; and (d) perform sensitivity and uncertainty analyses. *Sensitivity analysis* evaluates the effect of changes in input values or assumptions on a model's results. *Uncertainty analysis* investigates the effects of lack of knowledge and other potential sources of error in the model (e.g., the "uncertainty" associated with model parameter values). When conducted in combination, sensitivity and uncertainty analysis allow model users to be more informed about the confidence that can be placed in model results. A model's quality to support a decision becomes better known when information is available to assess these factors.

4. Model Evaluation

Summary of Recommendations for Model Evaluation

appropriately used to inform a decision.

- Model evaluation addresses the soundness of the science underlying a model, the quality and quantity of available data, the degree of correspondence with observed conditions, and the appropriateness of a model for a given application.
- Recommended components of the evaluation process include: (a) credible, objective peer review; (b) QA project planning and data quality assessment; (c) qualitative and/or quantitative model corroboration; and (d) sensitivity and uncertainty analyses.
- Quality is an attribute of models that is meaningful only within the context of a specific model application. Determining whether a model serves its intended purpose involves in-depth discussions between model developers and the users responsible for applying for the model to a particular problem.
- Information gathered during model evaluation allows the decision maker to be better positioned to formulate decisions and policies that take into account all relevant issues and concerns.

4.1 Introduction

> *Models will always be constrained by computational limitations, assumptions and knowledge gaps. They can best be viewed as tools to help inform decisions rather than as machines to generate truth or make decisions. Scientific advances will never make it possible to build a perfect model that accounts for every aspect of reality or to prove that a given model is correct in all aspects for a particular regulatory application. These characteristics…suggest that model evaluation be viewed as an integral and ongoing part of the life cycle of a model, from problem formulation and model conceptualization to the development and application of a computational tool.*
>
> — *NRC Committee on Models in the Regulatory Decision Process (NRC 2007)*

The natural complexity of environmental systems makes it difficult to mathematically describe all relevant processes, including all the intrinsic mechanisms that govern their behavior. Thus, policy makers often rely on models as tools to approximate reality when making decisions that affect environmental systems. The challenge facing model developers and users is determining when a model, despite its uncertainties, can be appropriately used to inform a decision. Model evaluation is the process used to make this determination. In this guidance, model evaluation is defined as *the process used to generate information to determine whether a model and its analytical results are of a quality sufficient to serve as the basis for a decision.* Model evaluation is conducted over the life cycle of the project, from development through application.

Box 5: Model Evaluation Versus Validation Versus Verification

Model evaluation should not be confused with model validation. Different disciplines assign different meanings to these terms and they are often confused. For example, Suter (1993) found that among models used for risk assessments, misconception often arises in the form of the question "Is the model valid?" and statements such as "No model should be used unless it has been validated." Suter further points out that "validated" in this context means (a) proven to correspond exactly to reality or (b) demonstrated through experimental tests to make consistently accurate predictions.

Because every model contains simplifications, predictions derived from a model can never be completely accurate and a model can never correspond exactly to reality. In addition, "validated models" (e.g., those that have been shown to correspond to field data) do not necessarily generate accurate predictions of reality for multiple applications (Beck 2002a). Thus, some researchers assert that no model is ever truly "validated"; models can only be invalidated for a specific application (Oreskes et al. 1994). Accordingly, this guidance focuses on process and techniques for *model evaluation* rather than model validation or invalidation.

"Verification" is another term commonly applied to the evaluation process. However, in this guidance and elsewhere, model verification typically refers to model code verification as defined in the model development section. For example, the NRC Committee on Models in the Regulatory Decision Process (NRC 2007) provides the following definition:

> *Verification* refers to activities that are designed to confirm that the mathematical framework embodied in the module is correct and that the computer code for a module is operating according to its intended design so that the results obtained compare favorably with those obtained using known analytical solutions or numerical solutions from simulators based on similar or identical mathematical frameworks.

In simple terms, model evaluation provides information to help answer four main questions (Beck 2002b):

1. How have the principles of sound science been addressed during model development?
2. How is the choice of model supported by the quantity and quality of available data?
3. How closely does the model approximate the real system of interest?
4. How does the model perform the specified task while meeting the objectives set by QA project planning?

These four factors address two aspects of model quality. The first factor focuses on the intrinsic mechanisms and generic properties of a model, *regardless of the particular task to which it is applied*. In contrast, the latter three factors are evaluated in the context of the use of a model *within a specific set of conditions*. Hence, it follows that model quality is an attribute that is meaningful only within the context of a *specific model application*. A model's quality to support a decision becomes known when information is available to assess these factors.

The NRC committee recommends that evaluation of a regulatory model continue throughout the life of a model and that an evaluation plan could:

- Describe the model and its intended uses.
- Describe the relationship of the model to data, including the data for both inputs and corroboration.

- Describe how such data and other sources of information will be used to assess the ability of the model to meet its intended task.
- Describe all the elements of the evaluation plan by using an outline or diagram that shows how the elements relate to the model's life cycle.
- Describe the factors or events that might trigger the need for major model revisions or the circumstances that might prompt users to seek an alternative model. These can be fairly broad and qualitative.
- Identify the responsibilities, accountabilities, and resources needed to ensure implementation of the evaluation plan.

As stated above, the goal of model evaluation is to ensure model quality. At EPA, quality is defined by the Information Quality Guidelines (IQGs) (EPA 2002a). The IQGs apply to all information that EPA disseminates, including models, information from models, and input data (see Appendix C, Box C4: Definition of Quality). According to the IQGs, quality has three major components: integrity, utility, and objectivity. This chapter focuses on addressing the four questions listed above by evaluating the third component, objectivity — specifically, how to ensure the objectivity of information from models by considering their accuracy, bias, and reliability.

- Accuracy, as described in Section 2.4, is the closeness of a measured or computed value to its "true" value, where the "true" value is obtained with perfect information.
- Bias describes any systematic deviation between a measured (i.e., observed) or computed value and its "true" value. Bias is affected by faulty instrument calibration and other measurement errors, systematic errors during data collection, and sampling errors such as incomplete spatial randomization during the design of sampling programs.
- Reliability is the confidence that (potential) users have in a model and its outputs such that they are willing to use the model and accept its results (Sargent 2000). Specifically, reliability is a function of the model's performance record and its conformance to best available, practicable science.

This chapter describes principles, tools, and considerations for model evaluation throughout all stages of development and application. Section 4.2 presents a variety of qualitative and quantitative best practices for evaluating models. Section 4.3 discusses special considerations for evaluating proprietary models. Section 4.4 explains why retrospective analysis of models, conducted after a model has been applied, can be important to improve individual models and regulatory policies and to systematically enhance the overall modeling field. Finally, Section 4.5 describes how the evaluation process culminates in a decision whether to apply the model to decision making. Section 4.6 reviews the key recommendations from this chapter.

4.2 Best Practices for Model Evaluation

The four questions listed above address the soundness of the science underlying a model, the quality and quantity of available data, the degree of correspondence with observed conditions, and the appropriateness of a model for a given application. This guidance describes several "tools" or best practices to address these questions: peer review of models; QA project planning, including data quality assessment; model corroboration (qualitative and/or quantitative evaluation of a model's accuracy and predictive capabilities); and sensitivity and uncertainty analysis. These tools and practices include both qualitative and quantitative techniques:

- Qualitative assessments: Some of the uncertainty in model predictions may arise from sources whose uncertainty cannot be quantified. Examples are uncertainties about the theory underlying the model, the manner in which that theory is mathematically expressed to represent the environmental components, and the theory being modeled. Subjective evaluation of experts may be needed to determine appropriate values for model parameters and inputs that cannot be directly observed or measured (e.g., air emissions estimates). Qualitative assessments are needed for these sources of uncertainty. These assessments may involve expert elicitation regarding the system's behavior and comparison with model forecasts.
- Quantitative assessments: The uncertainty in some sources — such as some model parameters and some input data — can be estimated through quantitative assessments involving statistical uncertainty and sensitivity analyses. These types of analyses can also be used to quantitatively describe how model estimates of current conditions may be expected to differ from comparable field observations. However, since model predictions are not directly observed, special care is needed when quantitatively comparing model predictions with field data.

As discussed previously, model evaluation is an iterative process. Hence, these tools and techniques may be effectively applied throughout model development, testing, and application and should not be interpreted as sequential steps for model evaluation.

Model evaluation should always be conducted using a graded approach that is adequate and appropriate to the decision at hand (EPA 2001, 2002b). This approach recognizes that model evaluation can be modified to the circumstances of the problem at hand and that programmatic requirements are varied. For example, a screening model (a type of model designed to provide a "conservative" or risk-averse answer) that is used for risk management should undergo rigorous evaluation to avoid false negatives, while still not imposing unreasonable data-generation burdens (false positives) on the regulated community. Ideally, decision makers and modeling staff work together at the onset of new projects to identify the appropriate degree of model evaluation (see Section 3.1).

External circumstances can affect the rigor required in model evaluation. For example, when the likely result of modeling will be costly control strategies and associated controversy, more detailed model evaluation may be necessary. In these cases, many aspects of the modeling may come under close scrutiny, and the modeler must document the findings of the model evaluation process and be prepared to answer questions that will arise about the model. A deeper level of model evaluation may also be appropriate when modeling unique or extreme situations that have not been previously encountered.

Finally, as noted earlier, some assessments require the use of multiple, linked models. This linkage has implications for assessing uncertainty and applying the system of models. Each component model as well as the full system of integrated models must be evaluated.

Sections 4.2.1 and 4.2.2, on peer review of models and quality assurance protocols for input data, respectively, are drawn from existing guidance. Section 4.2.3, on model corroboration activities and the use of sensitivity and uncertainty analysis, provides new guidance for model evaluation (along with Appendix D).

Box 6: Examples of Life Cycle Model Evaluation

The value in evaluating a model from the conceptual stage through the use stage is illustrated in a multi-year project conducted by the Organization for Economic Cooperation and Development (OECD). The project sought to develop a screening model that could be used to assess the persistence and long-range transport potential of chemicals. To ensure its effectiveness, the screening model needed to be a consensus model that had been evaluated against a broad set of available models and data.

This project began at a 2001 workshop to set model performance and evaluation goals that would provide the foundation for subsequent model selection and development (OECD 2002). OECD then established an expert group in 2002. This group began its work by developing and publishing a guidance document on using multimedia models to estimate environmental persistence and long-range transport. From 2003 to 2004, the group compared and assessed the performance of nine available multimedia fate and transport models (Fenner et al. 2005; Klasmeier et al. 2006). The group then developed a parsimonious consensus model representing the minimum set of key components identified in the model comparison. They convened three international workshops to disseminate this consensus model and provide an ongoing model evaluation forum (Scheringer et al. 2006).

In this example, more than half the total effort was invested in the conceptual and model formulation stages, and much of the effort focused on performance evaluation. The group recognized that each model's life cycle is different, but noted that attention should be given to developing consensus-based approaches in the model concept and formulation stages. Conducting concurrent evaluations at these stages in this setting resulted in a high degree of buy-in from the various modeling groups.

4.2.1 Scientific Peer Review

Peer review provides the main mechanism for independent evaluation and review of environmental models used by the Agency. Peer review provides an independent, expert review of the evaluation in Section 4.1; therefore, its purpose is two-fold:

- To evaluate whether the assumptions, methods, and conclusions derived from environmental models are based on sound scientific principles.
- To check the scientific appropriateness of a model for informing a specific regulatory decision. (The latter objective is particularly important for secondary applications of existing models.)

Information from peer reviews is also helpful for choosing among multiple competing models for a specific regulatory application. Finally, peer review is useful to identify the limitations of existing models. Peer review is *not* a mechanism to comment on the *regulatory decisions* or policies that are informed by models (EPA 2000c).

Peer review charge questions and corresponding records for peer reviewers to answer those questions should be incorporated into the quality assurance project plan, developed during assessment planning (see Section 4.2.2, below). For example, peer reviews may focus on whether a model meets the objectives or specifications that were set as part of the quality assurance plan (see EPA 2002b) (see Section 3.1).

All models that inform *significant*[2] regulatory decisions are candidates for peer review (EPA 2000c, 1993) for several reasons:

- Model results will be used as a basis for major regulatory or policy/guidance decision making.
- These decisions likely involve significant investment of Agency resources.
- These decisions may have inter-Agency or cross-agency implications/applicability.

Existing guidance recommends that a new model should be scientifically peer-reviewed prior to its first application; for subsequent applications, the program manager should consider the scientific/technical complexity and/or the novelty of the particular circumstances to determine whether additional peer review is needed (EPA 1993). To conserve resources, peer review of "similar" applications should be avoided.

Models used for secondary applications (existing EPA models or proprietary models) will generally undergo a different type of evaluation than those developed with a specific regulatory information need in mind. Specifically, these reviews may deal more with uncertainty about the appropriate application of a model to a specific set of conditions than with the science underlying the model framework. For example, a project team decides to assess a water quality problem using WASP, a well-established water quality model framework. The project team determines that peer review of the model framework itself is not necessary, and the team instead conducts a peer review on their specific application of the WASP framework.

The following aspects of a model should be peer-reviewed to establish scientific credibility (SAB 1993a, EPA 1993):

- Appropriateness of input data.
- Appropriateness of boundary condition specifications.
- Documentation of inputs and assumptions.
- Applicability and appropriateness of selected parameter values.
- Documentation and justification for adjusting model inputs to improve model performance (calibration).
- Model application with respect to the range of its validity.
- Supporting empirical data that strengthen or contradict the conclusions that are based on model results.

To be most effective and maximize its value, external peer review should begin as early in the model *development* phase as possible (EPA 2000b). Because peer review involves significant time and resources, these allocations must be incorporated into components of the project planning and any

[2] Executive Order 12866 (58 FR 51735) requires federal agencies to determine whether a regulatory action is "significant" and therefore, subject to the requirements of the Executive Order, including review by the Office of Management and Budget. The Order defines "significant regulatory action" as one "that is likely to result in a rule that may: (1) Have an annual effect on the economy of $100 million or more or adversely affect in a material way the economy, a sector of the economy, productivity, competition, jobs, the environment, public health or safety, or State, local, or tribal governments or communities; (2) Create a serious inconsistency or otherwise interfere with an action taken or planned by another agency; (3) Materially alter the budgetary impacts of entitlements, grants, user fees, or loan programs or the rights and obligations of recipients thereof; or (4) Raise novel legal or policy issues arising out of legal mandates, the President's priorities, or the principles set forth in [the] Order." Section 2(f).

related contracts. Peer review in the early stages of model development can help evaluate the conceptual basis of models and potentially save time by redirecting misguided initiatives, identifying alternative approaches, or providing strong technical support for a potentially controversial position (SAB 1993a, EPA 1993). Peer review in the later stages of model development is useful as an independent external review of model code (i.e., model verification). External peer review of the *applicability* of a model to a particular set of conditions should be considered well in advance of any decision making, as it helps avoid inappropriate applications of a model for specific regulatory purposes (EPA 1993).

The peer review logistics are left to the discretion of the managers responsible for applying the model results to decision making. Mechanisms for accomplishing external peer review include (but are not limited to):

- Using an ad hoc panel of scientists.[3]
- Using an established external peer review mechanism such as the SAB
- Holding a technical workshop.[4]

Several sources provide guidance for determining the qualifications and number of reviewers needed for a given modeling project (SAB 1993a; EPA 2000c, 1993, 1994a). Key aspects are summarized in Appendix D of this guidance.

4.2.2 Quality Assurance Project Planning and Data Quality Assessment

Like peer review, data quality assessment addresses whether a model has been developed according to the principles of sound science. While some variability in data is unavoidable (see Section 4.2.3.1), adhering to the tenets of data quality assessment described in other Agency guidance[5] (Appendix D, Box D2: Quality Assurance Planning and Data Acceptance Criteria) helps minimize data uncertainty.

Well-executed QA project planning also helps ensure that a model performs the specified task, which addresses the fourth model evaluation question posed in Section 4.1. As discussed above, evaluating the degree to which a modeling project has met QA objectives is often a function of the external peer review process. The *Guidance for Quality Assurance Project Plans for Modeling* (EPA 2002b) provides general information about how to document quality assurance planning for modeling (e.g., specifications

[3] The formation and use of an ad hoc panel of peer reviewers may be subject to the Federal Advisory Committee Act (FACA). Compliance with FACA's requirements is summarized in Chapter Two of the *Peer Review Handbook,* "Planning a Peer Review" (EPA 2000c). Guidance on compliance with FACA may be sought from the Office of Cooperative Environmental Management. Legal questions regarding FACA may be addressed to the Cross-Cutting Issues Law Office in the Office of General Counsel.

[4] Note that a technical workshop held for peer review purposes is not subject to FACA *if the reviewers provide individual opinions.* [Note that there is no "one time meeting" exemption from FACA. The courts have held that even a single meeting can be subject to FACA.] An attempt to obtain group advice, whether it be consensus or majority-minority views, likely would trigger FACA requirements.

[5] Other guidance that can help ensure the quality of data used in modeling projects includes:

- *Guidance for the Data Quality Objectives Process,* a systematic planning process for environmental data collection (EPA 2000a).
- *Guidance on Choosing a Sampling Design for Environmental Data Collection,* on applying statistical sampling designs to environmental applications (EPA 2002c).
- *Guidance for Data Quality Assessment: Practical Methods for Data Analysis,* to evaluate the extent to which data can be used for a specific purpose (EPA 2000b).

or assessment criteria development, assessments of various stages of the modeling process; reports to management as feedback for corrective action; and finally the process for acceptance, rejection, or qualification of the output for use) to conform with EPA policy and acquisition regulations. Data quality assessments are a key component of the QA plan for models.

Both the quality and quantity (representativeness) of supporting data used to parameterize and (when available) corroborate models should be assessed during all relevant stages of a modeling project. Such assessments are needed to evaluate whether the available data are sufficient to support the choice of the model to be applied (question 2, Section 4.1), and to ensure that the data are sufficiently representative of the true system being modeled to provide meaningful comparison to observational data (question 3, Section 4.1).

4.2.3 Corroboration, Sensitivity Analysis, and Uncertainty Analysis

The question "How closely does the model approximate the real system of interest?" is unlikely to have a simple answer. In general, answering this question is not simply a matter of comparing model results and empirical data. As noted in Section 3.1, when developing and using an environmental model, modelers and decision makers should consider what degree of uncertainty is acceptable within the context of a specific model application. To do this, they will need to understand the uncertainties underlying the model. This section discusses three approaches to gaining this understanding:

- Model corroboration (Section 4.2.3.2), which includes all quantitative and qualitative methods for evaluating the degree to which a model corresponds to reality.
- Sensitivity analysis (Section 4.2.3.3), which involves studying how changes in a model's input values or assumptions affect its output or response.
- Uncertainty analysis (Section 4.2.3.3), which investigates how a model might be affected by the lack of knowledge about a certain population or the real value of model parameters.

Where practical, the recommended analyses should be conducted and their results reported in the documentation supporting the model. Section 4.2.3.1 describes and defines the various types of uncertainty, and associated concepts, inherent in the modeling process that model corroboration and sensitivity and uncertainty analysis can help assess.

4.2.3.1 Types of Uncertainty

Uncertainties are inherent in all aspects of the modeling process. Identifying those uncertainties that *significantly* influence model outcomes (either qualitatively or quantitatively) and communicating their importance is key to successfully integrating information from models into the decision making process. As defined in Chapter 3, uncertainty is the term used in this guidance to describe incomplete knowledge about specific factors, parameters (inputs), or models. For organizational simplicity, uncertainties that affect model quality are categorized in this guidance as:

- **Model framework uncertainty,** resulting from incomplete knowledge about factors that control the behavior of the system being modeled; limitations in spatial or temporal resolution; and simplifications of the system.

- **Model input uncertainty,** resulting from data measurement errors, inconsistencies between measured values and those used by the model (e.g., in their level of aggregation/averaging), and parameter value uncertainty.
- **Model niche uncertainty,** resulting from the use of a model outside the system for which it was originally developed and/or developing a larger model from several existing models with different spatial or temporal scales.

Box 7: Example of Model Input Uncertainty

The NRC's *Models in Environmental Regulatory Decision Making* provides a detailed example, summarized below, of the effect of model input uncertainty on policy decisions.

The formation of ozone in the lower atmosphere (troposphere) is an exceedingly complex chemical process that involves the interaction of oxides of nitrogen (NO_x), volatile organic compounds (VOCs), sunlight, and dynamic atmospheric processes. The basic chemistry of ozone formation was known in the early 1960s (Leighton 1961). Reduction of ozone concentrations generally requires controlling either or both NO_x and VOC emissions. Due to the nonlinearity of atmospheric chemistry, selection of the emission-control strategy traditionally relied on air quality models.

One of the first attempts to include the complexity of atmospheric ozone chemistry in the decision making process was a simple observation-based model, the so-called Appendix J curve (36 Fed. Reg. 8166 [1971]). The curve was used to indicate the percentage VOC emission reduction required to attain the ozone standard in an urban area based on peak concentration of photochemical oxidants observed in that area. Reliable NO_x data were virtually nonexistent at the time; Appendix J was based on data from measurements of ozone and VOC concentrations from six U.S. cities. The Appendix J curve was based on the hypothesis that reducing VOC emissions was the most effective emission-control path, and this conceptual model helped define legislative mandates enacted by Congress that emphasized controlling these emissions.

The choice in the 1970s to concentrate on VOC controls was supported by early results from models. Though new results in the 1980s showed higher-than-expected biogenic VOC emissions, EPA continued to emphasize VOC controls, in part because the schedule that Congress and EPA set for attaining the ozone ambient air quality standards was not conducive to reflecting on the basic elements of the science (Dennis 2002).

VOC reductions from the early 1970s to the early 1990s had little effect on ozone concentrations. Regional ozone models developed in the 1980s and 1990s suggested that controlling NO_x emissions was necessary in addition to, or instead of, controlling VOCs to reduce ozone concentrations (NRC 1991). The shift in the 1990s toward regulatory activities focusing on NO_x controls was partly due to the realization that historical estimates of emissions and the effectiveness of various control strategies in reducing emissions were not accurate. In other words, ozone concentrations had not been reduced as much as hoped over the past three decades, in part because emissions of some pollutants were much higher than originally estimated.

Regulations may go forward before science and models are perfected because of the desire to mitigate the potential harm from environmental hazards. In the case of ozone modeling, the model inputs (emissions inventories in this case) are often more important than the model science (description of atmospheric transport and chemistry in this case) and require as careful an evaluation as the evaluation of the model. These factors point to the potential synergistic role that measurements play in model development and application.

In reality, all three categories are interrelated. Uncertainty in the underlying model structure or model framework uncertainty is the result of incomplete scientific data or lack of knowledge about the factors

that control the behavior of the system being modeled. Model framework uncertainty can also be the result of simplifications needed to translate the conceptual model into mathematical terms as described in Section 3.3. In the scientific literature, this type of uncertainty is also referred to as structural error (Beck 1987), conceptual errors (Konikow and Bredehoeft 1992), uncertainties in the conceptual model (Usunoff et al. 1992), or model error/uncertainty (EPA 1997; Luis and McLaughlin 1992). Structural error relates to the mathematical construction of the algorithms that make up a model, while the conceptual model refers to the science underlying a model's governing equations. The terms "model error" and "model uncertainty" are both generally synonymous with model framework uncertainty.

Many models are developed iteratively to update their underlying science and resolve existing model framework uncertainty as new information becomes available. Models with long lives may undergo important changes from version to version. The MOBILE model for estimating atmospheric vehicle emissions, the CMAQ (Community Multi-scale Air Quality) model, and the QUAL2 water quality models are examples of models that have had multiple versions and major scientific modifications and extensions in over two decades of their existence (Scheffe and Morris 1993; Barnwell et al. 2004; EPA 1999c, as cited in NRC 2007).

When an appropriate model framework has been developed, the model itself may still be highly uncertain if the input data or database used to construct the application tool is not of sufficient quality. The quality of empirical data used for both model parameterization and corroboration tests is affected by both uncertainty and variability. This guidance uses the term "data uncertainty" to refer to the uncertainty caused by measurement errors, analytical imprecision, and limited sample sizes during data collection and treatment.

In contrast to data uncertainty, variability results from the inherent randomness of certain parameters, which in turn results from the heterogeneity and diversity in environmental processes. Examples of variability include fluctuations in ecological conditions, differences in habitat, and genetic variances among populations (EPA 1997). Variability in model parameters is largely dependent on the extent to which input data have been aggregated (both spatially and temporally). Data uncertainty is sometimes referred to as reducible uncertainty because it can be minimized with further study (EPA 1997). Accordingly, variability is referred to as irreducible because it can be better characterized and represented but not reduced with further study (EPA 1997).

A model's application niche is the set of conditions under which use of the model is scientifically defensible (EPA 1994b). Application niche uncertainty is therefore a function of the appropriateness of a model for use under a specific set of conditions. Application niche uncertainty is particularly important when (a) choosing among existing models for an application that lies outside the system for which the models were originally developed and/or (b) developing a larger model from several existing models with different spatial or temporal scales (Levins 1992).

The SAB's review of MMSOILS (Multimedia Contaminant Fate, Transport and Exposure Model) provides a good example of application niche uncertainty. The SAB questioned the adequacy of using a screening-level model to characterize situations where there is substantial subsurface heterogeneity or where non-aqueous phase contaminants are present (conditions differ from default values) (SAB 1993b). The SAB considered the MMSOILS model acceptable within its original application niche, but unsuitable for more heterogeneous conditions.

4.2.3.2 Model Corroboration

> *The interdependence of models and measurements is complex and iterative for several reasons. Measurements help to provide the conceptual basis of a model and inform model development, including parameter estimation. Measurements are also a critical tool for corroborating model results. Once developed, models can derive priorities for measurements that ultimately get used in modifying existing models or in developing new ones. Measurement and model activities are often conducted in isolation…Although environmental data systems serve a range of purposes, including compliance assessment, monitoring of trends in indicators, and basic research performance, the importance of models in the regulatory process requires measurements and models to be better integrated. Adaptive strategies that rely on iterations of measurements and modeling, such as those discussed in the 2003 NRC report titled* Adaptive Monitoring and Assessment for the Comprehensive Everglades Restoration Plan, *provide examples of how improved coordination might be achieved.*
>
> *— NRC Committee on Models in the Regulatory Decision Process (NRC 2007)*

Model corroboration includes all quantitative and qualitative methods for evaluating the degree to which a model corresponds to reality. The rigor of these methods varies depending on the type and purpose of the model application. Quantitative model corroboration uses statistics to estimate how closely the model results match measurements made in the real system. Qualitative corroboration activities may include expert elicitation to obtain beliefs about a system's behavior in a data-poor situation. These corroboration activities may move model forecasts toward consensus.

For newly developed model frameworks or untested mathematical processes, formal corroboration procedures may be appropriate. Formal corroboration may involve formulation of hypothesis tests for model acceptance, tests on datasets independent of the calibration dataset, and quantitative testing criteria. In many cases, collecting independent datasets for formal model corroboration is extremely costly or otherwise unfeasible. In such circumstances, model evaluation may be appropriately conducted using a combination of other evaluation tools discussed in this section.

Robustness is the capacity of a model to perform equally well across the full range of environmental conditions for which it was designed (Reckhow 1994; Borsuk et al. 2002). The degree of similarity among datasets available for calibration and corroboration provides insight into a model's robustness. For example, if the dataset used to corroborate a model is identical or statistically similar to the dataset used to calibrate the model, then the corroboration exercise has provided neither an independent measure of the model's performance nor insight into the model's robustness. Conversely, when corroboration data are significantly different from calibration data, the corroboration exercise provides a measure of both model performance and robustness.

Quantitative model corroboration methods are recommended for choosing among multiple models that are available for the same application. In such cases, models may be ranked on the basis of their statistical performance in comparison to the observational data (e.g., EPA 1992). EPA's Office of Air and Radiation evaluates models in this manner. When a single model is found to perform better than others in a given category, OAR recommends it in the *Guidelines on Air Quality Models* as a preferred model for

application in that category (EPA 2003a). If models perform similarly, then the preferred model is selected based on other factors, such as past use, public familiarity, cost or resource requirements, and availability.

Box 8: Example: Comparing Results from Models of Varying Complexity

(From Box 5-4 in NRC's *Models in Environmental Regulatory Decision Making*)

The Clean Air Mercury Rule[6] requires industry to reduce mercury emissions from coal-fired power plants. A potential benefit is the reduced human exposure and related health impacts from methylmercury that may result from reduced concentrations of this toxin in fish. Many challenges and uncertainties affect assessment of this benefit. In its assessment of the benefits and costs of this rule, EPA used multiple models to examine how changes in atmospheric deposition would affect mercury concentrations in fish, and applied the models to assess some of the uncertainties associated with the model results (EPA 2005).

EPA based its national-scale benefits assessment on results from the mercury maps (MMaps) model. This model assumes a linear, steady-state relationship between atmospheric deposition of mercury and mercury concentrations in fish, and thus assumes that a 50% reduction in mercury deposition rates results in a 50% decrease in fish mercury concentrations. In addition, MMaps assumes instantaneous adjustment of aquatic systems and their ecosystems to changes in deposition — that is, no time lag in the conversion of mercury to methylmercury and its bioaccumulation in fish. MMaps also does not deal with sources of mercury other than those from atmospheric deposition. Despite those limitations, the Agency concluded that no other available model was capable of performing a national-scale assessment.

To further investigate fish mercury concentrations and to assess the effects of MMaps' assumptions, EPA applied more detailed models, including the spreadsheet-based ecological risk assessment for the fate of mercury (SERAFM) model, to five well-characterized ecosystems. Unlike the steady-state MMaps model, SERAFM is a dynamic model which calculates the temporal response of mercury concentrations in fish tissues to changes in mercury loading. It includes multiple land-use types for representing watershed loadings of mercury through soil erosion and runoff. SERAFM partitions mercury among multiple compartments and phases, including aqueous phase, abiotic participles (for example, silts), and biotic particles (for example, phytoplankton). Comparisons of SERAFM's predictions with observed fish mercury concentrations for a single fish species in four ecosystems showed that the model under-predicted mean concentrations for one water body, over-predicted mean concentrations for a second water body, and accurately predicted mean concentrations for the other two. The error bars for the observed fish mercury concentrations in these four ecosystems were large, making it difficult to assess the models' accuracy. Modeling the four ecosystems also showed how the assumed physical and chemical characteristics of the specific ecosystem affected absolute fish mercury concentrations and the length of time before fish mercury concentrations reached steady state.

Although EPA concluded that the best available science supports the assumption of a linear relationship between atmospheric deposition and fish mercury concentrations for broad-scale use, the more detailed ecosystem modeling demonstrated that individual ecosystems were highly sensitive to uncertainties in model parameters. The Agency also noted that many of the model uncertainties could not be quantified. Although the case studies covered the bulk of the key environmental characteristics, EPA found that extrapolating the individual ecosystem case studies to account for the variability in ecosystems across the country indicated that those case studies might not represent extreme conditions that could influence how atmospheric mercury deposition affected fish mercury concentrations in

[6] On February 8, 2008, the U.S. Court of Appeals for the District of Columbia Circuit vacated the Clean Air Mercury Rule. The DC Circuit's vacatur of this rule was unrelated to the modeling conducted in support of the rule.

a water body.

This example illustrates the usefulness of investigating a variety of models at varying levels of complexity. A hierarchical modeling approach, such as that used in the mercury analysis, can provide justification for simplified model assumptions or potentially provide evidence for a consistent bias that would negate the assumption that a simple model is appropriate for broad-scale application.

4.2.3.3 Sensitivity and Uncertainty Analysis

Sensitivity analysis is the study of how a model's response can be apportioned to changes in model inputs (Saltelli et al. 2000a). Sensitivity analysis is recommended as the principal evaluation tool for characterizing the most and least important sources of uncertainty in environmental models.

Uncertainty analysis investigates the lack of knowledge about a certain population or the real value of model parameters. Uncertainty can sometimes be reduced through further study and by collecting additional data. EPA guidance (e.g., EPA 1997) distinguishes uncertainty analysis from methods used to account for variability in input data and model parameters. As mentioned earlier, variability in model parameters and input data can be better characterized through further study but is usually not reducible (EPA 1997).

Although sensitivity and uncertainty analysis are closely related, sensitivity is algorithm-specific with respect to model "variables" and uncertainty is parameter-specific. Sensitivity analysis assesses the "sensitivity" of the model to specific parameters and uncertainty analysis assesses the "uncertainty" associated with parameter values. Both types of analyses are important to understand the degree of confidence a user can place in the model results. Recommended techniques for conducting uncertainty and sensitivity analysis are discussed in Appendix D.

The NRC committee pointed out that uncertainty analysis for regulatory environmental modeling involves not only analyzing uncertainty, but also communicating the uncertainties to policy makers. To facilitate communication of model uncertainty, the committee recommends using hybrid approaches in which unknown quantities are treated probabilistically *and* explored in scenario-assessment mode by decision makers through a range of plausible values. The committee further acknowledges (NRC 2007) that:

> *Effective uncertainty communication requires a high level of interaction with the relevant decision makers to ensure that they have the necessary information about the nature and sources of uncertainty and their consequences. Thus, performing uncertainty analysis for environmental regulatory activities requires extensive discussion between analysts and decision makers.*

4.3 Evaluating Proprietary Models

This guidance defines proprietary models as those computer models for which the source code is not universally shared. To promote the transparency with which decisions are made, EPA prefers using non-proprietary models when available. However, the Agency acknowledges there will be times when the use of proprietary models provides the most reliable and best-accepted characterization of a system.

When a proprietary model is used, its use should be accompanied by comprehensive, publicly available documentation. This documentation should describe:

- The conceptual model and the theoretical basis (as described in Section 3.3.1) for the model.
- The techniques and procedures used to verify that the proprietary model is free from numerical problems or "bugs" and that it truly represents the conceptual model (as described in Section 3.3.3).
- The process used to evaluate the model (as described in Section 4.2) and the basis for concluding that the model and its analytical results are of a quality sufficient to serve as the basis for a decision (as described in Section 4.1).
- To the extent practicable, access to input and output data such that third parties can replicate the model results.

4.4 Learning From Prior Experiences — Retrospective Analyses of Models

The NRC Committee on Models in the Regulatory Decision Process emphasized that the final issue in managing the model evaluation process is the learning that comes from examining prior modeling experiences. Retrospective analysis of models is important to individual models and regulatory policies and to systematically enhance the overall modeling field. The committee pointed out that retrospective analyses can be considered from various perspectives:

- They can investigate the systematic strengths and weaknesses that are characteristic of broad classes of models — for example, models of ground water flow, surface water, air pollution, and health risks assessment. For example, a researcher estimated that in 20 to 30 percent of ground water modeling efforts, surprising occurrences indicated that the conceptual model underlying the computer model was invalid (Bredehoeft 2003, 2005, in NRC 2007).

- They can study the processes (for example, approaches to model development and evaluation) that lead to successful model applications.

- They can examine models that have been in use for years to determine how well they work. Ongoing evaluation of the model against data, especially data taken under novel conditions, offers the best chance to identify and correct conceptual errors. This type of analysis is referred to as a model "post-audit" (see Section 5.5)

The results of retrospective evaluations of individual models and model classes can be used to identify priorities for improving models.

Box 9: Example of a Retrospective Model Analysis at EPA
(From Box 4-6 in NRC's *Models in Environmental Regulatory Decision Making*)

EPA's Model Evaluation and Applications Research Branch has been performing a retrospective analysis of the CMAQ model's ability to simulate the change in a pollutant associated with a known change in emissions (A. Gilliland, EPA, personal commun., May 19, 2006 and March 5, 2007). This study, which EPA terms a "dynamic evaluation" study, focuses on a rule issue by EPA in 1998 that required 22 states and the District of Columbia to submit State Implementation Plans providing NO_x emission reductions to mitigate ozone transport in the eastern United States. This rule, known as the NO_x SIP Call, requires emission reductions from the utility sector and large industrial boilers in the eastern and midwestern United States by 2004. Since theses sources are equipped with continuous emission monitoring systems, the NO_x SIP call represents a special opportunity to directly measure the emission changes and incorporate them into model simulations with reasonable confidence.

Air quality model simulations were developed for the summers of 2002 and 2004 using the CMAQ model, and the resulting ozone predictions were compared to observed ozone concentrations. Two series of CMAQ simulations were developed to test two different chemical mechanisms in CMAQ. This allowed an evaluation of the uncertainty associated with the model's representation of chemistry. Since the model's prediction of the relative change in pollutant concentrations provides input for regulatory decision making, this type of dynamic evaluations is particularly relevant to how the model is used.

4.5 Documenting the Model Evaluation

In its *Models in Environmental Regulatory Decision Making* report, the NRC summarizes the key elements of a model evaluation (NRC 2007). This list provides a useful framework for documenting the results of model evaluation as the various elements are conducted during model development and application:

- **Scientific basis.** The scientific theories that form the basis for models.
- **Computational infrastructure.** The mathematical algorithms and approaches used in executing the model computations.
- **Assumptions and limitations.** The detailing of important assumptions used in developing or applying a computational model, as well as the resulting limitations that will affect the model's applicability.
- **Peer review.** The documented critical review of a model or its application conducted by qualified individuals who are independent of those who performed the work, but who collectively have at least equivalent technical expertise to those who performed the original work. Peer review attempts to ensure that the model is technically adequate, competently performed, properly documented, and satisfies established quality requirements through the review of assumptions, calculations, extrapolations, alternate interpretations, methodology, acceptance criteria, and/or conclusions pertaining from a model or its application (modified from EPA 2006).
- **Quality assurance and quality control (QA/QC).** A system of management activities involving planning, implementation, documentation, assessment, reporting, and improvement to ensure that a model and its components are of the type needed and expected for its task and that they meet all required performance standards.
- **Data availability and quality.** The availability and quality of monitoring and laboratory data that can be used for both developing model input parameters and assessing model results.

- **Test cases.** Basic model runs where an analytical solution is available or an empirical solution is known with a high degree of confidence to ensure that algorithms and computational processes are implemented correctly.
- **Corroboration of model results with observations.** Comparison of model results with data collected in the field or laboratory to assess the model's accuracy and improve its performance.
- **Benchmarking against other models.** Comparison of model results with other similar models.
- **Sensitivity and uncertainty analysis.** Investigation of the parameters or processes that drive model results, as well as the effects of lack of knowledge and other potential sources of error in the model.
- **Model resolution capabilities.** The level of disaggregation of processes and results in the model compared to the resolution needs from the problem statement or model application. The resolution includes the level of spatial, temporal, demographic, or other types of disaggregation.
- **Transparency.** The need for individuals and groups outside modeling activities to comprehend either the processes followed in evaluation or the essential workings of the model and its outputs.

4.6 Deciding Whether to Accept the Model for Use in Decision Making

The model development and evaluation process culminates in a decision to accept (or not accept) the model for use in decision making. This decision is made by the program manager charged with making regulatory decisions, in consultation with the model developers and project team. It should be informed by good communication of the key findings of the model evaluation process, including the critical issue of uncertainty. The project team should gain model acceptance before applying the model to decision making to avoid confusion and potential re-work.

5. Model Application

5.1 Introduction

Once a model has been accepted for use by decision makers, it is applied to the problem that was identified in the first stages of the modeling process. Model application commonly involves a shift from the *hindcasting* (testing the model against past observed conditions) used in the model development and evaluation phases to *forecasting* (predicting a future change) in the application phase. This may involve a collaborative effort between modelers and program staff to devise management scenarios that represent different regulatory alternatives. Some model applications may entail trial-and-error model simulations, where model inputs are changed iteratively until a desired environmental condition is achieved.

Using a model in a proposed decision requires that the model application be transparently incorporated into the public process. This is accomplished by providing written documentation of the model's relevant characteristics in a style and format accessible to the interested public, and by sharing specific model files and data with external parties, such as technical consultants and university scientists, upon request. This chapter presents best practices and other recommendations for integrating the results of environmental models into Agency decisions. Section 5.2 describes how to achieve and document a transparent modeling process, Section 5.3 reviews situations when use of multiple models may be appropriate, and Section 5.4 discusses the use of post-audits to determine whether the actual system response concurs with that predicted by the model.

Box 10: Examples of Major EPA Documents That Incorporate a Substantial Amount of Computational Modeling Activities

(From Table 2-2 in NRC's *Models in Environmental Regulatory Decision Making*)

Air Quality

Criteria Documents and Staff Paper for Establishing NAAQS

Summarize and assess exposures and health impacts for the criteria air pollutants (ozone, particulate matter, carbon monoxide, lead, nitrogen dioxide, and sulfur dioxide). Criteria documents include results from exposure and health modeling studies, focusing on describing exposure-response relationships. For example, the particulate matter criteria document placed emphasis on epidemiological models of morbidity and mortality (EPA 2004c). The Staff Paper takes this scientific foundation a step further by identifying the crucial health information and using exposure modeling to characterize risks that serve as the basis for the staff recommendation of the standards to the EPA Administrator. For example, models of the number of children exercising outdoors during those parts of the day when ozone is elevated had a major influence on decisions about the 8-hour ozone national ambient air quality standard (EPA 1996).

State Implementation Plan (SIP) Amendments

A detailed description of the scientific methods and emissions reduction programs a state will use to carry out its responsibilities under the CAA for complying with NAAQS. A SIP typically relies on results from activity, emissions, and air quality modeling. Model-generated emissions inventories serve as input to regional air quality models and are used to test alternative emission-reduction schemes to see whether they will result in air quality standards being met (e.g., ADEC 2001; TCEQ 2004). Regional-scale modeling has become part of developing state implementation plans

for the new 8-hour ozone and fine particulate matter standards. States, local governments, and their consultants do this analysis.

Regulatory Impact Assessments (RIAs) for Air Quality Rules

RIAs for air quality regulations document the costs and benefits of major emission control regulations. Recent RIAs have included emissions, air quality, exposure, and health and economic impacts modeling results (e.g., EPA 2004b)

Water Regulations

Total Maximum Daily Load (TMDL) Determinations

For each impaired water body, a TMDL identifies (a) the water quality standard that is not being attained and the pollutant causing the impairment (b) and the total loading of the pollutant that the water may receive and still meet the water quality standard and (c) allocates that total loading among the point and nonpoint sources of the pollutant discharging to the water. Establishment of TMDLs may utilize water quality and/or nutrient loading models. States establish most TMDLs and therefore state and their consultants can be expected to do the majority of this modeling, with EPA occasionally doing the modeling for particularly contentious TMDLs (EPA 2002b; George 2004; Shoemaker 2004; Wool 2004).

Leaking Underground Storage Tank Program

Assesses the potential risks associated with leaking underground gasoline storage tanks. At an initial screening level, it may assess one-dimensional transport of a conservative contaminant using an analytical model (Weaver 2004).

Development of Maximum Contaminant Levels for Drinking Water

Assess drinking water standards for public water supply systems. Such assessments can include exposure, epidemiology, and dose-response modeling (EPA 2002c; NRC 2001b, 2005b).

Pesticides and Toxic Substances Program

Pre-manufacturing Notice Decisions

Assess risks associated with new manufactured chemicals entering the market. Most chemicals are screened initially as to their environmental and human health risks using structure-activity relationship models.

Pesticide Reassessments

Requires that all existing pesticides undergo a reassessment based on cumulative (from multiple pesticides) and aggregate (exposure from multiple pathways) health risk. This includes the use of pesticide exposure models.

Solid and Hazardous Wastes Regulations

Superfund Site Decision Documents

Includes the remedial investigation, feasibility study, proposed plan, and record-of-decision documents that address the characteristics and cleanup of Superfund sites. For many hazardous waste sites, a primary modeling task is using groundwater modeling to assess movement of toxic substances through the substrate (Burden 2004). The remedial investigation for a mining megasite might include water quality, environmental chemistry, human health risk, and ecological risk assessment modeling (NRC 2005a).

Human Health Risk Assessment

Benchmark Dose (BMD) Technical Guidance Document

EPA relies on both laboratory animal and epidemiological studies to assess the noncancer effects of chronic exposure to pollutants (that is, the reference dose [RfD] and the inhalation reference concentration, [RfC]). These data are modeled to estimate the human dose-response. EPA recommends the use of BMD modeling, which essentially fits the experimental data to use as much of the available data as possible (EPA 2000).

Ecological Risk Assessment

The ecological risk assessment guidelines provide general principles and give examples to show how ecological risk assessment can be applied to a wide range of systems, stressors, and biological, spatial, and temporal scales. They describe the strengths and limitations of alternative approaches and emphasize processes and approaches for analyzing data rather than specifying data collection techniques, methods or models (EPA 1998).

5.2 Transparency

The objective of transparency is to enable communication between modelers, decision makers, and the public. Model transparency is achieved when the modeling processes are documented with clarity and completeness at an appropriate level of detail. When models are transparent, they can be used reasonably and effectively in a regulatory decision.

5.2.1 Documentation

Documentation enables decision makers and other model users to understand the process by which a model was developed and used. During model development and use, many choices must be made and options selected that may bias the model results. Documenting this process and its limitations and uncertainties is essential to increase the utility and acceptability of the model outcomes. Modelers and project teams should document all relevant information about the model to the extent practicable, particularly when a controversial decision is involved. In legal proceedings, the quality and thoroughness of the model's written documentation and the Agency's responses to peer review and public comments on the model can affect the outcome of the legal challenge.

The documentation should include a clear explanation of the model's relationship to the scenario of the particular application. This explanation should describe the limitations of the available information when applied to other scenarios. Disclosure about the state of science used in a model and future plans to update the model can help establish a record of reasoned, evidence-based application to inform decisions. For example, EPA successfully defended a challenge to a model used in its TMDL program when it explained that it was basing its decision on the best available scientific information and that it intended to refine its model as better information surfaced.[7]

When a court reviews EPA modeling decisions, they generally give some deference to EPA's technical expertise, unless it is without substantial basis in fact. As discussed in Section 4.2.3 regarding corroboration, deviations from empirical observations are to be expected. In substantive legal disputes, the courts generally examine the record supporting EPA's decisions for justification as to why the model was reasonable.[8] The record should contain not only model development, evaluation, and application but also the Agency's responses to comments on the model raised during peer review and the public process. The organization of this guidance document offers a general outline for model documentation. Box 11 provides a more detailed outline. These elements are adapted from EPA Region 10's standard practices for modeling projects.

[7] *Natural Resources Defense Council v. Muszynski*, 268 F.3d 91 (2d Cir. 2001).
[8] *American Iron and Steel Inst. v. EPA*, 115 F.3d 979 (D.C. Cir. 1997).

Box 11: Recommended Elements for Model Documentation

1. Management Objectives
- Scope of problem
- Technical objectives that result from management objectives
- Level of analysis needed
- Level of confidence needed

2. Conceptual Model
- System boundaries (spatial and temporal domain)
- Important time and length scales
- Key processes
- System characteristics
- Source description
- Available data sources (quality and quantity)
- Data gaps
- Data collection programs (quality and quantity)
- Mathematical model
- Important assumptions

3. Choice of Technical Approach
- Rationale for approach in context of management objectives and conceptual model
- Reliability and acceptability of approach
- Important assumptions

4. Parameter Estimation
- Data used for parameter estimation
- Rationale for estimates in the absence of data
- Reliability of parameter estimates

5. Uncertainty/Error
- Error/uncertainty in inputs, initial conditions, and boundary conditions
- Error/uncertainty in pollutant loadings
- Error/uncertainty in specification of environment
- Structural errors in methodology (e.g., effects of aggregation or simplification)

6. Results
- Tables of all parameter values used for analysis
- Tables or graphs of all results used in support of management objectives or conclusions
- Accuracy of results

7. Conclusions of analysis in relationship to management objectives

8. Recommendations for additional analysis, if necessary

Note: The QA project plan for models (EPA 2002b) *includes a documentation and records component that also describes the types of records and level of detailed documentation to be kept depending on the scope and magnitude of the project.*

5.2.2 Effective Communication

The modeling process should effectively communicate uncertainty to anyone interested in the model results. All technical information should be documented in a manner that decision makers and stakeholders can readily interpret and understand. Recommendations for improving clarity, adapted from the Risk Characterization Handbook (EPA 2000d), include the following:

- Be as brief as possible while still providing all necessary details.

- Use plain language that modelers, policy makers, and the informed lay person can understand.
- Avoid jargon and excessively technical language. Define specialized terms upon first use.
- Provide the model equations.
- Use clear and appropriate methods to efficiently display mathematical relationships.
- Describe quantitative outputs clearly.
- Use understandable tables and graphics to present technical data (see Morgan and Henrion, 1990, for suggestions).

The conclusions and other key points of the modeling project should be clearly communicated. The challenge is to characterize these essentials for decision makers, while also providing them with more detailed information about the modeling process and its limitations. Decision makers should have sufficient insight into the model framework and its underlying assumptions to be able to apply model results appropriately. This is consistent with QA planning practices that assert that all technical reports must discuss the data quality and any limitations with respect to their intended use (EPA 2000e).

5.3 Application of Multiple Models

As mentioned in earlier chapters, multiple models sometimes apply to a certain decision making need; for example, several air quality models, each with its own strengths and weaknesses, might be applied for regulatory purposes. In other situations, stakeholders may use alternative models (developed by industry and academic researchers) to produce alternative risk assessments (e.g., CARES pesticide exposure model developed by industry). One approach to address this issue is to use multiple models of varying complexities to simulate the same phenomena (NRC 2007). This may provide insight into how sensitive the results are to different modeling choices and how much trust to put in the results from any one model. Experience has shown that running multiple models can increase confidence in the model results (Manno et al. 2008) (see Box 8 in Chapter 4 for an example). However, resource limitations or regulatory time constraints may limit the capacity to fully evaluate all possible models.

5.4 Model Post-Audit

Due to time complexity, constraints, scarcity of resources, and/or lack of scientific understanding, technical decisions are often based on incomplete information and imperfect models. Further, even if model developers strive to use the best science available, scientific knowledge and understanding are continually advancing. Given this reality, decision makers should use model results in the context of an iterative, ever-improving process of continuous model refinement to demonstrate the accountability of model-based decisions. This process includes conducting model post-audits to assess and improve a model and its ability to provide valuable predictions for management decisions. Whereas corroboration (discussed in Section 4.2.3.2) demonstrates the degree to which a model corresponds to past system behavior, a model post-audit assesses its ability to model future conditions (Anderson and Woessner 1992).

A model post-audit involves monitoring the modeled system, after implementing a remedial or management action, to determine whether the actual system response concurs with that predicted by the model. Post-auditing of all models is not feasible due to resource constraints, but targeted audits of commonly used models may provide valuable information for improving model frameworks and/or model parameter estimates. In its review of the TMDL program, the NRC recommended that EPA implement

this approach by selectively targeting “some post-implementation TMDL compliance monitoring for verification data collection to assess model prediction error” (NRC 2001). The post-audit should also evaluate how effectively the model development and use process engaged decision makers and other stakeholders (Manno et al. 2008).

Appendix D: Best Practices for Model Evaluation

D.1 Introduction

This appendix presents a practical guide to the best practices for model evaluation (please see Section 4.1 for descriptions of these practices). These best practices are:

- Scientific peer review (Section 4.1.1)
- Quality assurance project planning (Section 4.1.2)
- Corroboration (Section 4.1.3)
- Sensitivity analysis (Section 4.1.3)
- Uncertainty analysis (Section 4.1.3)

The objective of model evaluation is to determine whether a model is of sufficient quality to inform a regulatory decision. For each of these best practices, this appendix provides a conceptual overview for model evaluation and introduces a suite of "tools" that can be used in partial fulfillment of the best practice. The appropriate use of these tools is discussed and citations to primary references are provided. Users are encouraged to obtain more complete information about tools of interest, including their theoretical basis, details of their computational methods, and the availability of software.

Figure D.1.1 provides an overview of the steps in the modeling process that are discussed in this guidance. Items in bold in the figure, including peer review, model corroboration, uncertainty analysis, and sensitivity analysis, are discussed in this section on model evaluation.

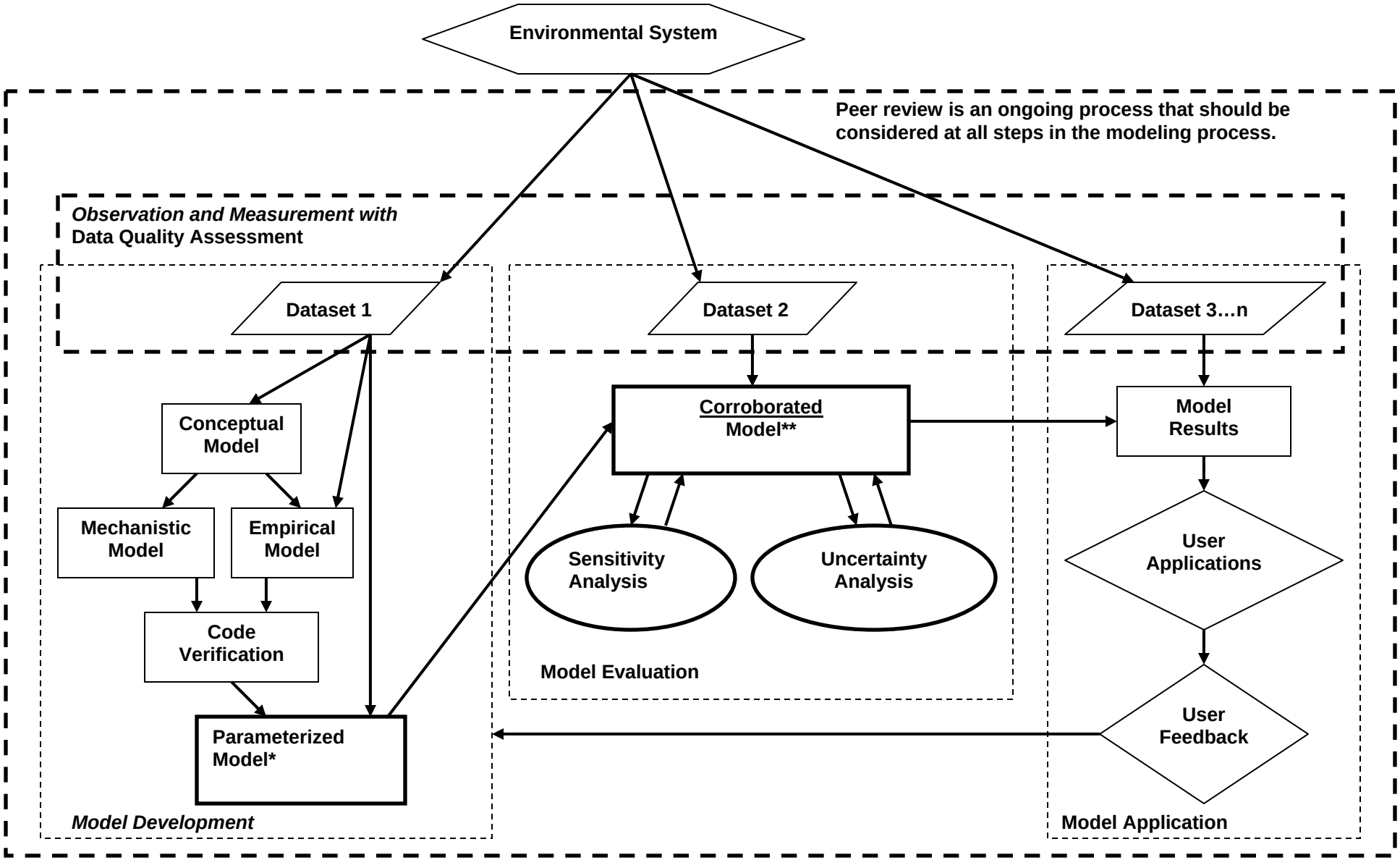


Figure D.1.1. The modeling process.

* In some disciplines parameterization may include, or be referred to as, calibration.

** Qualitative and/or quantitative corroboration should be performed when necessary.

D.2 Scientific Peer Review

EPA policy states that major science-based and technical products related to Agency decisions should normally be peer-reviewed. Agency managers determine and are accountable for the decision whether to employ peer review in particular instances and, if so, its character, scope, and timing. EPA has published guidance for program managers responsible for implementing the peer review process for models (Beck et al. 1994). This guidance discusses peer review mechanisms, the relationship of external peer review to the process of environmental regulatory model development and application, documentation of the peer review process, and specific elements of what could be covered in an external peer review of model development and application.

The general process for external peer review of models is as follows (Beck et al. 1994, Press 1992):

- Step 0: The program manager within the originating office (AA-ship or Region) identifies elements of the regulatory process that would benefit from the use of environmental models. A review/solicitation of currently available models and related research should be conducted. If it is concluded that the development of a new model is necessary, a research/development work plan is prepared.
- Step 0b (optional): The program manager may consider internal and/or external peer review of the research/development concepts to determine whether they are of sufficient merit and whether the model is likely to achieve the stated purpose.
- Step 1: The originating office develops a new or revised model or evaluates the possible novel application of a model developed for a different purpose.
- Step 1b (optional): The program manager may consider internal and/or external peer review of the technical or theoretical basis prior to final development, revision, or application at this stage. For model development, this review should evaluate the stated application niche.
- Step 2: Initial Agency-wide (internal) peer review/consultation of model development and/or proposed application may be undertaken by the developing originating office. Model design, default parameters, etc., and/or intended application are revised (if necessary) based on consideration of internal peer review comments.
- Step 3: The origination office considers external peer review. Model design, default parameters, etc., and/or intended application are revised (if necessary) based on consideration of internal peer review comments.
- Step 4: Final Agency-wide evaluation/consultation may be implemented by the originating office. This step should consist of consideration of external peer review comments and documentation of the Agency's response to scientific/technical issues.

(Note: Steps 2 and 4 are relevant when there is either an internal Agency standing or an ad hoc peer review committee or process).

Box D1: Elements of External Peer Review for Environmental Regulatory Models (Box 2-4 from NRC's *Models in Environmental Regulatory Decision Making*)

Model Purpose/Objectives

- What is the regulatory context in which the model will be used and what broad scientific question is the model intended to answer?
- What is the model's application niche?
- What are the model's strengths and weaknesses?

Major Defining and Limiting Considerations

- Which processes are characterized by the model?
- What are the important temporal and spatial scales?
- What is the level of aggregation?

Theoretical Basis for the Model — formulating the basis for problem solution

- What algorithms are used within the model and how were they derived?
- What is the method of solution?
- What are the shortcomings of the modeling approach?

Parameter Estimation

- What methods and data were used for parameter estimation?
- What methods were used to estimate parameters for which there were no data?
- What are the boundary conditions and are they appropriate?

Data Quality/Quantity

Questions related to model design include:

- What data were utilized in the design of the model?
- How can the adequacy of the data be defined taking into account the regulatory objectives of the model?

Questions related to model application include:

- To what extent are these data available and what are the key data gaps?
- Do additional data need to be collected and for what purpose?

Key Assumptions

- What are the key assumptions?
- What is the basis for each key assumption and what is the range of possible alternatives?
- How sensitive is the model toward modifying key assumptions?

Model Performance Measures

- What criteria have been used to assess model performance?
- Did the data bases used in the performance evaluation provide an adequate test of the model?
- How does the model perform relative to other models in this application niche?

Model Documentation and Users Guide

- Does the documentation cover model applicability and limitations, data input, and interpretation of results?

Retrospective

- Does the model satisfy its intended scientific and regulatory objectives?
- How robust are the model predictions?
- How well does the model output quantify the overall uncertainty?

Source: EPA 1994b.

D.3 Quality Assurance Project Planning

Box D2: Quality Assurance Planning and Data Acceptance Criteria
The QA Project Plan needs to address four issues regarding information on how non-direct measurements are acquired and used on the project (EPA 2002d):

- The need and intended use of each type of data or information to be acquired.
- How the data will be identified or acquired, and expected sources of these data.
- The method of determining the underlying quality of the data.
- The criteria established for determining whether the level of quality for a given set of data is acceptable for use on the project.

Acceptance criteria for individual data values generally address issues such as the following:

Representativeness: Were the data collected from a population sufficiently similar to the population of interest and the model-specified population boundaries? Were the sampling and analytical methods used to generate the collected data acceptable to this project? How will potentially confounding effects in the data (e.g., season, time of day, location, and scale incompatibilities) be addressed so that these effects do not unduly impact the model output?

Bias: Would any characteristics of the dataset directly impact the model output (e.g., unduly high or low process rates)? For example, has bias in analysis results been documented? Is there sufficient information to estimate and correct bias? If using data to develop probabilistic distributions, are there adequate data in the upper and lower extremes of the tails to allow for unbiased probabilistic estimates?

Precision: How is the spread in the results estimated? Is the estimate of variability sufficiently small to meet the uncertainty objectives of the modeling project as stated in Element A7 (Quality Objectives and Criteria for Model Inputs/Outputs) (e.g., adequate to provide a frequency of distribution)?

Qualifiers: Have the data been evaluated in a manner that permits logical decisions on the data's applicability to the current project? Is the system of qualifying or flagging data adequately documented to allow data from different sources to be used on the same project (e.g., distinguish actual measurements from estimated values, note differences in detection limits)?

Summarization: Is the data summarization process clear and sufficiently consistent with the goals of this project (e.g., distinguish averages or statistically transformed values from unaltered measurement values)? Ideally, processing and transformation equations will be made available so that their underlying assumptions can be evaluated against the objectives of the current project.

D.4 Corroboration

In this guidance, "corroboration" is defined as all quantitative and qualitative methods for evaluating the degree to which a model corresponds to reality. In practical terms, it is the process of "confronting models with data" (Hilborn and Mangel 1997). In some disciplines, this process has been referred to as validation. In general, the term "corroboration" is preferred because it implies a claim of usefulness and not truth.

Corroboration is used to understand how consistent the model is with data. However, uncertainty and variability affect how accurately both models and data represent reality because both models and data (observations) are approximations of some system. Thus, to conduct corroboration meaningfully (i.e., as a tool to assess how well a model represents the system being modeled), this process should begin by characterizing the uncertainty and variability in the corroboration data. As discussed in Section 4.1.3.1,

variability stems from the natural randomness or stochasticity of natural systems and can be better captured or characterized in a model but not reduced. In contrast, uncertainty can be minimized with improvements in model structure (framework), improved measurement and analytical techniques, and more comprehensive data for the system being studied. Hence, even a "perfect" model (that contains no measurement error and predicts the correct ensemble average) may deviate from observed field measurements at a given time.

Depending on the type (qualitative and/or quantitative) and availability of data, corroboration can involve hypothesis testing and/or estimates of the likelihood of different model outcomes.

D.4.1 Qualitative Corroboration

Qualitative model corroboration involves expert judgment and tests of intuitive behavior. This type of corroboration uses "knowledge" of the behavior of the system in question, but is not formalized or statistics-based. Expert knowledge can establish model reliability through *consensus* and *consistency*. For example, an expert panel consisting of model developers and stakeholders could be convened to determine whether there is agreement that the methods and outputs of a model are consistent with processes, standards, and results used in other models. Expert judgment can also establish model credibility by determining if model-predicted behavior of a system agrees with best-available understanding of internal processes and functions.

D.4.2 Quantitative Methods

When data are available, model corroboration may involve comparing model predictions to independent empirical observations to investigate how well a model's description of the world fits the observational data. This involves using both statistical measures for goodness of fit and numerical procedures to facilitate these calculations. The can be done graphically or by calculating various statistical measures of fit of a model's results to data.

Recall that a model's *application niche* is the set of conditions under which the use of a model is scientifically defensible (Section 5.2.3); it is the domain of a model's intended applicability. If the model being evaluated purports to estimate an average value across the entire system, then one method to deal with corroboration data is to stratify model results and observed data into "regimes," subsets of data within which system processes operate similarly. Corroboration is then performed by comparing the average of model estimates and observed data within each regime (ASTM 2000).

D.4.2.1 Graphical Methods

Graphical methods can be used to compare the *distribution* of model outputs to independent observations. The degree to which these two distributions overlap, and their respective shapes, provide an indication of model performance with respect to the data. Alternately, the differences between observed and predicted data pairs can be plotted and the resulting probability density function (PDF) used to indicate precisions and bias. Graphical methods for model corroboration can be used to indicate bias, skewness, and kurtosis of model results. Skewness indicates the relative precision of model results, while bias is a reflection of accuracy. Kurtosis refers to the amplitude of the PDF.

D.4.2.2 Deviance Measures

Methods for calculating model bias:

Mean error calculates the average deviation between models and data (e = model-data) by dividing the sum of errors (Σe) by total number of data points compared (m).

$$MeanError = \frac{\Sigma e}{m}$$ (in original measurement units)

Similarly, **mean % error** provides a unit-less measure of model bias:

$$MeanError(\%) = \frac{\Sigma e / s}{m} * 100 ,$$

where "s" is the sample or observational data in original units.

Methods for calculating bias and precision:
Mean square error (MSE):

$$MSE = \frac{\Sigma e^2}{m}$$

(Large deviations in any single data pair (model-data) can dominate this metric.)

Mean absolute error:

$$MeanAbsError = \frac{\Sigma |e|}{m}$$

D.4.2.3 Statistical Tests
A more formal hypothesis testing procedure can also be used for model corroboration. In such cases, a test is performed to determine if the model outputs are statistically significantly different from the empirical data. Important considerations in these tests are the probability of making type I and type II errors and the shape of the data distributions, as most of these metrics assume the data are distributed normally. The test-statistic used should also be based on the number of data-pairs (observed and predicted) available.

There are a number of comprehensive texts that may help analysts determine the appropriate statistical and numerical procedures for conducting model corroboration. These include:

- Efron, B., and R. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Gelman, A.J.B., H.S. Carlin, and D.B. Rubin. 1995. *Bayesian Data Analysis*. New York: Chapman and Hall.
- McCullagh, P., and J.A. Nelder. 1989. *Generalized Linear Models*. New York: Chapman and Hall.
- Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. 1986. *Numerical Recipes*. Cambridge, UK: Cambridge University Press.
- Snedecor, G.W., and W.G. Cochran. 1989. *Statistical Methods*. Eighth Ed. Iowa State University Press.

D.4.3 Evaluating Multiple Models

> *Models are metaphorical (albeit sometimes accurate) descriptions of nature, and there can never be a "correct" model. There may be a "best" model, which is more consistent with the data than any of its competitors, or several models may be contenders because each is consistent in some way with the data and none clearly dominates the others. It is the job of the ecological detective to determine the support that the data offer for each competing model or hypothesis.*
> — Hillborn and Mangel 1997, *Ecological Detective*

In the simplest sense, a first cut of model performance is obtained by examining which model minimizes the sum of squares (SSq) between observed and model-predicted data.

$$SSq = \sum (pred - obs)^2$$

The SSq is equal to the squared differences between model-predicted values and observational values. If data are used to fit models and estimate parameters, the fit will automatically improve with each higher-order model — e.g., simple linear model, $y = a + bX$, vs. a polynomial model, $y = a + bX + cX^2$.

It is therefore useful to apply a penalty for additional parameters to determine if the improvement in model performance (minimizing SSq deviation) justifies an increase in model complexity. The question is essential whether the decrease in the sum of squares is statistically significant.

The SSq is best applied when comparing several models using a single dataset. However, if several datasets are available the Normalized Mean Square Error (NMSE) is typically a better statistic, as it is normalized to the product of the means of the observed and predicted values (see discussion and references, Section D.4.4.4).

D.4.4 An Example Protocol for Selecting a Set of Best Performing Models

During the development phase of an air quality dispersion model and in subsequent upgrades, model performance is constantly evaluated. These evaluations generally compare simulation results using simple methods that do not account for the fact that models only predict a portion of the variability seen in the observations. To fill a part of this void, the U.S. Environmental Protection Agency (EPA) developed a standard that has been adopted by the ASTM International, designation D 6589–00 for Statistical Evaluation of Atmospheric Dispersion Model Performance (ASTM 2000). The following discussion summarizes some of the issues discussed in D 6589.

D.4.4.1 Define Evaluation Objectives

Performing a statistical model evaluation involves defining those evaluation objectives (features or characteristics) within the pattern of observed and modeled concentration values that are of interest to compare. As yet, no one feature or characteristic has been found that can be defined within a concentration pattern that will fully test a model's performance. For instance, the maximum surface concentration may appear unbiased through a compensation of errors in estimating the lateral extent of the dispersing material and in estimating the vertical extent of the dispersing material. Adding into consideration that other biases may exist (e.g., in treatment of the chemical and removal processes during transport, in estimating buoyant plume rise, in accounting for wind direction changes with height, in accounting for penetration of material into layers above the current mixing depth, in systematic variation in all of these biases as a function of atmospheric stability), one can appreciate that there are many ways that a model can falsely give the appearance of good performance.

In principle, modeling diffusion involves characterizing the size and shape of the volume into which the material is dispersing as well as the distribution of the material within this volume. Volumes have three dimensions, so a model evaluation will be more complete if it tests the model's ability to characterize diffusion along more than one of these dimensions.

D.4.4.2 Define Evaluation Procedures

Having selected evaluation objectives for comparison, the next step is to establish an evaluation procedure (or series of procedures), which defines how each evaluation objective will be derived from the available information. Development of statistical model evaluation procedures begins with technical definitions of the terminology used in the goal statement. In the following discussion, we use a plume dispersion model example, but the thought process is valid as well for regional photochemical grid models.

Suppose the evaluation goal is to test models' ability to replicate the average centerline concentration as a function of transport downwind and as a function of atmospheric stability. Several questions must be answered to achieve this goal: What is an "average centerline concentration"? What is "transport downwind"? How will "stability" be defined?

What questions arise in defining the average centerline concentration? Given a sampling arc of concentration values, it is necessary to decide whether the centerline concentration is the maximum value

seen anywhere along the arc or that seen near the center of mass of the observed lateral concentration distribution. If one chooses the latter concept, one needs a definition of how ”near” the center of mass one has to be, to be representative of a centerline concentration value. One might decide to select all values within a specific range (nearness to the center of mass). In such a case, either a definition or a procedure will be needed to define how this specific range will be determined. A decision will have to be made on the treatment of observed zero (and near measurement threshold) concentrations. To discard such values is to say that low concentrations cannot occur near a plume’s center of mass, which is a dubious assumption. One might test to see if conclusions reached regarding the “best performing model” are sensitive to the decision made on the treatment of near-zero concentrations.

What questions arise in defining “transport downwind”? During near-calm wind conditions, when transport may have favored more than one direction over the sampling period, ”downwind” is not well described by one direction. If plume models are being tested, one might exclude near-calm conditions, since plume models are not meant to provide meaningful results during such conditions. If puff models or grid models are being tested, one might sort the near-calm cases into a special regime for analysis.

What questions arise in defining “stability”? For surface releases, surface-layer Monin-Obukhov length, L, has been found to adequately define stability effects; for elevated releases, Z_i/L, where Z_i is the mixing depth, has been found to be a useful parameter for describing stability effects. Each model likely has its own meteorological processor. It is likely that different processors will have different values for L and Z_i for each of the evaluation cases. There is no one best way to deal with this problem. One solution might be to sort the data into regimes using each of the models’ input values, and see if the conclusions reached as to best performing model are affected.

What questions arise if one is grouping data together? If one is grouping data together for which the emission rates are different, one might choose to resolve this difference by normalizing the concentration values by dividing by the respective emission rates. To divide by the emission rate, either one has a constant emission rate over the entire release or the downwind transport is sufficiently obvious that one can compute an emission rate, based on travel time, that is appropriate for each downwind distance.

Characterizing the plume transport direction is highly uncertain, even with meteorological data collected specific for the purpose. Thus, we expect that the simulated position of the plume will not overlap the observed position of the plume. One must decide how to compare a feature (or characteristic) in a concentration pattern, when uncertainties in transport direction are large. Will the observed and modeled patterns be shifted, and if so, in what manner?

This discussion is not meant to be exhaustive, but to be illustrative of how the thought process might evolve. When terms are defined, other questions arise that — when resolved — eventually produce an analysis that will compute the evaluation objective from the available data. There likely is more than one answer to the questions that develop. This may cause different people to develop different objectives and procedures for the same goal. If the same set of models is chosen as the best-performing, regardless of which path is chosen, one can likely be assured that the conclusions reached are robust.

D.4.4.3 Define Trends in Modeling Bias

In this discussion, references to observed and modeled values refer to the observed and model evaluation objectives (e.g., regime averages). A plot of the observed and modeled values as a function of one of the model input parameters is a direct means for detecting model bias. Such comparison has been recommended and employed in a variety of investigations, e.g., Fox (1981), Weil et al. (1992), Hanna (1993) In some cases the comparison is the ratio formed by dividing the modeled value by the observed value, plotted as a function of one or more of the model input parameters. If the data have been stratified into regimes, one can also display the standard error estimates on the respective modeled and observed regime averages. If the respective averages are encompassed by the error bars (typically plus and minus two times the standard error estimates), one can assume the differences are not significant. As Hanna [11] describes, this a “seductive” inference. Procedures to provide a robust assessment of the significance of the differences are defined in ASTM D 6589 (ASTM 2000).

D.4.4.4 Summary of Performance

As an example of overall summary of performance, we will discuss a procedure constructed using the scheme introduced by Cox and Tikvart (1990) as a template. The design for statistically summarizing model performance over several regimes is envisioned as a five-step procedure.

1. Form a replicate sample using concurrent sampling of the observed and modeled values for each regime. Concurrent sampling associates results from all models with each observed value, so that selection of an observed value automatically selects the corresponding estimates by all models.
2. Compute the average of observed and modeled values for each regime.
3. Compute the normalized mean square error, NMSE, using the computed regime averages, and store the value of the NMSE computed for this pass of the bootstrap sampling.
4. Repeat steps 1 through 3 for all bootstrap sampling passes (typically of order 500).
5. Implement the procedure described in ASTM D 6589 (ASTM 2000) to detect which model has the lowest computed NMSE value (call this the “base” model) and which models have NMSE values that are significantly different from the ”base” model.

In the Cox and Tikvart (1990) analysis, the data were sorted into regimes (defined in terms of Pasquill stability category and low/high wind speed classes), and bootstrap sampling was used to develop standard error estimates on the comparisons. The performance measure was the robust highest concentration (computed from the raw observed cumulative frequency distribution), which is a comparison of the highest concentration values (maxima), which most models do not contain the physics to simulate. This procedure can be improved if intensive field data are used and the performance measure is the NMSE computed from the modeled and observed regime averages of centerline concentration values as a function of stability along each downwind arc, where each regime is a particular distance downwind for a defined stability range.

The data demands are much greater for using regime averages than for using individual concentrations. Procedures that analyze groups (regimes) of data include intensive tracer field studies, with a dense receptor network, and many experiments. Whereas, Cox and Tikvart (1990) devised their analysis to make use of very sparse receptor networks having one or more years of sampling results. With dense receptor networks, attempts can be made to compare average modeled and ”observed” centerline concentration values, but only a few of these experiments have sufficient data to allow stratification of the data into regimes for analysis. With sparse receptor networks, there are more data for analysis, but there is insufficient information to define the observed maxima relative to the dispersing plume’s center of mass. Thus, there is uncertainty as to whether or not the observed maxima are representative of centerline concentration values. It is not obvious that the average of the *n* (say 25) observed maximum hourly concentration values (for a particular distance downwind and narrowly defined stability range) is the ensemble average centerline concentration the model is predicting. In fact, one might anticipate that the average of the *n* maximum concentration values is likely to be higher than the ensemble average of the centerline concentration. Thus the testing procedure outlined by Cox and Tikvart (1990) may favor selection of poorly formed models that routinely underestimate the lateral diffusion (and thereby overestimate the plume centerline concentration). This in turn, may bias such models’ ability to characterize concentration patterns for longer averaging times.

It is therefore concluded that once a set of “best-performing models” has been selected from an evaluation using intensive field data that tests a model’s ability to predict the average characteristics to be seen in the observed concentration patterns, evaluations using sparse networks are seen as useful extensions to further explore the performance of well-formulated models for other environs and purposes.

D.5 Sensitivity Analysis

This section provides a broad overview of uncertainty and sensitivity analyses and introduces various methods used to conduct the latter. A table at the end of this section summarizes these methods’ primary features and citations to additional resources for computational detail.

D.5.1 Introducing Sensitivity Analyses and Uncertainty Analysis

A model approximates reality in the face of scientific uncertainties. Section 4.1.3.1 identifies and defines various sources of model uncertainty. External peer reviewers of EPA models have consistently recommended that EPA communicate this uncertainty through uncertainty analysis and sensitivity analysis, two related disciplines. Uncertainty analysis investigates the effects of lack of knowledge or potential errors of model inputs (e.g., the "uncertainty" associated with parameter values); when combined with sensitivity analysis, it allows a model user to be more informed about the confidence that can be placed in model results. Sensitivity analysis measures the effect of changes in input values or assumptions (including boundaries and model functional form) on the outputs (Morgan and Henrion 1990); it is the study of how uncertainty in a model output can be systematically apportioned to different sources of uncertainty in the model input (Beck et al. 1994). By investigating the "relative sensitivity" of model parameters, a user can become knowledgeable of the relative importance of parameters in the model.

Consider a model represented as a function *f*, with inputs x_1 and x_2, and with output *y*, such that $y = f(x_1,x_2)$. Figure D.5.1 schematically depicts how uncertainty analysis and sensitivity analysis would be conducted for this model. Uncertainty analysis would be conducted by determining how *y* responds to variation in inputs x_1 and x_2, the graphic depiction of which is referred to as the model's response surface. Sensitivity analysis would be conducted by apportioning the respective contributions of x_1 and x_2 to changes in *y*. The schematic should *not* be construed to imply that uncertainty analysis and sensitivity analysis are sequential events. Rather, they are generally conducted by trial and error, with each type of analysis informing the other. Indeed, in practice, the distinction between these two related disciplines may be irrelevant. For purposes of clarity, the remainder of this appendix will refer exclusively to sensitivity analysis.

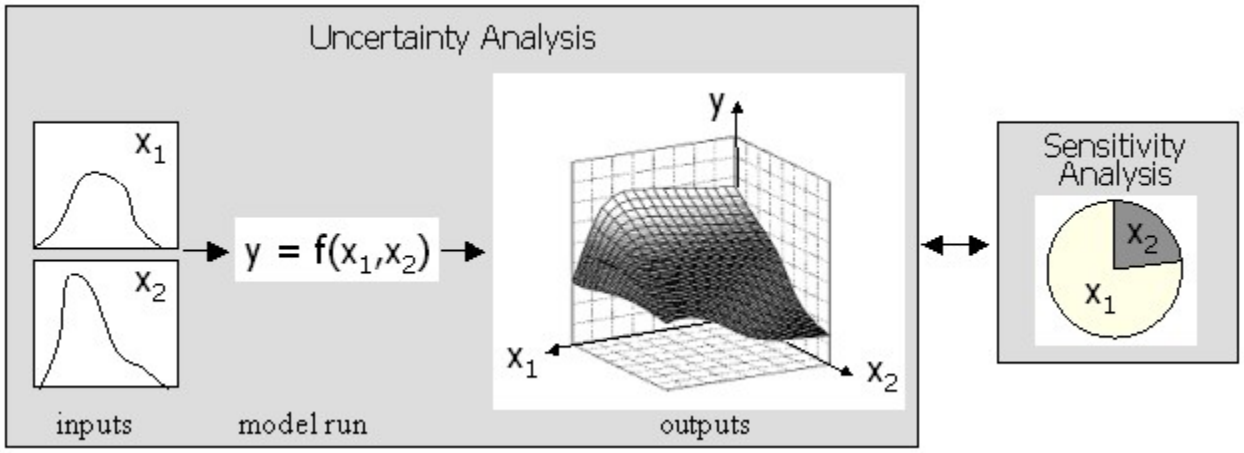


Figure D.5.1. Uncertainty and sensitivity analyses. Uncertainty analysis investigates the effects of lack of knowledge or potential errors of model inputs. Sensitivity analysis evaluates the respective contributions of inputs x_1 and x_2 to output *y*.

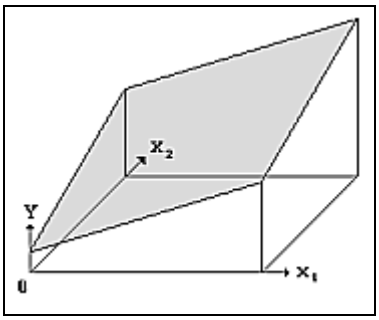
D.5.2 Sensitivity Analysis and Computational Complexity

Choosing the appropriate uncertainty analysis/sensitivity analysis method is often a matter of trading off between the amount of information one wants from the analyses and the computational difficulties of the analyses. These computational difficulties are often inversely related to the number of assumptions one is willing or able to make about the shape of a model's response surface.

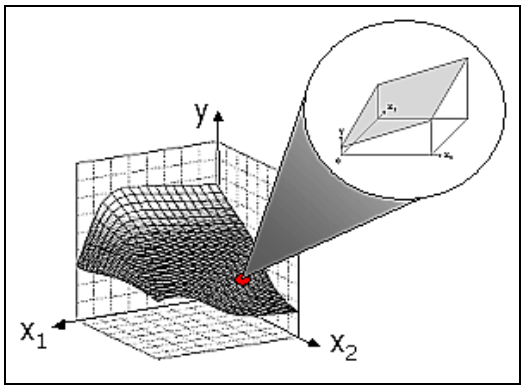
Consider once again a model represented as a function *f*, with inputs x_1 and x_2 and with output *y*, such that $y = f(x_1,x_2)$. *Sensitivity* measures how output changes with respect to an input. This is a straightforward enough procedure with differential analysis if the analyst:

- Can assume that the model's response surface is a hyperplane, as in Figure D.5.2(1);
- Accepts that the results apply only to specific points on the response surface and that these points are monotonic first order, as in Figure D.5.2 (2);[10] or
- Is unconcerned about interactions among the input variables.

Otherwise, sensitivity analysis may be more appropriately conducted using more intensive computational methods.



(1)



(2)

Figure D.5.2. It's hyperplane and simple. (1) A model response surface that is a hyperplane can simplify sensitivity analysis computations. (2) The same computations can also be used for other response surfaces, but only as approximations around a single locus.

This guidance suggests that, depending on assumptions underlying the model, the analyst should use non-intensive sensitivity analysis techniques to initially identify those inputs that generate the most sensitivity, then apply more intensive methods to this smaller subset of inputs. It may therefore be useful to categorize the various sensitivity analysis techniques into methods that (a) can be quickly used to screen for the more important input factors; (b) are based on differential analyses; (c) are based on sampling; and (d) are based on variance methods.

D.5.3 Screening Tools

D.5.3.1 Tools That Require No Model Runs

Cullen and Frey (1999) suggest that summary statistics measuring input uncertainty can serve as preliminary screening tools without additional model runs (and if the models are simple and linear), indicating proportionate contributions to output uncertainty:

- *Coefficient of variation*. The coefficient of variation is the standard deviation normalized to the mean (σ/μ) in order to reduce the possibility that inputs that take on large values are given undue importance.
- *Gaussian approximation.* Another approach to apportioning input variance is Gaussian approximation. Using this method, the variance of a model's output is estimated as the sum of the variances of the inputs (for additive models) or the sum of the variances of the log-transformed inputs (for multiplicative models), weighted by the squares on any constants which may be multiplied by the inputs as they occur in the model.

D.5.3.2 Scatterplots

Cullen and Frey (1999) suggest that a high correlation between an input and an output variable may indicate substantial dependence of the variation in output and the variation of the input. A simple, visual

[10] Related to this issue are the terms "local sensitivity analysis" and "global sensitivity analysis." The former refers to sensitivity analysis conducted around a nominal point of the response surface, while the latter refers to sensitivity analysis across the entire surface.

assessment of the influence of an input on the output is therefore possible using scatterplots, with each plot posing a selected input against the output, as in Figure D.5.3.

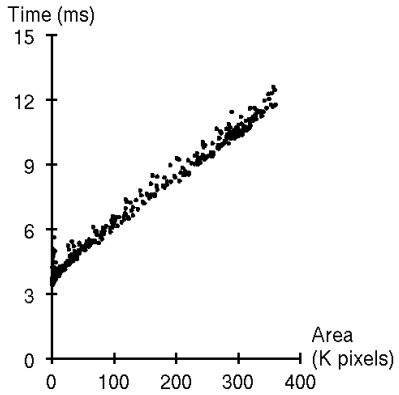


Figure D.5.3. Correlation as indication of input effect. The high correlation between the input variable area and the output variable time (holding all other variables fixed) is an indication of the possible effect of area's variation on the output.

D.5.3.3 Morris's OAT

The key concept underlying one-at-a-time (OAT) sensitivity analyses is to choose a base case of input values and to perturb each input variable by a given percentage away from the base value while holding all other input variables constant. Most OAT sensitivity analysis methods yield *local* measures of sensitivity (see footnote 9) that depend on the choice of base case values. To avoid this bias, Saltelli et al. (2000b) recommend using Morris's OAT for screening purposes because it is a *global* sensitivity analysis method — it entails computing a number of local measures (randomly extracted across the input space) and then taking their average.

Morris's OAT provides a measure of the importance of an input factor in generating output variation, and while it does not quantify interaction effects, it does provide an indication of the presence of interaction. Figure D.5.4 presents the results that one would expect to obtain from applying Morris's OAT (Cossarini et al. 2002). Computational methods for this technique are described in Saltelli et al. 2000b.

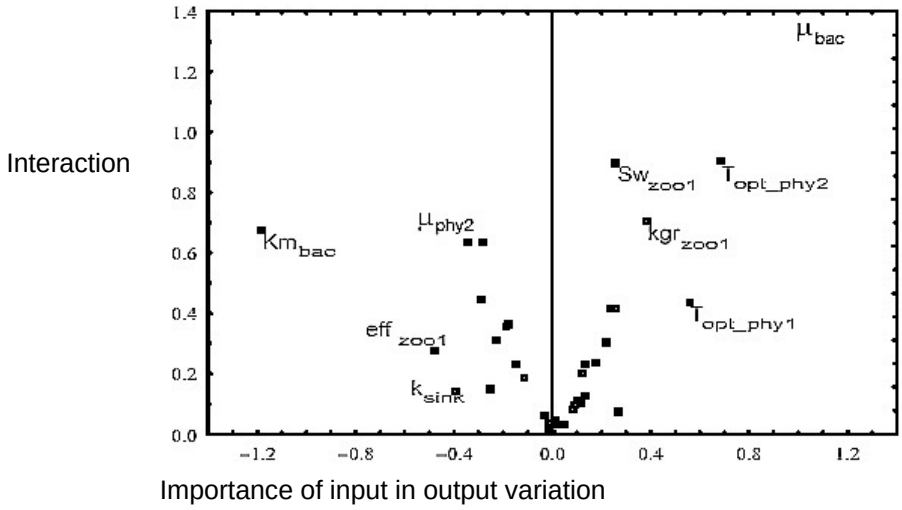


Figure D.5.4. An application of Morris's OAT. Cossarini et al. (2002) investigated the influence of various ecological factors on energy flow through a food web. Their sensitivity analysis indicated that maximum bacteria growth and bacteria mortality (μ_{bac} and Km_{bac}, respectively) have the largest (and opposite) effects on energy flow, as indicated by their values on the horizontal axis. These effects, as indicated by their values on the vertical axis, resulted from interactions with other factors.

D.5.4 Methods Based on Differential Analysis

As noted previously, differential analyses may be used to analyze sensitivity if the analyst is willing either to assume that the model response surface is hyperplanar or to accept that the sensitivity analysis results are local and that they are based on hyperplanar approximations tangent to the response surface at the nominal scenario (Morgan and Henrion 1990; Saltelli et al. 2000b).

Differential analyses entail four steps. First, select base values and ranges for input factors. Second, using these input base values, develop a Taylor series approximation to the output. Third, estimate uncertainty in output in terms of its expected value and variance using variance propagation techniques. Finally, use the Taylor series approximations to estimate the importance of individual input factors (Saltelli et al. 2000b). Computational methods for this technique are described in Morgan and Henrion 1990.

D.5.5 Methods Based on Sampling

One approach to estimating the impact of input uncertainties is to repeatedly run a model using randomly sampled values from the input space. The most well-known method using this approach is Monte Carlo analysis. In a Monte Carlo simulation, a model is run repeatedly. With each run, different input values are drawn randomly from the probability distribution functions of each input, thereby generating multiple output values (Morgan and Henrion 1990; Cullen and Frey 1999). One can view a Monte Carlo simulation as a process through which multiple scenarios generate multiple output values; although each execution of the model run is deterministic, the set of output values may be represented as a cumulative distribution function and summarized using statistical measures (Cullen and Frey 1999).

EPA proposes several best principles of good practice for the conduct of Monte Carlo simulations (EPA 1997). They include the following:

- Conduct preliminary sensitivity analyses to identify significant model components and input variables that make important contributions to model uncertainty.
- When deciding upon a probability distribution function (PDF) for input variables, consider the following questions: Is there any mechanistic basis for choosing a distributional family? Is the PDF likely to be dictated by physical, biological, or other properties and mechanisms? Is the variable

discrete or continuous? What are the bounds of the variable? Is the PDF symmetric or skewed, and if skewed, in which direction?
- Base the PDF on empirical, representative data.
- If expert judgment is used as the basis for the PDF, document explicitly the reasoning underlying this opinion.
- Discuss the presence or absence of covariance among the input variables, which can significantly affect the output.

The preceding points merely summarize some of the main points raised in EPA's Guidance on Monte Carlo Analysis. That document should be consulted for more detailed guidance. Conducting Monte Carlo analysis may be problematic for models containing a large number of input variables. Fortunately, there are several approaches to dealing with this problem:

- *Brute force approach*. One approach is to increase sheer computing power. For example, EPA's ORD is developing a Java-based tool that facilitates Monte Carlo analyses across a cluster of PCs by harnessing the computing power of multiple workstations to conduct multiple runs for a complex model (Babendreier and Castleton 2002).
- *Smaller, structured trials*. The value of Monte Carlo lies not in the randomness of sampling, but in achieving representative properties of sets of points in the input space. Therefore, rather than sampling data from entire input space, computations may be through *stratified sampling* by dividing the input sample space into strata and sampling from within each stratum. A widely used method for stratified sampling is *Latin hypercube sampling,* comprehensively described in Cullen and Frey 1999.
- *Response surface model surrogate*. The analyst may also choose to conduct Monte Carlo not on the complex model directly, but rather on a response surface representation of it. The latter is a simplified representation of the relationship between a selected number of model outputs and a selected number of model inputs, with all other model inputs held at fixed values (Morgan and Henrion 1990; Saltelli et al. 2000b).

D.5.6 Methods Based on Variance

Consider once again a model represented as a function f, with inputs x_1 and x_2 and with output y, such that $y = f(x_1,x_2)$. The input variables are affected by uncertainties and may take on any number of possible values. Let X denote an input vector randomly chosen from among all possible values for x_1 and x_2. The output y for a given X can also be seen as a realization of a random variable Y. Let $E(Y|X)$ denote the expectation of Y conditional on a fixed value of X. If the total variation in y is matched by the variability in $E(Y|X)$ as x_1 is allowed to vary, this is an indication that variation in x_1 significantly affects y.

The variance-based approaches to sensitivity analysis are based on the estimation of what fraction of total variation of y is attributable to variability in $E(Y|X)$ as a subset of input factors are allowed to vary. Three methods for computing this estimation (correlation ratio, Sobol, and Fourier amplitude sensitivity test) are featured in Saltelli et al. 2000b.

D.5.7 Which Method to Use?

A panel of experts was recently assembled to review various sensitivity analysis methods. The panel refrained from explicitly recommending a "best" method and instead developed a list of attributes for preferred sensitivity analysis methods. The panel recommended that methods should preferably be able to deal with a model regardless of assumptions about a model's linearity and additivity, consider interaction effects among input uncertainties, cope with differences in the scale and shape of input PDFs, cope with differences in input spatial and temporal dimensions, and evaluate the effect of an input while all other inputs are allowed to vary as well (Frey 2002; Saltelli 2002). Of the various methods discussed above, only those based on variance (Section D.5.6) are characterized by these attributes. When one or more of the criteria are not important, the other tools discussed in this section will provide a reasonable sensitivity assessment.

As mentioned earlier, choosing the most appropriate sensitivity analysis method will often entail a trade-off between computational complexity, model assumptions, and the amount of information needed from

the sensitivity analysis. As an aid to sensitivity analysis method selection, the table below summarizes the features and caveats of the methods discussed above.

Method	Features	Caveats	Reference
Screening methods	May be conducted independent of model run	Potential for significant error if model is non-linear	Cullen and Frey 1999, pp. 247-8.
Morris's one-at-a-time	Global sensitivity analysis	Indicates, but does not quantify interactions	Saltelli et al. 2000b, p. 68.
Differential analyses	Global sensitivity analysis for linear model; local sensitivity analysis for nonlinear model	No treatment of interactions among inputs Assumes linearity, monotonicity, and continuity	Cullen and Frey 1999, pp. 186-94. Saltelli et al. 2000b, pp. 183-91
Monte Carlo analyses	Intuitive No assumptions regarding response surface	Depending on number of input variables, may be time-consuming to run, but methods to simplify are available May rely on assumptions regarding input PDFs	Cullen and Frey 1999, pp. 196-237 Morgan and Henrion 1990, pp. 198-216.
Variance-based	Robust and independent of model assumptions Addresses interactions	May be computationally difficult.	Saltelli et al. 2000b, pp. 167-97

D.6 Uncertainty Analysis

D.6.1 Model Suitability

An evaluation of model suitability to resolve application niche uncertainty (Section 4.1.3.1) should precede any evaluation of data uncertainty and model performance. The extent to which a model is suitable for a proposed application depends on:

- Mapping of model attributes to the problem statement
- The degree of certainty needed in model outputs
- The amount of reliable data available or resources available to collect additional data
- Quality of the state of knowledge on which the model is based
- Technical competence of those undertaking simulation modeling

Appropriate data should be available before any attempt is made to apply a model. A model that needs detailed, precise input data should not be used when such data are unavailable.

D.6.2 Data Uncertainty

There are two statistical paradigms that can be adopted to summarize data. The first employs classical statistics and is useful for capturing the most likely or "average" conditions observed in a given system. This is known as the "frequentist" approach to summarizing model input data. Frequentist statistics rely on measures of central tendency (median, mode, mean values) and represent uncertainty as the deviation from these metrics. A frequentist or "deterministic" model produces a single set of solutions for each model run. In contrast, the alternate statistical paradigm employs a probabilistic framework, which summarizes data according to their "likelihood" of occurrence. Input data are represented as distributions rather than a single numerical value and models outputs capture a range of possible values.

The classical view of probability defines the probability of an event occurring by the value to which the long run frequency of an event or quantity converges as the number of trials increases (Morgan and Henrion 1990). Classical statistics relies on measures of central tendency (mean, median, mode) to

define model parameters and their associated uncertainty (standard deviation, standard error, confidence intervals).

In contrast to the classical view, a subjectivist or Bayesian view is that the probability of an event is the current degree of belief that a person has that it will occur, given all of the relevant information currently known to that person. This framework involves the use of probability distributions based on likelihoods functions to represent model input values and employs techniques like Bayesian updating and Monte Carlo methods as statistical evaluation tools (Morgan and Henrion 1990).