Draft Guidance on the Development, Evaluation, and Application of Regulatory Environmental Models

Prepared by:

The Council for Regulatory Environmental Modeling

Principal Authors:

Pasky Pascual
Neil Stiber
Elsie Sunderland

Office of Science Policy,
Office of Research and Development
Washington, D.C. 20460

Contributors:

Thomas Barnwell, ORD
Ed Bender, ORD
Nancy Bethune, R4
Lawrence Burns, ORD
Ming Chang, OEI
Alan Cimorelli, R3
Ben Cope, R10
Evangeline Cummings, OEI
Lynn Delpire, OPPTS
Alan Dixon, OPPTS
David Frank, R10
Bertram Frey, ORC
Kathryn Gallagher, ORD
Kenneth Galluppi, ORD
Sharon E. Hayes, OW
Brian Hennessey, R1
John Irwin, OAR
Marjorie Jones, OW
Brenda Johnson, R4
Linda Kirkland, OEI
Stephen Kroner, OSWER
Gerry Laniak, ORD
Wen-Hsiung Lee, OPPT
Lauren MacWilliams, OW
Lisa McGuire, OW
Mike Messner, OW
Vince Nabholz, OPPT
James Nelson, OGC
Rosella O'Connor, R2
Barbara Pace, OGC
James F. Pendergast, OW
Randolph Perfetti, OPPTS
Scott Prothero, OPPTS
Randy Robinson, R5
Donald Rodier, OPPTS
Stephen Roy, R5
Kenneth Schere, ORD
Subhas Sikdar, ORD
Joe Tikvart, OAR
Nancy Wentworth, OEI
Jason West, OAR
Mary White, R5
Joe Williams, ORD
Richard Winfield, R2
Tim Wool, R4
James Yarborough, R6
John Yearsley, R10
Larry Zaragoza, OSWER

Note to Reviewers

This draft Guidance Document on Environmental Models was prepared in response to the EPA Administrator's request that the EPA Council for Regulatory Environmental Modeling (CREM) help continue to strengthen EPA's development, evaluation and use of models. (Please see http://www.epa.gov/osp/crem/library/whitman.PDF.) An independent panel of experts established by EPA's Science Advisory Board is currently reviewing this draft Guidance Document (Please see http://cfpub.epa.gov/crem/crem_sab.cfm). Following this independent review, the CREM intends to make any necessary changes to the draft Guidance Document and to ask for public comments on the resulting final draft through a Federal Register Notice.

****

Foreword

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

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, the definition of a model is a representation of the behavior of an object or process, often in mathematical or statistical terms. This Guidance provides recommendations for environmental models drawn from Agency white papers, EPA's Science Advisory Board reports, and peer-reviewed literature. For organizational simplicity, these recommendations are categorized into the following sections: *model development*, *model evaluation*, and *model application*.

Model Development can be viewed as a process that is achieved by following four main steps: (a) identify the environmental issue (or set of issues) the model is intended to address; (b) develop the conceptual model; (c) construct the model framework (develop the mathematical model), and (d) 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 to determine whether a model and its analytical results are of a quality sufficient 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 assess the following factors: (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 QA 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) and when conducted in combination with sensitivity analysis allows a model user to be more informed about the confidence that can be placed in model results. A model's quality to support a decision becomes known when information is available to assess these factors.

Table of Contents

1.0 Introduction

The mission of the U.S. Environmental Protection Agency is to protect human health and to safeguard the natural environment — air, water, and land — upon which life depends [1]. Environmental models are one source of information for decision makers who need to consider many competing objectives (Fig. 1). For the purposes of this guidance, a model is a representation of the behavior of an object or process, often in mathematical or statistical terms [2]. EPA uses a wide range of models to inform decisions that are important for human health and the environment, including: atmospheric and indoor air models, chemical equilibrium models, economic models, exposure models, leaching and runoff models, multi-media models, risk assessment models, ground water and surface water models, and toxicokinetic models. These models range from simple to complex and may employ a combination of scientific, economic, socio-economic, or other types of data.

Models have a number of other useful applications outside of the regulatory context. For example, because models include explicit mathematical statements about system mechanics, they serve as research tools for exploring new scientific issues and screening tools for simplifying and/or refining existing scientific paradigms or software [3, 4]. Models can also help to study the behavior of ecological systems, design field studies, interpret data, and generalize results.

All underlined terms in this document are defined in the Glossary (Appendix A: Glossary of Frequently Used Terms).

1.1 Purpose

This Guidance (hereon referred to as "guidance") presents recommendations that are drawn from Agency white papers on environmental modeling, EPA Science Advisory Board (SAB) reports, and peer-reviewed literature. It provides an overview of *best practices* for evaluating the quality of environmental models. These practices, in turn, support the execution of mandatory Agency quality assurance (QA) processes for planning, implementing, and assessing of modeling projects that produce quality documentation (Box 1: Background on EPA Quality System).

QA plans should contain performance criteria or ("specifications") for a model in the context of their intended use that are developed at the onset of each project. When documented using a series of associated tests of model quality ("checks") during model evaluation, these specifications provide a record of how well a model meets its intended use. These checks are the basis for a decision on model acceptability and provide a description of model pedigree.

The purpose of this guidance is to provide specific advice on how these "checks" are best performed throughout the development, evaluation and application of models. Following the best practices emphasized in this document in conjunction with a well-documented QA project plan helps to ensure that information EPA disseminates in support of model development and decisions that are informed by models heed the principles set by the Agency's Information Quality Guidelines [5].

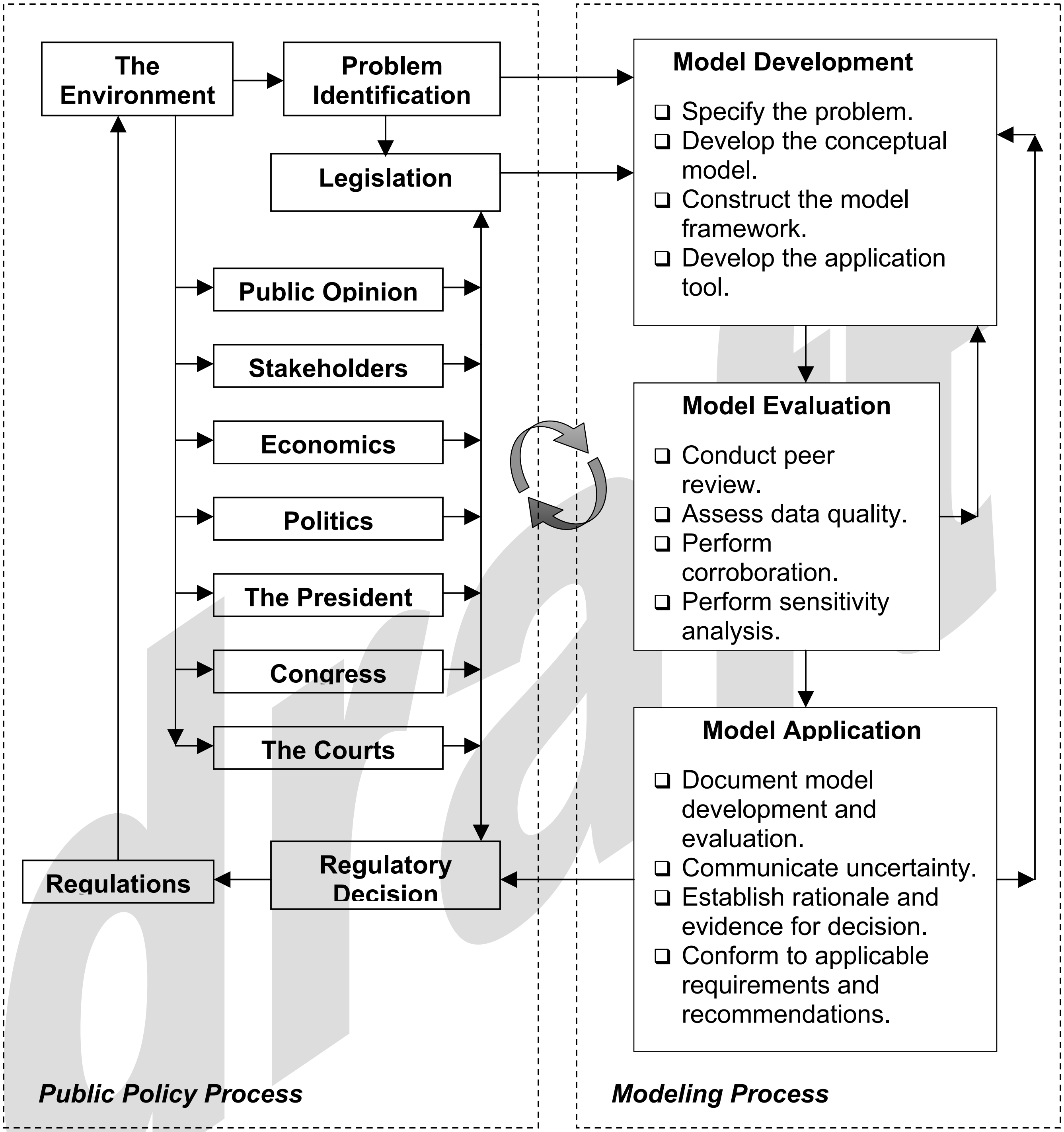


Figure 1. The Role of Modeling in the Public Policy Process. This guidance recommends best practices to develop, evaluate, and apply models that are to be used in the public policy process.

1.2 Intended Audience

This guidance is intended for model developers, computer programmers, model users, and policy makers who work with models used for EPA decisions.

The main body of this document provides a general overview of some principles of good modeling for all users. The appendices contain technical information and examples that may be more appropriate for specific user groups. As indicated by the use of non-mandatory language such as "may," "should," and "can," this guidance provides recommendations and suggestions and does not create legal rights or impose legally binding requirements on EPA or the public.

1.3 Scope of Guidance

The principles and practices described in the guidance are intended to be generally applicable to all models that are used to inform Agency decisions, regardless of domain, mode, conceptual basis, or form [6]. Within the Agency, models are used to simulate many different processes: natural (chemical, physical, and biological) systems, economic phenomena, and decision processes. Models can be designed to represent phenomena in different modes. Prognostic (or predictive) models are designed to forecast outcomes and future events, while diagnostic models work "backwards" to assess causes and precursor conditions. Due to varying levels of complexity and available knowledge for the systems being modeled, the conceptual basis for models is either empirical or mechanistic. Regardless of the domain, mode, and conceptual basis of models, they can generally be represented and solved in different forms: analytic, stochastic, and simulation.

1.4 Organizational Framework

For organizational simplicity, the main body of this guidance is divided into three sections on model development, model evaluation and model application (Fig. 1). However, evaluating a model and its input data to ensure their quality is in fact a process that should be undertaken and documented during all stages of model development and application.

Chapter 1 serves as a general introduction and outlines the scope of this guidance. Chapters 2 and 3 provide guidance on elements of model development and evaluation, respectively. Finally, Chapter 4 recommends practices for most effectively incorporating information from environmental models into policy decisions made by the Agency. The role of information from models in Agency decisions is illustrated in Figure 1. Several appendices referred to in the text contain more detailed technical information and examples that complement each of these chapters. As mentioned above, Appendix A is a glossary with definitions for all of the underlined terms in the guidance. Appendix B presents additional background information on the QA program and other relevant topics. Appendix C presents an overview of best practices that may be used to evaluate models including more detailed information on the peer review process for models and specific technical guidance on tools for model evaluation.

2.0 Model Development

Model development can be viewed as a four-step process: (a) identify the environmental issue (or set of issues) the model is intended to address; (b) develop the conceptual model; (c) construct the model framework (develop the mathematical model), and (d) parameterize the model to develop the application tool. Each step in this process provides opportunities for feedback and iteration.

As defined in this guidance, a "model" is an application tool, while a "model framework" is the system of governing equations. For many Agency models, the application tool is synonymous with the model framework (e.g., Office of Air and Radiation (OAR) air quality models). For other Offices, (e.g., Office of Water), it is common practice to calibrate an existing model framework to different water bodies and ecosystems to develop site-specific tools. Finally, for other Agency applications the main purpose of a model is data retrieval and parameterization. In such cases the model is only partially parameterized as part of the model development process.

Recognizing the diversity of modeling applications throughout the Agency, the following sections are intended to outline some general principles that support the development process. These principles complement the systematic QA planning process for modeling projects that is outlined in existing guidance [7].

2.1 Problem Identification

Modeling projects should clearly state the problem, or set of problems, of interest and describe how model output(s) will inform regulatory decisions [7]. This ideally is a collaborative effort among model developers, intended users, and decision makers but will not be possible in some instances. In all cases, the model's documentation should provide a clear understanding of why and how the model will and can be used. The modeling literature should be reviewed before developing a new model to determine whether an existing model meets the identified needs and requirements.

It is useful to qualitatively or quantitatively specify the acceptable range of uncertainty during the problem identification stage. Uncertainty is the term used in this guidance to describe *lack of knowledge* about models, parameters, constants, data, and beliefs. Defining the ranges of acceptable uncertainty helps project planners generate "specifications" for quality assurance planning and partially determines the appropriate boundary conditions and complexity for the model being developed. The spatial and temporal domain of a model is specified to determine the boundary conditions of the model. Boundary conditions are sets of values for state variables and their rates along problem domain boundaries, sufficient to determine the state of the system within the problem domain.

Data Quality Objectives (DQOs)[1] [8] enable the development of specifications for model quality and associated checks (Appendix B, Box 1: Background on EPA Quality System) during the

[1] The DQOs provide guidance on how to state data needs when limiting decision errors (false positives or false negatives) relative to a given decision. False rejection decision errors (false positives) occur when the null-hypothesis (or baseline condition) is incorrectly rejected based on the sample data. The decision is made assuming the alternate condition or hypothesis to be true when in reality it is false. False acceptance decision errors (false negatives) occur when the null hypothesis (or baseline condition) cannot be rejected based on the available sample data. The decision is made assuming the baseline condition is true when in reality it is false.

problem identification stage. The DQOs provide guidance on how to state data needs when limiting decision errors (false positives or false negatives) relative to a given decision. False rejection decision errors (false positives) occur when the null-hypothesis (or baseline condition) is incorrectly rejected based on the sample data. The decision is made assuming the alternate condition or hypothesis to be true when in reality it is false. False acceptance decision errors (false negatives) occur when the null hypothesis (or baseline condition) cannot be rejected based on the available sample data. The decision is made assuming the baseline condition is true when in reality it is false.

Well-defined DQOs are invaluable during later stages of model development (e.g., calibration and verification) and often attract the attention of peer reviewers. Included in these objectives should be a statement about the acceptable level of total uncertainty that will still enable model results to be used for the intended purpose (Appendix B, Box 2: Configuration Tests Specified in the QA program).

2.2 Conceptual Model Development

Once the need for a model has been identified, a conceptual model should be developed to represent the most important behaviors of the object or process relevant to the problem of interest. Literature, fieldwork, applicable anecdotal evidence, and relevant historical modeling projects should be reviewed while developing the conceptual model.

The model developer should present a clear statement and description (in words, functional expressions, diagrams, and/or graphs) of each element of the conceptual model. For each element, the modeler should document the science behind the conceptual model (e.g., laboratory experiments, mechanistic evidence, empirical data supporting the hypothesis, peer reviewed literature) in mathematical form, when possible. To the extent feasible, the modeler should provide information on assumptions, scale, feedback mechanisms, and static/dynamic behaviors. When it is relevant, an appraisal of the strengths and weaknesses of each constituent hypothesis should be provided.

2.3 Model Framework Construction

The model framework is a formal mathematical specification of the concepts and procedures of the conceptual model. The mathematical framework is usually translated into a form amenable to running on a computer (i.e., algorithm development and model coding). Several issues considered during model framework construction include:

- Does sound science support the underlying hypothesis? (Sound science may be defined in part, but is not restricted to, peer reviewed theory and equations.)
- Is the complexity of the model appropriate for the problem at hand?
- Do the quality and quantity of data support the choice of model?
- Does the model structure reflect all the relevant inputs based on the conceptual model?
- Has the model code been developed? If so, has it been verified?

If multiple viable competing hypotheses about the system behavior exist, it may be useful to statistically compare the performance of these competing models with observational, field, or laboratory data (Chapter 3). The principles of scientific hypothesis testing [9] should be applied to model development using an iterative approach to model evaluation [10].

2.3.1 Model Complexity

Due to the inherent trade-off between model framework uncertainty and data uncertainty, an optimal level of model complexity exists for every model (Fig. 2).

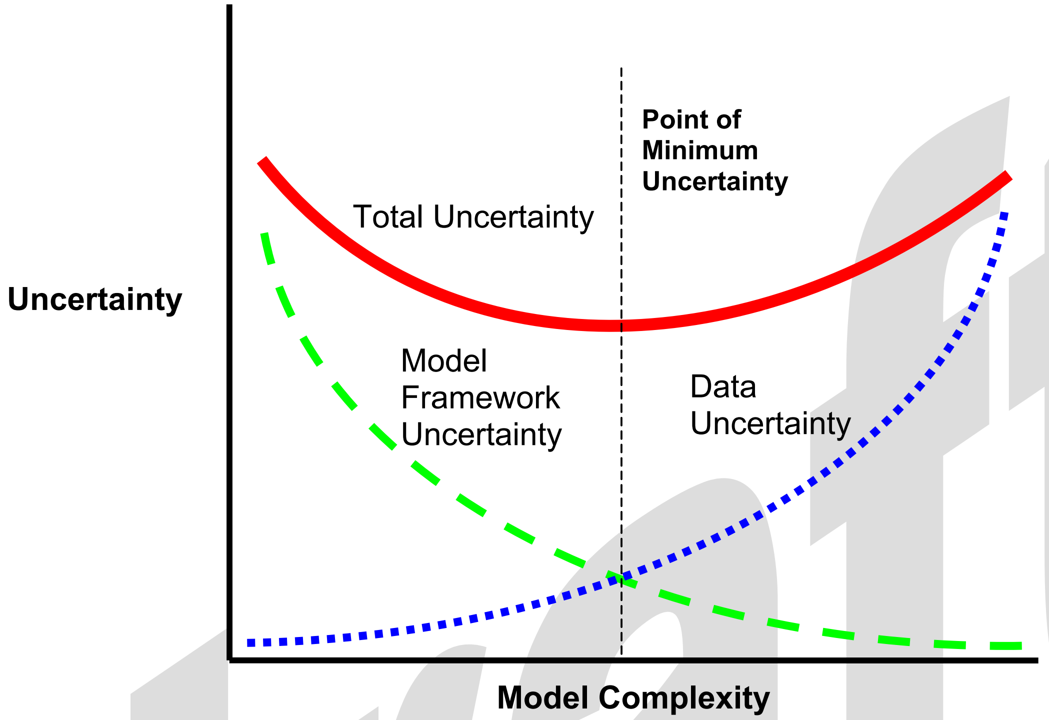


Figure 2. Illustration of relationship between model framework uncertainty and data uncertainty, and the combined effect on total model uncertainty. In this guidance, uncertainty is the term used to describe incomplete knowledge about specific factors, parameters (inputs) or models. Model framework uncertainty is a function of the soundness of the underlying scientific foundations of the model. Data uncertainty measurement errors, analytical imprecision and limited sample sizes during the collection and treatment of data that are used to characterize model parameters. An optimal level of complexity exists for every model, which is represented in the figure as the "point of minimum uncertainty." (Adapted from Hanna [11]).

For example, within the field of air-quality modeling it is sometimes necessary to compromise when choosing among the physical processes that will be treated explicitly in the model. If the objective is to estimate the pattern of concentration values in the near vicinity of one (or several) source(s), then typically chemistry is of little importance. For such situations, travel times from sources to receptors are often too short for chemical formation and destruction to greatly affect results. However, such situations demand that the air quality model properly treat near-field source emission effects such as: building wakes, initial characterization of source release conditions and size, rates of diffusion of pollutants released as they transport downwind, and land use effects on plume transport. If chemistry is to be treated explicitly, initial source release effects are typically unimportant because the pollutants are well-mixed over some volume of the atmosphere before the chemistry of interest has greatly affected the results. To date, attempts to treat both near-field dispersion effects and chemistry have been found to be inefficient and slow on desktop computers.

Because of these competing objectives, parsimony (economy or simplicity of assumptions) is a desirable modeling trait. As illustrated in Figure 2, the likelihood of degrading performance due to input uncertainty increases as a model's formulation increases in complexity (to explicitly treat more physical processes). This decrease in performance is a function of the increase in the number of input variables and thus an increase in data uncertainty.

Because there may not be a "best" model for a given decision and different models contain different types and ranges of uncertainty, sensitivity analysis at early stages in the model development phase is useful for identifying the relative importance of model parameters. Sensitivity analysis is the computation of the effect of changes in input values or assumptions (including boundaries and model functional form) on the outputs [12].

When sensitivity analyses (Chapter 3/Appendix C) show that a model parameter has an insignificant effect on outputs and there is no process-based rationale for inclusion in the model-framework, it may be appropriate to eliminate this parameter in order to constrain model complexity. However, because sensitivity surfaces are often extremely irregular, a variable of little significance during one application of a model may be more important in a different application. In past reviews of Agency models, the SAB has supported the general guiding principle of simplifying complex models, where possible, for the sake of transparency [13] but has emphasized that care should be taken not to eliminate important parameters from process-based models simply because data are unavailable or difficult to obtain [4]. In any case, the quality and resolution of available data will ultimately constrain the type of model that can be applied. Hence, it is important to identify the existing data and and/or field collection efforts that are needed to adequately parameterize the model framework and support the application of a model.

2.3.2 Model Coding and Verification

Model coding translates the mathematical equations that constitute the model framework into functioning computer code.

Code verification ascertains that the computer code has no inherent numerical problems with obtaining a solution. Code verification tests whether the code performs according to its design specifications. It should include an examination of the numerical technique in the computer code for consistency with the conceptual model and governing equations [14]. Independent testing of the code once it is fully developed can be useful as an additional check of integrity and quality.

Steps taken early in the development of the computer code can help minimize programming errors later on and facilitate the code verification process. *Using "comment" lines to describe the purpose of each component within the code* during development makes future revisions and improvements by different modelers and programmers more efficient.

Using a flow chart during the development of the conceptual model before beginning coding helps to show the overall structure of the model program. This provides a simplified description of the calculations that will be performed in each step of the model.

Breaking the program/model into component parts or modules is also useful for careful consideration of model behavior in an encapsulated way. This allows the modeler to test the behavior of each of the sub-components separately, expediting testing and increasing confidence in the program. A module is an independent piece of software that forms part of one or more larger programs. It is useful to break large models into discrete modules because testing and locating/correcting errors ("debugging") is more difficult with large programs. The approach also makes it easier to re-use relevant modules in future modeling projects, or to update/add/remove sections of the model without altering the overall structure of the program.

Time and resources can often be saved by *using generic algorithms for common tasks*, allowing efforts to be focused on developing and improving the original aspects of a new model. An algorithm is a precise rule (or set of rules) for solving some problem. Commonly used algorithms are often published as "recipes" with publicly available code (e.g., [15]). Existing Agency models and code should also be reviewed to minimize duplication in effort. The CREM models knowledge base, which will contain a web-accessible inventory of models, will provide a resource for model developers to support these efforts.

Software engineering has evolved rapidly in the past ten years and continues to advance rapidly with changes in technology and user platforms. For example, some of the general recommendations for developing computer code given above do not apply to models that are developed using object-oriented platforms. Object-oriented platforms model systems use a collection of cooperating "objects." These objects are treated as instances of a class within a class hierarchy, where a class is a set of objects that share a common structure and behavior. The structure of a class is determined by the class variables, which represent the state of an object of that class and the behavior is given by the set of methods associated with the class [16]. When models are developed with object oriented platforms, the user should print out the actual mathematical relationships that the platform generates and this should be reviewed as part of the code validation process.

There are many available references on programming style and conventions that provide specific, technical suggestions for developing and testing computer code (e.g., *The Elements of Programming Style* [17]). In addition to these recommendations, the *Guidance for Quality Assurance Project Plans for Modeling* [7] suggests a number of practices during code verification to "check" how well it follows the "specifications" laid out during QA planning (Appendix B, Box 2: Configuration Tests Specified in the QA program).

2.4 Development of Application Tools

Assigning values to model parameters allows the model framework to be converted into an application tool. Model parameters are terms in the model that are fixed during a model run or simulation but can be changed in different runs as a method for conducting sensitivity analysis or to achieve calibration (defined below) goals. Parameters can be quantities estimated from sample data that characterize statistical populations or constants such as the speed of light and gravitational force. Other activities at this stage of model development include creating a user guide for the model, assembling datasets for model input parameters, and determining hardware requirements.

The accuracy and precision of input data used to characterize model parameters is a major source of uncertainty. Accuracy refers to the closeness of a measured or computed value to its "true" value, where the "true" value is obtained with perfect information. Due to the natural heterogeneity and random variability (stochasticity) of many environmental systems, this "true" value exists as a distribution rather than a discrete value. Variability is the term used in this guidance to describe differences that are attributable to true heterogeneity or diversity in model parameters. Because of variability, the “true” value of model parameters is often a function of the degree of spatial and temporal aggregation. Precision refers to the quality of being reproducible in outcome or performance. With models and other forms of quantitative information, precision often refers to the number of decimal places to which a number is computed. This is a measure of the “preciseness” or “exactness” of the model.

The most appropriate data, as defined by QA protocols for field sampling, data collection, and analysis [18, 19, 20] should always be selected for use in modeling analyses. Whenever possible, all parameters should be directly measured in the system of interest.

Some models are “calibrated” to set parameters. Guidance on model calibration as a QA project plan element is given in Appendix B (Box 3: Quality Assurance Planning Suggestions for Model Calibration Activities). In this guidance, calibration is defined as the process of adjusting model parameters within physically defensible ranges until the resulting predictions give the best possible fit to the observed data [21]. In some disciplines, calibration is also referred to as parameter estimation [14]. In cases where a calibration database is developed and improved over time, it may be necessary to periodically reevaluate initial adjustments and estimates. When data for quantifying one or more parameter values are limited, calibration exercises can also be used to find solutions that result in the ‘best fit’ of the model. However, this type of calibration should be undertaken with caution, as these solutions will not provide meaningful information unless they are based on *measured* physically defensible ranges.

Because of these concerns, the use of calibration to improve model performance varies among EPA offices and regions. For a particular model, the appropriateness of calibration may be a function of the modeling activities undertaken. For example, it is standard practice for the Office of Water to calibrate well-established model frameworks such as CE-QUAL-W2 (a model for predicting temperature fluctuations in rivers), to a specific system (e.g., the Snake River). This calibration generates a site-specific tool (e.g., the “Snake River Temperature” model). In contrast, the Office of Air and Radiation (OAR) more commonly uses model frameworks and models that do not need site-specific adjustments. For example, certain types of air models (e.g., gaussian plume) are parameterized for a range of meteorological conditions, and thus, do not need to be “re-calibrated” for different geographic locations (assuming that the range of conditions is appropriate for the model). This, in combination with a desire to avoid artificial improvements in model performance by adjusting model inputs outside of the ranges supported by the empirical databases, prompted OAR to issue the following statement on model calibration in their *Guideline on Air Quality Models* [22]:

> Calibration of models is not common practice and is subject to much error and misunderstanding. There have been attempts by some to compare model estimates and measurements on an event-by-event basis and then calibrate a model with results of that comparison. This approach is severely limited by uncertainties in both source and meteorological data and therefore it is difficult to precisely estimate the concentration at an exact location for a specific increment of time. Such uncertainties make calibration of models of questionable benefit. Therefore, model calibration is unacceptable.

In general, however, models benefit from thoughtful adaptation to respond adequately to the specifics of each regulatory problem to which they are applied.

2.5 Summary of Recommendations for Model Development

The following points summarize the recommendations for model development discussed in this chapter of the guidance:

- Present a clear statement and description (in words, functional expressions, diagrams, and graphs, as necessary) of each element of the conceptual model. For each element, the modeler should document the science behind the conceptual model.
- When possible, test simple competing conceptual models/hypotheses.
- Use sensitivity analysis early and often.
- Determine the optimal level of model complexity by making appropriate tradeoffs among competing objectives.
- Where possible, model parameters should be characterized using direct measurements of sample populations.
- All input data should meet data quality acceptance criteria in the QA project plan for modeling.

3.0 Model Evaluation

The natural complexity of environmental systems means that it is difficult to develop complete mathematical descriptions of relevant processes, including all of the intrinsic mechanisms that govern their behavior. Thus, policy-makers often rely on models as tools to approximate reality when implementing 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 provides a vehicle for dealing with this problem. 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.*

Different disciplines assign different meanings to the terms “model evaluation” and “model validation.” For example, Suter [23] 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.” The author 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 the model can never be completely accurate and the model can never correspond exactly to reality. Additionally, “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 [24]. Thus, some researchers assert that no model is ever truly “validated,” but can only be invalidated for a specific application [25]. Accordingly, this guidance focuses on the process and techniques that can be used for *model evaluation* rather than model validation or invalidation.

In simple terms, model evaluation provides information to assess the following factors [26]:

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 components 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.

As stated above, the goal of model evaluation is to ensure model quality. At EPA, quality is a concept defined by the Information Quality Guidelines (IQGs) [5]. The IQG applies to all information that is disseminated by EPA, including models, information from models and input data (Appendix B, Box 4: Definition of Quality). According to the IQG, quality has three major components: integrity, utility, and objectivity. Ensuring the objectivity of information from models by considering their accuracy, bias and reliability is emphasized in this guidance as part of the model evaluation process that addresses the questions listed above. While accuracy was defined in section 2.4, for the purposes of this guidance bias and reliability are defined as follows:

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 in the outputs of the model such that they are willing to use the model and accept its results [27]. Specifically, reliability is a function of the performance record of a model and its conformance to best available, practicable science.

3.1 Best Practices for Model Evaluation

This guidance provides an overview of model evaluation that can help answer the questions posed in Chapter 3.0 to determine when a model, despite its uncertainties, can be appropriately used to inform a decision. In summary, these questions are intended to 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.

The proposed "tools" or best practices emphasized in this guidance are: peer review of models, QA project planning including data quality assessment, model corroboration and sensitivity and uncertainty analysis. In this guidance, corroboration is defined as a qualitative and/or quantitative evaluation of the accuracy and predictive capabilities of a model.

As discussed in previous sections, the process of model evaluation is iterative in nature. Hence, the proposed qualitative and quantitative assessment techniques discussed below may be effectively applied throughout model development, testing and application and should not be interpreted as sequential steps for model evaluation.

The distinction between qualitative and quantitative assessments is given below:

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 theory being modeled. The subjective evaluations 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. Qualitative, corroboration activities 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. In addition, comparisons can be made for the special purpose of quantitatively describing the differences to be expected between model estimates of current conditions and comparable field observations. However, model predictions are not what are directly observed, so special care needs to be taken in any exercise that attempts to make quantitative comparisons of model predictions with field data.

Model evaluation should always be conducted using a graded approach that is adequate and appropriate to the decision at hand [6, 7]. This approach recognizes that model evaluation can be modified to the circumstances of the problem at hand and that programmatic needs vary in complexity and requirements. For example, a "screening" model 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. A screening model is a type of model designed to provide a "conservative" or risk-averse answer. The appropriate degree of model evaluation is ideally identified by both decision-makers and modeling staff at the onset of new projects (§2.1).

The substance of the following discussion on peer review of models and quality assurance protocols for input data is drawn from existing guidance. In addition, this chapter (along with Appendix C) provides new guidance on model corroboration activities and the use of sensitivity and uncertainty analysis during model evaluation.

3.1.1 Scientific Peer-Review

Peer review provides the main mechanism for independent evaluation and review of environmental models used by the Agency. Peer review addresses questions (1) and (4) posed during model evaluation (§3.0). Thus, the purpose of peer review is twofold. First, peer review evaluates whether the assumptions, methods, and conclusions derived from environmental models are based on sound scientific principles. Secondly, peer review provides a useful check on the scientific appropriateness of a model for informing a specific regulatory decision. This is particularly important for secondary applications of existing models. As part of this second objective, peer reviews may focus on whether a model meets the objectives or specifications set as part of the quality assurance plan (see [7]) at the onset of the modeling project (§2.1). Peer review is *not* a mechanism to comment on the *regulatory decisions* or policies that are informed by models [28]. Peer review charter questions and corresponding records for peer reviewers to answer those questions need to be incorporated into assessment planning as part of developing the QA Project Plan.

All models that inform *significant*[2] regulatory decisions are candidates for peer review [28, 29]. There are a number of reasons for initiating this review:

- Use of model results as a basis for major regulatory or policy/guidance decision-making
- Significant investment of Agency resources
- Inter-Agency or cross-Agency implications/applicability

[2] Under Executive Order 12866 (58 FR 51735), federal agencies must determine whether a regulatory action is "significant" and therefore, its *underlying scientific basis* is subject to review by the Office of Management and Budget (OMB) and the requirements of the Executive Order. The rule or regulation itself is *not* subject to peer review. 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 .

Existing guidance recommends that the first application of a new model should undergo scientific peer review but for subsequent applications, the program manager should consider the scientific/technical complexity and/or the novelty of the particular circumstances [29]. In the interests of conserving resources, peer-review of "similar" applications should be avoided. When a modeling project uses a well-established model framework (e.g., WASP), it is left to the discretion of project managers within the individual program offices and regions to determine when a modeling product is sufficiently similar to past applications and/or does not affect "significant" actions or policies, such that it should not be subject to peer review.

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. By their nature, reviews of secondary applications models 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. Information from peer reviews is also helpful for choosing among multiple competing models for a specific regulatory application. Finally, peer review is a useful mechanism for identifying the limitations of existing models

Aspects of a model that should be reviewed in this process to establish scientific credibility are: (a) appropriateness of input data, (b) appropriateness of boundary condition specification, (c) documentation of inputs and assumptions, (d) the applicability and appropriateness of using default values, (e) documentation and justification for adjusting model inputs to improve model performance (calibration), and (f) model application with respect to the range of its validity, (g) supporting empirical data that strengthen or contradict the "conclusions" that are based on model results [3, 29].

To be most effective and maximize its value, external peer review should begin as early in the model *development* phase as possible [20]. Because peer review involves significant time and resources, these allocations need to be components of the project planning and any related contracts. In the early stages, peer review can help to review 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 [3, 29]. Peer review at the development stage is also 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 to avoid inappropriate applications of a model for specific regulatory purposes [29].

The logistics of the peer review process are left to the discretion of the office managers responsible for use and application of the model for a given decision. Mechanisms for accomplishing external peer review include, but are not limited to, the following**:**

- Using an ad hoc panel of at least three scientists[3]
- Using an established external peer review mechanism such as the Science Advisory Board or the Science Advisory Panel
- Holding a technical workshop[4]

[3] The selection of an ad hoc panel of peer reviewers may create legal concerns under the Federal Advisory Committee Act (FACA). Compliance with this statute's requirements is best summarized in the Chapter two of the *Peer Review Handbook,* "Planning a Peer Review" [28]. Guidance may also be sought from the Cross-Cutting Issues Law Office of the Office of General Council.

Specific techniques for selecting the qualifications and number of reviewers needed for a given modeling project can be found in the guidelines for peer review [3, 28, 29, 30] and are summarized in Appendix C of this guidance.

3.1.2 Quality Assurance Project Planning and Data Quality Assessment

Another aspect of model evaluation that addresses the issue of whether a model has been developed according to the principles of sound science is data quality. Some variability in data is unavoidable (see § 3.1.3.1), but adhering to the tenets of data quality assessment described in other Agency guidance[5] (Appendix C, Box 5: Quality Assurance Planning and Data Acceptance Criteria) helps to minimize data uncertainty.

Well-executed QA project planning also helps to ensure that a model performs the specified task, which was the fourth question related to model evaluation posed in Section 3.0. 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* [7] provides general information about how to document quality assurance planning for modeling (e.g., specifications 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.

The quality and quantity (representativeness) of supporting data used to parameterize and (when available) corroborate models should be evaluated 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 two, §3.0). In addition, model outputs cannot be meaningfully compared to observational data (question three, §3.0) unless these data are representative of the true system being modeled.

3.1.3 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 most cases, it will not simply be a matter of comparing model results and empirical data. As noted in Section 2.1, in developing and using an environmental model it is important that modelers and decision-makers consider the acceptable degree of uncertainty within the context of a specific model application. An understanding of the uncertainties underlying a model is needed to meaningfully address this question. Where

[4] Use of a 'one-shot' technical workshop does not implicate the same concerns under FACA, especially if EPA personnel see only the individual opinions of the workshop attendees. However, repeated meetings of the workshop group or an attempt to forge a group consensus at the end of a one-day meeting might implicate FACA requirements.

[5] Other guidance that can help to 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 [8].
- *Guidance on Choosing a Sampling Design for Environmental Data Collection*, on applying statistical sampling designs to environmental applications [18].
- *Guidance for Data Quality Assessment: Practical Methods for Data Analysis*, to evaluate the extent to which data can be used for a specific purpose [20].

practical, the recommended analyses should be conducted and their results reported in the documentation supporting the model.

3.1.3.1 Types of Uncertainty

Uncertainties in the scientific sense are a component of all aspects of the modeling process. However, identifying the types of uncertainty that significantly influence model outcomes (qualitatively or quantitatively) and communicating their importance is key to successfully integrating information from models into the decision-making process. As defined in Section 2.0, 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: (a) uncertainty in the underlying science and algorithms of a model (model framework uncertainty), (b) data uncertainty, and (c) uncertainty regarding the appropriate application of a model (application niche uncertainty). 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 necessary to translate the conceptual model into mathematical terms as described in Section 2.3. In the scientific literature this type of uncertainty is also referred to as structural error [31], conceptual errors [32], uncertainties in the conceptual model [33], or model error/uncertainty [34, 35]. 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. Model error and model uncertainty are both generally synonymous with model framework uncertainty.

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 that is caused by measurement errors, analytical imprecision and limited sample sizes during the collection and treatment of data.

In contrast to data uncertainty, variability results from the inherent randomness of certain parameters that is attributable to the heterogeneity and diversity in environmental processes. Variability includes: fluctuations in ecological conditions, differences in habitat, and genetic variances among populations [34]. 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 [34]. Accordingly, variability is referred to as irreducible because it can be better characterized and represented but cannot be reduced with further study [34].

A model's application niche is the set of conditions under which the use of a model is scientifically defensible [30]. 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 choosing among existing models for an application outside of the system for which it was originally developed and/or developing a larger model from several existing models with different spatial or temporal scales [36].

A good example of application niche uncertainty is given in the SAB review of MMSOILS (Multimedia Contaminant Fate, Transport and Exposure Model) where they address 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) [37]. In this example, the SAB considered the MMSOILS model acceptable within its original application niche, but unsuitable for more heterogeneous conditions.

3.1.3.2 Model Corroboration

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

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 data sets independent of the calibration data set, and quantitative testing criteria. In many cases, collecting independent data sets for formal model corroboration is extremely costly or otherwise unfeasible. In such circumstances, model evaluation may be appropriately conducted using a combination of other tools for model evaluation that are discussed in this section.

The degree of similarity between calibration data and corroboration data provides a measure of robustness of model performance [38, 39]. Robustness is defined in this guidance as the capacity of a model to perform equally well across the full range of environmental conditions for which it was designed. The degree of similarity among data sets available for calibration and corroboration provides insight into the robustness of the model. For example, if the dataset used to calibrate a model is identical or statistically similar to the dataset used to corroborate a model, an independent measure of the model's performance has not been provided. In this case, the exercise has provided no insight into model robustness. Conversely, when model outputs are similar to corroboration data that are significantly different from the calibration data, the corroboration exercise provides a measure of both model performance and robustness.

Quantitative model corroboration methods are also 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., [40]). The Office of Air and Radiation evaluates models in this manner and when a single model is found to perform better than others in a given category, it is recommended in the *Guidelines on Air Quality Models* for application in that category as a preferred model [22]. If no model is found to perform better through the evaluation exercise, then the preferred model is selected on the basis of other factors such as past use, public familiarity, cost or resource requirements, and availability.

3.1.3.3 Sensitivity and Uncertainty Analysis

Sensitivity analysis is the study of how the response of a model can be apportioned to changes in a model's inputs [41]. A model's sensitivity describes the degree to which the model result is affected by changes in a selected input parameter [14]. 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 is sometimes reducible through further study and with the collection of additional data. Existing Agency guidance (e.g., [34]) distinguishes uncertainty analysis from methods that are used to account for variability in input data and model parameters. Variability in model parameters and input data can be better characterized through further study but is usually not reducible [34].

Although uncertainty and sensitivity analysis are closely related, uncertainty is parameter specific, and sensitivity is algorithm-specific with respect to model "variables." By investigating the "relative sensitivity" of model parameters, a user can become knowledgeable of the relative importance of parameters in the model. By knowing the "uncertainty" associated with parameter values and the "sensitivity" of the model to specific parameters, a user will be more informed regarding the confidence that can be placed in model results. Recommended techniques for conducting uncertainty and sensitivity analysis are discussed in Appendix C.

3.2 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, the Agency has a preference for using non-proprietary models when available. The Agency acknowledges there will be times when the use of proprietary models provides the most reliable and best accepted modeling alternatives.

When proprietary models are used, their use should be accompanied by comprehensive, publicly available documentation that the proprietary models adhere to the recommendations within this guidance. That is, the propriety models should be accompanied by documentation that describes:

- The conceptual model and theoretical basis (as described in §2.2) 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 §2.3.2);
- The process used to evaluate the model (as described in §3.1) 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 §3.0); and
- To the extent practicable, access to input and output data such that third parties can replicate results derived from the use of the proprietary model.

3.3 Summary of Model Evaluation

To summarize, model evaluation provides information to help answer the four questions in Section 3.0. Recommended components of the evaluation process that help to determine when a model, despite its *uncertainties*, can be appropriately used to inform a decision that were reviewed in this section 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.

As discussed in Section 3.0, quality is an attribute of models that is meaningful only within the context of a specific model application. Deciding whether a model serves its intended purpose is answered through the discretion of the decision maker. 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. The following section recommends best practices for integrating models into environmental decisions.

4.0 Model Application

This chapter presents best practices and other recommendations for integrating the results of environmental models into Agency decisions. Environmental models should provide decision makers with meaningful outputs and enable them to understand the modeling processes that generated these outputs. Decision makers need to understand the relevant environmental processes at a level that is appropriate for the decision of interest. In other words, decision makers should be empowered by being shown the inside of the "black box," as well as its outputs.

This need was emphasized in a National Research Council (NRC) assessment of the TMDL program. The NRC suggested that strengthening the program's scientific basis would not just be a matter of focusing on analytical results. More importantly, the program needs to ensure that it conforms to the *scientific method* [42]. The following sections describe how the transparency of a modeling process can be achieved and documented (§4.1) and how reasoned, evidence-based decision making (§4.2) can be implemented.

4.1 Transparency

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

4.1.1 Documentation

Documentation enables decision makers and other users of models to understand the process by which a model was developed, its intended application niche, and the limitations of its applicable domain. One of the major objectives of documentation should be the reduction of application niche uncertainty.

Modelers should document the following elements of technical analysis that are relevant to each modeling project (see following page). These elements are adapted from EPA Region 10's standard practices for modeling projects.

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 [7] *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.*

4.1.2 Effective Communication of Uncertainty

Reliance on model results is conditioned by one's ability to understand the uncertainty of those results. Consequently, the modeling process should include effective communication of uncertainty such that information about uncertainty is provided to anyone who will use model results. At a minimum, the modeler should communicate uncertainty to the decision maker and any future users. All technical information should documented in a manner that decision makers can readily interpret and understand Recommendations for improving clarity that have been adapted from the Risk Characterization Handbook [43] include the following:

- Be brief as possible while still providing all necessary details.
- Use plain language that is understood by modelers, policy makers, and the informed lay person.
- 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 [12] for suggestions).

It is important to clearly identify the conclusions of the modeling project and other relevant points of the modeling project when communicating with decision makers. The challenge is to characterize these essentials for decision makers, while 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 that they can apply model results appropriately. This is consistent with QA planning practices that assert that the quality of data and any limitations on their use should be discussed with respect to their intended use in all technical reports [44].

4.2 Reasoned, Evidence-Based Decision-Making

4.2.1 Model Evolution and the State of Scientific Knowledge

Due to time constraints, scarcity of resources, and/or lack of scientific understanding, technical decisions are often based on incomplete information and imperfect models. Furthermore, even if model developers strive to use the best science available, advances in knowledge and understanding are ongoing. Given this reality, decision makers should use model results in the context of an iterative, ever-improving process of continuous model refinement.

Models should include a clear explanation of their 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 to establish the 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 [45].

4.2.2 Explaining Deviations Between Models and Reality

Models represent reality and necessarily depend on applicable conditions. Hence, applying a model to conditions that are different from those of its design deserves scrutiny. In general, it may be desirable to apply a model to industries, activities, or locations that vary from the conditions used to corroborate the model. Any variance in use should be evaluated properly.

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 3.1.3 regarding corroboration, deviations from empirical observations are to be expected.

If the challengers make a strong case against EPA's decisions, then the courts generally look to the record supporting EPA's decisions. The record will be examined for justification as to why the model was reasonable [46]. The challenger's case will be strong if model results deviate significantly from the scenario to which it was applied such that the model does not rationally relate to the situation or facts. This outcome can be avoided by providing good documentation of the rationale for a model's use and the appropriate context for its application.

4.3 Appropriate Implementation of Guidance

Program and regional offices may rely on the details of this guidance, while modifying and clarifying recommendations, as appropriate and necessary. Each EPA office should be responsible for implementing this model guidance in an appropriate manner to meet its needs.

Appendix A: Glossary of Frequently Used Terms

Accuracy: Closeness of a measured or computed value to its "true" value, where the "true" value is obtained with perfect information. Due to the natural heterogeneity and stochasticity of many environmental systems, this "true" value exists as a distribution rather than a discrete value. In these cases, the "true" value will be a function of spatial and temporal aggregation.

Algorithm: A precise rule (or set of rules) for solving some problem.

Analytical Models: Models that can be solved mathematically in closed form. For example, some model algorithms that are based on relatively simple differential equations can be solved analytically to provide a single solution.

Applicability and Utility: One of EPA's five Assessment Factors (see definition) that describes the extent to which the information is relevant for the Agency's intended use [47].

Application Niche: The set of conditions under which the use of a model is scientifically defensible. The identification of application niche is a key step during model development. Peer review should include an evaluation of application niche. An explicit statement of application niche helps decision makers to understand the limitations of the scientific basis of the model [29].

Application Niche Uncertainty: Uncertainty as to the appropriateness of a model for use under a specific set of conditions (see application niche).

Assessment Factors: Considerations recommended by EPA for evaluating the quality and relevance of scientific and technical information. These include: (1) soundness, (2) applicability and utility, (3) clarity and completeness, (4) uncertainty and variability, (5) evaluation and review [47].

Bias: *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.

Boundaries: The spatial and temporal conditions and practical constraints under which environmental data are collected. Boundaries specify the area or volume (spatial boundary) and the time period (temporal boundary) to which a decision will apply [8].

Boundary Conditions: Sets of values for state variables and their rates along problem domain boundaries, sufficient to determine the state of the system within the problem domain.

Calibration: The process of adjusting model parameters within physically defensible ranges until the resulting predictions give the best possible fit to the observed data [21]. In some disciplines, calibration is also referred to as "parameter estimation" [14].

Checks: Specific tests in a quality assurance plan that are used to evaluate whether the specifications (performance criteria) for the project developed at its onset have been met.

Clarity and Completeness: One of EPA's five Assessment Factors (see definition) that describes the degree of clarity and completeness with which the data, assumptions, methods, quality assurance, sponsoring organizations and analyses employed to generate the information are documented [47].

Class (see object oriented platform): A set of objects that share a common structure and behavior. The structure of a class is determined by the class variables, which represent the state of an object of that class and the behavior is given by the set of methods associated with the class [16].

Code: Instructions, written in the syntax of a computer language, which provide the computer with a logical process. Code may also be referred to as computer program. The term code describes the fact that computer languages use a different vocabulary and syntax than algorithms that may be written in standard language.

Complexity: The opposite of simplicity. Complex systems tend to have a large number of variables, multiple parts, mathematical equations of a higher order, and are more difficult to solve. In relation to computer models, complexity generally refers to the level in difficulty in solving mathematically posed problems as measured by the time, number of steps or arithmetic operations, or memory space required (called time complexity, computational complexity, and space complexity, respectively).

Conceptual Basis: This is the underlying scientific foundation of model algorithms or governing equations. The conceptual basis for models is either empirical (based on statistical relationships between observations) or mechanistic (process-based). See definitions for: empirical model and mechanistic model.

Conceptual Model: A hypothesis regarding the important factors that govern the behavior of an object or process of interest. This can be an interpretation or working description of the characteristics and dynamics of a physical system [21].

Confounding Errors: Errors induced by unrecognized effects from variables that are not included in the model. The unrecognized, uncharacterized nature of these errors makes them more difficult to describe and account for in statistical analysis of uncertainty [48].

Constants: Quantities with have fixed values (e.g., the speed of light and the gravitational force) representing known physical, biological, or ecological activities.

Corroboration (model): Quantitative and qualitative methods for evaluating the degree to which a model corresponds to reality. 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.

Data Uncertainty: Uncertainty (see definition) that is caused by measurement errors, analytical imprecision and limited sample sizes during the collection and treatment of data. Data uncertainty, in contrast to variability (see definition) is the component of total uncertainty that is "reducible" through further study.

Debug: The identification and removal of bugs from computer code. Bugs are errors in computer code that range from typos to misuse of concepts and equations.

Deterministic Model: A model that provides a single solution for the state variables. Because this type of model does not explicitly simulate the effects of data uncertainty or variability, changes in model outputs are solely due to changes in model components.

Domain (spatial and temporal): The limits of space and time that are specified within a model's boundary conditions (see boundary conditions).

Domain Boundaries (spatial and temporal): The spatial and temporal domain of a model are the limits of extent and resolution with respect to time and space for which the model has been developed and over which it should be evaluated.

Empirical Model: An empirical model is one where the structure is determined by the observed relationship among experimental data [23]. These models can be used to develop relationships that are useful for forecasting and describing trends in behavior but they are not necessarily mechanistically relevant.

Environmental Data: Information collected directly from measurements, produced from models, and compiled from other sources such as databases and literature [5].

Evaluation (model): The process used to generate information to determine whether a model and its results are of a quality sufficient to serve as the basis for a regulatory decision.

Evaluation and Review: One of EPA's five Assessment Factors (see definition) that describes the extent of independent verification, validation and peer review of the information or of the procedures, measures, methods or models [47].

Extrapolation: Extrapolation is a process that uses assumptions about fundamental causes underlying the observed phenomena in order to project beyond the range of the data. In general, extrapolation is not considered a reliable process for prediction; however, there are situations where it may be necessary and useful.

Expert Elicitation: a process for obtaining expert beliefs about subjective quantities and probabilities. Typically, structured interviews and/or questionnaires are used to elicit the necessary knowledge. Expert elicitations may also include "coaching" techniques to help the expert conceptualize, visualize, and quantify the knowledge being sought.

False Positives: Also known as false rejection decision errors. False positives occur when the null-hypothesis or baseline condition is incorrectly rejected based on the sample data. The decision is made assuming the alternate condition or hypothesis to be true when in reality it is false [8].

False Negatives: Also known as false acceptance decision errors. False negatives occur when the null hypothesis or baseline condition cannot be rejected based on the available sample data. The decision is made assuming the baseline condition is true when in reality it is false [8].

Forcing/Driving Variables: External or exogenous (outside the model framework) factors that influence the state variables calculated within the model. These may include, for example, climatic or environmental conditions (temperature, wind flow, oceanic circulation, etc.).

Function: A mathematical relationship between variables.

Forms (models): Models can be represented and solved in different forms, including: analytic, stochastic, and simulation.

Graded approach: process of basing the level of application of managerial controls applied to an item or work according to the intended use of results and degree of confidence needed in the results [7].

Integrity: One of three main components of quality in EPA's *Information Quality Guidelines*. Integrity refers to the protection of information from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification [5].

Intrinsic Variation: The variability (see definition) or inherent randomness in the real-world processes.

Loading: The rate of release of a constituent of interest to a particular receiving medium.

Measurement Errors: Errors in the observed data that are a function of human or instrumental error during collection. Such errors may be independent or random. When a persistent bias or miscalibration is present in the measurement device, measurement errors may be correlated among observations [48]. In some disciplines, measurement error may be referred to as observation error.

Mechanistic Model: A model that has a structure that explicitly represents an understanding of physical, chemical, and/or biological processes. Mechanistic models quantitatively describe the relationship between some phenomenon and underlying first principles of cause. Hence, in theory, they are useful for inferring solutions outside of the domain that the initial data was collected and used to parameterize the mechanisms.

Model Coding: The process of translating the mathematical equations that constitute the model framework into a functioning computer program.

Model Framework: The model framework is the system of governing equations that make up the mathematical model. It is a formal mathematical specification of the concepts and procedures of the conceptual model consisting of generalized algorithms (computer code/software) for different site or problem-specific simulations [21].

Model Framework Uncertainty: The uncertainty in the underlying science and algorithms of a model. 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 necessary to translate the conceptual model into mathematical terms.

Model: A representation of the behavior of an object or process, often in mathematical or statistical terms. Models can also be physical or conceptual [2].

Model Pedigree: A qualitative or quantitative determination of the rigor with which a model has been developed and evaluated. In some cases, a model's pedigree may be represented as a quantitative score that reflects the quality of a model's development and evaluation. Model pedigree is concerned with the source of data used in model development, the origin of the model framework, and the extent of evaluation performed on the model.

Modes (of models): Manner in which a model operates. Models can be designed to represent phenomena in different modes. Prognostic (or predictive) models are designed to forecast outcomes and future events, while diagnostic models work "backwards" to assess causes and precursor conditions.

Module: An independent or self contained component of a model which is used in combination with other components and forms part of one or more larger programs.

Noise: Inherent variability that the model does not characterize (see definition for variability).

Objectivity: One of three main components of quality in EPA's *Information Quality Guidelines*. Objectivity includes whether disseminated information is being presented in an accurate, clear, complete and unbiased manner. In addition, objectivity involves a focus on ascertaining accurate, reliable and unbiased information [5].

Object-Oriented Platforms: Type of user interface that models systems using a collection of cooperating "objects." These objects are treated as instances of a class within a class hierarchy, where a class is a set of objects that share a common structure and behavior. The structure of a class is determined by the class variables, which represent the state of an object of that class and the behavior is given by the set of methods associated with the class [16].

Parameters: Terms in the model that are fixed during a model run or simulation but can be changed in different runs as a method for conducting sensitivity analysis or to achieve calibration goals.

Parametric Variation: When the value of a parameter itself is not a constant and includes natural variability. Consequently, the parameter should be described as a distribution [49].

Parameter Uncertainty: See Data Uncertainty.

Perfect Information: The state of information where there is no uncertainty. The current and future values for all parameters are known with certainty. The state of perfect information includes knowledge about the values of parameters with natural variability.

Precision: The quality of being reproducible in amount or performance. With models and other forms of quantitative information, precision refers specifically to the number of decimal places to which a number is computed as a measure of the "preciseness" or "exactness" with which a number is computed.

Probability Density Function: Mathematical, graphical, or tabular expression of the relative likelihoods with which an unknown or variable quantity may take various values. The sum (or integral) of all likelihoods equals one for discrete (continous) random variables [50]. These distributions arise from the fundamental properties of the quantities we are attempting to represent. For example, quantities formed from adding many uncertain parameters tend to be normally distributed, and quantities formed from multiplying uncertain quantities tend to be lognormal [12].

Programs (computer): Instructions, written in the syntax of a computer language, that provide the computer with a step-by-step logical process. Computer programs are also referred to as code.

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 theory being modeled. The subjective evaluations 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, corroboration activities may involve the elicitation of expert judgment on the true behavior of the system and agreement with model-forecasted behavior.

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. In addition, comparisons can be made for the special purpose of quantitatively describing the differences to be expected between model estimates of current conditions and comparable field observations.

Quality: A broad term that includes notions of integrity, utility, and objectivity [5].

Reducible Uncertainty: Uncertainty in models that can be minimized or even eliminated with further study and additional data [34]. See data uncertainty.

Reliability: The confidence that (potential) users have in a model and in the information derived from the model such that they are willing to use the model and the derived information [27]. Specifically, reliability is a function of the performance record of a model and its conformance to best available, practicable science.

Robustness: The capacity of a model to perform equally well across the full range of environmental conditions for which it was designed.

Screening Model: A type of model designed to provide a "conservative" or risk-averse answer. Because screening models can be used with limited information and are conservative, they can be used in lieu of refined models and in some cases, even when time or resources are not limited.

Sensitivity: The degree to which the model outputs are affected by changes in a selected input parameters [14].

Sensitivity Analysis: The computation of the effect of changes in input values or assumptions (including boundaries and model functional form) on the outputs [12]. The study of how uncertainty in a model output can be systematically apportioned to different sources of uncertainty in the model input [41]. By investigating the "relative sensitivity" of model parameters, a user can become knowledgeable of the relative importance of parameters in the model.

Sensitivity Surface: A theoretical multi-dimensional "surface" that describes the response of a model to changes in its parameter values. A sensitivity surface is also known as a response surface.

Simulation Models: Simulation models are used to obtain solutions for more models that are too complex to be solved analytically. In general, simulation models provide approximations of the mathematical solutions. For most situations, where a differential equation is being approximated, the simulation model will use finite time step (or spatial step) to "simulate" changes in state variables over time (or space).

Soundness: One of EPA's five Assessment Factors (see definition) that describes the extent to which the scientific and technical procedures, measures, methods or models employed to generate the information are reasonable for and consistent with, the intended application [47].

Specifications: Acceptance criteria set at the onset of a quality assurance plan that help to determine if the intended objectives of the project have been met. Specifications are evaluated using a series of associated checks (see definition).

State variables: The dependent variables calculated within the model, which are also often the performance indicators of the models that change over the simulation.

Statistical Models: Simple linear or multivariate regression models obtained by fitting observational data to a mathematical function.

Stochasticity: Fluctuations in ecological processes that are due to natural variability and inherent randomness.

Stochastic Model: A model that includes variability (see definition) in model parameters. This variability is a function of: 1) changing environmental conditions, 2) spatial and temporal aggregation within the model framework, 3) random variability. The solutions obtained by the model or output is therefore a function of model components and random variability.

Transparency: The clarity and completeness with which data, assumptions and methods of analysis are documented. Experimental replication is possible when information about modeling processes is properly and adequately communicated [5].

Uncertainty: The term used in this guidance to describe *lack of knowledge* about models, parameters, constants, data, and beliefs. There are many sources of uncertainty, including: the science underlying a model, uncertainty in model parameters and input data, observation error, and code uncertainty. Additional study and collecting more information allows error that stems from uncertainty to be minimized/reduced (or eliminated). In contrast, variability (see definition) is irreducible but can be better characterized or represented with further study [7, 49].

Uncertainty Analysis: Investigates the effects of lack of knowledge or potential errors on the model (e.g, the "uncertainty" associated with parameter values) and when conducted in combination with sensitivity analysis (see definition) allows a model user to be more informed about the confidence that can be placed in model results.

Uncertainty and Variability: One of EPA's five Assessment Factors (see definition) that describes the extent to which the variability and uncertainty (quantitative and qualitative) in the information or in the procedures, measures, methods or models are evaluated and characterized [47].

Utility: One of three main components of quality in EPA's Information Quality Guidelines. Utility refers to the usefulness of the information to the intended users [5].

Variable: A measured or estimated quantity which describes an object or can be observed in a system and which is subject to change.

Variability: Variability refers to observed differences attributable to *true heterogeneity* or diversity. Variability is the result of natural random processes and is usually not reducible by further measurement or study (although it can be better characterized) [34].

Verification (code): Examination of the algorithms and numerical technique in the computer code to ascertain that they truly represent the conceptual model and that there are no inherent numerical problems with obtaining a solution [14].

Appendix B: Supplementary Material on Quality Assurance Planning and Protocols

This section consists of a series of text boxes meant to supplement concepts and references made in the main body of the document. They are not meant to provide a comprehensive discussion on QA practices, and each box should be considered as a discrete unit. Individually, each of the text boxes provides additional background material for specific sections of the main document. The complete QA manuals for each subject area discussed in this guidance and referred to below should be consulted for more complete information on QA planning and protocols.

Box 1: Background on EPA Quality System

The EPA Quality System defined in EPA Order 5360.1 A2, *Policy and Program Requirements for the Mandatory Agency-wide Quality System* [44], covers environmental data produced from models as well as "any measurement or information that environmental processes, location, or conditions; ecological or health effects and consequences; or the performance of environmental technology." For EPA, environmental data includes information collected directly from measurements, produced from models, and compiled from other sources such as databases and literature.

The EPA Quality System is based on an American National Standard, ANSI/ASQC E4-1994 [51]. Consistent with minimum specifications of this standard, §6.a.(7) of EPA Order 5360.1 A2 states that EPA organizations will develop a Quality System that includes "approved Quality Assurance (QA) Project Plans, or equivalent documents defined by the Quality Management Plan, for all applicable projects and tasks involving environmental data with review and approval having been made by the EPA QA Manager (or authorized representative defined in the Quality Management Plan). The approval of the QA Project Plan containing the specifications for the product(s) and the checks against those specifications (assessments) for implementation is an important management control assuring records to avoid fiduciary "waste and abuse" (Federal Managers' Financial Integrity Act of 1982 [52] with annual declarations including conformance to the EPA Quality System). The assessments (including peer review) support the product acceptance for models and their outputs and approval for use such as supporting environmental management decisions by answering questions, characterizing environmental processes or conditions and direct decision support such as economic analyses (process planned in Group D in the Guidance for QA Project Plans for Modeling). EPA's policies for QA Project Plans are provided in Chapter 5 of the *EPA Manual 5360 A1* [44], *EPA Quality Manual for Environmental Programs* [53] for in-house modeling and *Requirements for Quality Assurance Project Plans (QA/R-5)* [7] for modeling done through extramural agreements (e.g., contracts 48 CFR 46, grants and cooperative agreements 40 CFR 30, 31, and 35). For Interagency Agreements QA requirements need to be negotiated and written into the agreement if the project is funded by EPA but if funds are received by EPA the *EPA Manual 5360 A1* [44] applies.

EPA Order 5360.1 A2 also states that EPA organization's Quality Systems include "use of a systematic planning approach to develop acceptance or performance criteria for all work covered" and "assessment of existing data, when used to support Agency decisions or other secondary purposes, to verify that they are of sufficient quantity and adequate quality for their intended use."

Box 2: Configuration Tests Specified in the QA program

During code verification the final set of computer code is scrutinized to assure that the equations are programmed correctly and that sources of error, such as rounding, are minimal. This process is likely to be more extensive for new computer code. For existing code, the criteria used for previous verification, if known, can be described or cited. Any additional criteria specific to the modeling project can be specified, along with how the criteria were established. Possible departures from the criteria are discussed, along with how the departures can affect the modeling process.

Software code development inspections: Software requirements, software design, or code are examined by an independent person or groups other than the author(s) to detect faults, programming errors, violations of development standards, or other problems. All errors found are recorded at the time of inspection, with later verification that all errors found have been successfully corrected.

Software code performance testing: Software used to compute model predictions is tested to assess its performance relative to specific response times, computer processing usage, run time, convergence to solutions, stability of the solution algorithms, the absence of terminal failures, and other quantitative aspects of computer operation.

Tests for individual model module: Checks ensure that the computer code for each module is computing module outputs accurately and within any specific time constraints. (Modules are different segments or portions of the model linked together to obtain the final model prediction.)

Model framework testing: The full model framework is tested as the ultimate level of integration testing to verify that all project-specific requirements have been implemented as intended.

Integration tests: The computational and transfer interfaces between modules need to allow an accurate transfer of information from one module to the next, and ensure that uncertainties in one module are not lost or changed when that information is transferred to the next module. These tests detect unanticipated interactions between modules and track down cause(s) of those interactions. (Integration tests should be designed and applied in a hierarchical way by increasing, as testing proceeds, the number of modules tested and the subsystem complexity.)

Regression tests: All testing performed on the original version of the module or linked modules is repeated to detect new “bugs” introduced by changes made in the code to correct a model.

Stress testing (of complex models): This ensures that the maximum load (e.g., real-time data acquisition and control systems) does not exceed limits. The stress test should attempt to simulate the maximum input, output, and computational load expected during peak usage. The load can be defined quantitatively using criteria such as the frequency of inputs and outputs or the number of computations or disk accesses per unit of time.

Acceptance testing: Certain contractually required testing may be needed before the new model or the client accepts model application. Specific procedures and the criteria for passing the acceptance test are listed before the testing is conducted. A stress test and a thorough evaluation of the user interface is a recommended part of the acceptance test.

Beta testing of the pre-release hardware/software: Persons outside the project group use the software as they would in normal operation and record any anomalies encountered or answer questions provided in a testing protocol by the regulatory program. The users report these observations to the regulatory program or specified developers, who address the problems before release of the final version.

Reasonableness checks: These checks involve items like order-of-magnitude, unit, and other checks to ensure that the numbers are in the range of what is expected.

Note: This section is adapted from [7].

Box 3: Quality Assurance Planning Suggestions for Model Calibration Activities

Information related to objectives and acceptance criteria for calibration activities that generally appear at the beginning of this QA Project Plan element includes the following:

Objectives of model calibration: This includes expected accomplishments of the calibration and how the predictive quality of the model might be improved as a result of implementing the calibration procedures.

Acceptance criteria: The specific limits, standards, goodness-of-fit, or other criteria on which a model will be judged as being properly calibrated (e.g., the percentage difference between reference data values from the field or laboratory and predicted results from the model). This includes a mention of the types of data and other information that will be necessary to acquire in order to determine that the model is properly calibrated (e.g., field data, laboratory data, predictions from other accepted models). In addition to addressing these questions when establishing acceptance criteria, the QA Project Plan can document the likely consequences (e.g., incorrect decision-making) for selecting data that do not satisfy one or more of these areas (e.g., are non-representative, are inaccurate), as well as procedures in place to minimize the likelihood of selecting such data.

Justifying the calibration approach and acceptance criteria: Each time a model is calibrated, it is potentially altered. Therefore, it is important that the different calibrations, the approaches taken (e.g., qualitative versus quantitative), and their acceptance criteria are properly justified. This justification can refer to the overall quality of the standards being used as a reference or of the quality of the input data (e.g., whether data are sufficient for statistical tests to achieve desired levels of accuracy).

Box 4: Definition of Quality

As defined by EPA's Information Quality Guidelines [5], quality is a broad-term that includes notions of integrity, utility, and objectivity. Integrity refers to the protection of information from unauthorized access or revision to ensure that it is not compromised through corruption or falsification. In the context of environmental models, often integrity is most relevant to protection of code from unauthorized or inappropriate manipulation (see Box 2). Utility refers to the usefulness of the information to the intended users. The utility of modeling projects is aided by the implementation of a systematic planning approach that includes the development of acceptance or performance criteria (see Box 1). Objectivity involves two distinct elements, presentation and substance. Objectivity includes whether disseminated information is being presented in an accurate, clear, complete and unbiased manner. In addition, objectivity involves a focus on ascertaining accurate, reliable and unbiased information.

EPA's five general assessment factors [47] for evaluating the quality and relevance of scientific and technical information supporting Agency actions are: (a) soundness, (b) applicability and utility, (c) clarity and completeness, (d) uncertainty and variability, (e) evaluation and review. Soundness refers to the extent to which a model is appropriate for its intended application and is a reasonable representation of reality. Applicability and utility describe the extent to which the information is relevant and appropriate for the Agency's intended use. Clarity and completeness refer to documentation of the data, assumptions, methods, quality controls, and analysis employed to generate the model outputs. Uncertainty and variability highlight the extent to which limitations in knowledge and information and natural randomness in input data and models are evaluated and characterized. Evaluation and review evaluate the extent of independent application, replication, evaluation, validation and peer review of the information or of the procedures, measures, methods or models employed to generate the information.

Appendix C: Best Practices for Model Evaluation

C.1 Introduction

This appendix presents a practical guide to the best practices for model evaluation (please see §3.1 for descriptions of these practices). These best practices are:

- Scientific peer review (§3.1.1)
- Quality assurance project planning (§3.1.2)
- Corroboration (§3.1.3)
- Sensitivity Analysis (§3.1.3)
- Uncertainty Analysis (§3.1.3)

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 may 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 C.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.

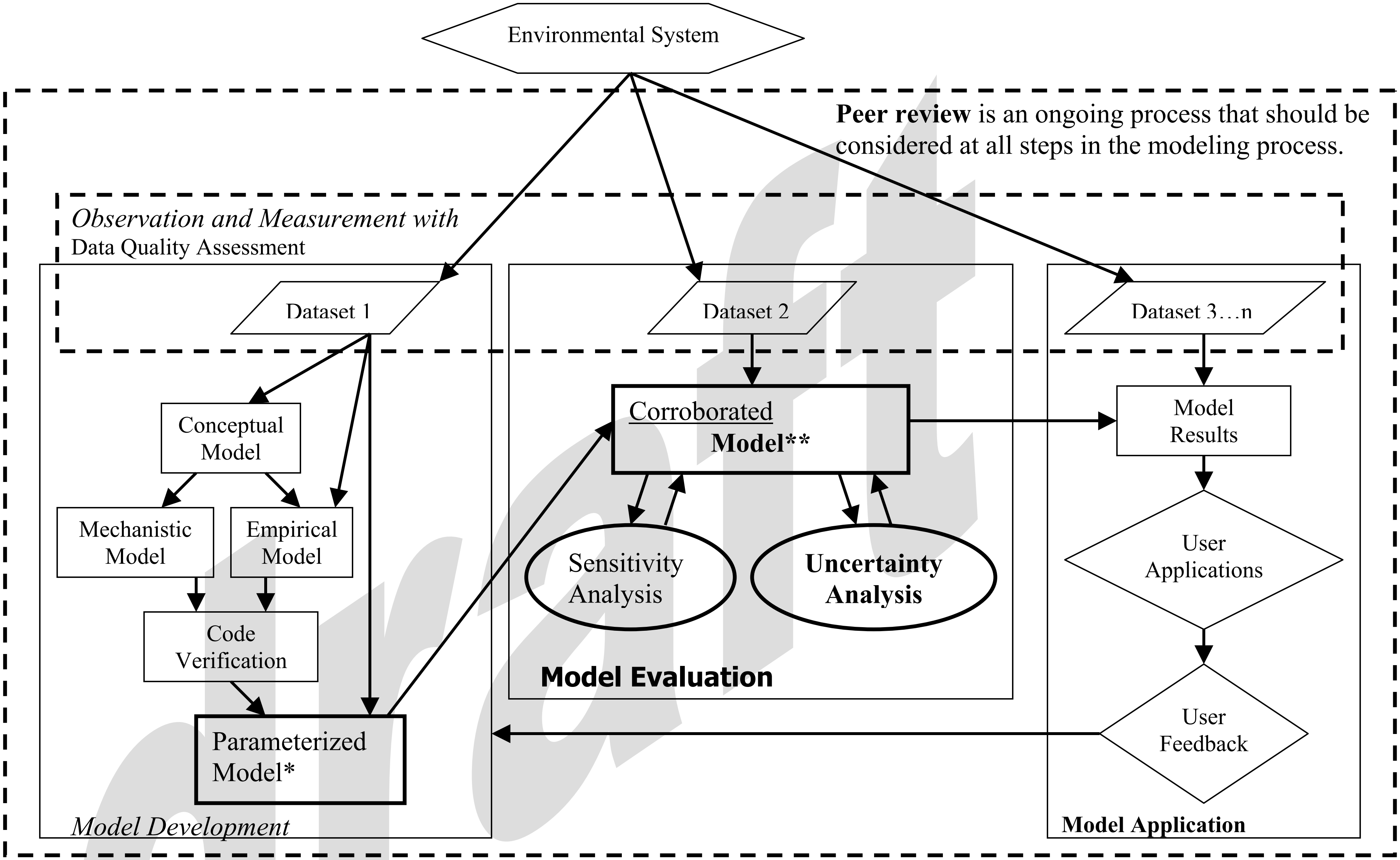


Figure C.1.1. The Modeling Process. * In some disciplines parameterization may include, or be referred to as, calibration.
** Qualitative and/or quantitative corroboration should be performed when necessary.

C.2 Scientific Peer Review

EPA policy states that major scientifically and technically based products related to Agency decisions normally should 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 to provide a resource for program managers responsible for implementing the peer review process for models [14]. More specifically, this guidance discussed 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 [14, 15]:

- 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 would be 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 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: External peer review is considered by the originating office. 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 ad hoc peer review committee or process).

C.3 Quality Assurance Project Planning

Box 5: Quality Assurance Planning and Data Acceptance Criteria

The QA Project Plan needs to address the following four issues regarding information on how non-direct measurements are acquired and used on the project [19]:

- 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 data set 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.

C.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" [10]. 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 consist 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 3.1.3.1, 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, 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.

C.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 does not treat model corroboration in a formalized statistical manner. Expert knowledge can establish model reliability through: (a) *consensus* and (b) *consistency*. For example, an expert panel consisting of model developers, other parties 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.

C.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 the use of 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 (§3.2.3); 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 [54].

C.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, precision 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.

C.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 unitless 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: $MeanAbsError = \frac{\Sigma |e|}{m}$

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

C.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, *Ecological Detective* [10]

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

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

Sum of squares 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 data set. However, if several data sets 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 §C.4.4.4).

C.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 D6589 – 00 for Statistical Evaluation of Atmospheric Dispersion Model Performance [54]. The following discussion summarizes some of the issues discussed in D6589.

C.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.

C.4.4.2 Define Evaluation Procedures

Having selected evaluation objectives for comparison, the next step would be to define a evaluation procedure (or series of procedures), which define how each evaluation objective will be derived from the available information. Development of statistical model evaluation procedures begins by providing 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 the ability of models to replicate the average centerline concentration as a function of transport downwind and as a function of atmospheric stability. Several items are defined to achieve the stated goal, namely: 1) what is an 'average centerline concentration', 2) what is 'transport downwind', and 3) how will 'stability' be defined?

What questions arise in defining the average centerline concentration? Given a sampling arc of concentration values, a decision is needed of whether the centerline concentration is the maximum value seen anywhere along the arc, or whether the centerline concentration is that seen near the center of mass of the observed lateral concentration distribution. If one chooses the latter concept, then a definition is needed 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 '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 the 'stability'? For surface releases, surface-layer Monin-Obukhov length, L, has been found to adequately define stability effects, whereas, 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 model's 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 by normalizing the concentration values by dividing by the respective emission rates. To divide by the emission rate, one has either 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. A decision will have to be made as to how one will 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. It is seen that in defining terms, other questions arise that when resolved will eventually develop 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.

C.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 have been recommended and employed in a variety of investigations, e.g., Fox [55], Weil et al. [56], Hanna [57]. 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 described by Hanna [11], this a 'seductive' inference. Procedures to provide a robust assessment of the significance of the differences are defined in ASTM D6589 [54].

C.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 [58] 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 Normalize 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 [54] to detect: a) which model has the lowest computed *NMSE* value (call this the 'base' model); b) which models have *NMSE* values that are significantly different from the 'base' model.

In the Cox and Tikvart [58] 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 regimes 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 [58] 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 there are only a few of these experiments that 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 [58] 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 the performance of such models in their 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, then evaluations using sparse networks are seen as useful extensions to further explore the performance of well-formulated models for other environs and purposes.

C.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.

C.5.1 Introducing Sensitivity Analyses and Uncertainty Analysis

Section 1.4 of this guidance defines a model as a representation of the behavior of an object or process, often in mathematical or statistical terms. As such, a model approximates reality in the face of scientific uncertainties.

Section 3.1.3.1 identifies and defines various sources of model uncertainty. External peer reviewers of EPA models have consistently recommended that EPA communicate this uncertainty by conducting uncertainty and sensitivity analyses, 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) and when conducted in combination with sensitivity analysis (see definition) 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 [12]; it is the study of how uncertainty in a model output can be systematically apportioned to different sources of uncertainty in the model input [14]. 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 C.1 schematically depicts how uncertainty analysis and sensitivity analysis would be conducted for this model. Uncertainty analysis would be conducted by determining how the model output 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, uncertainty analysis and sensitivity analysis 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.

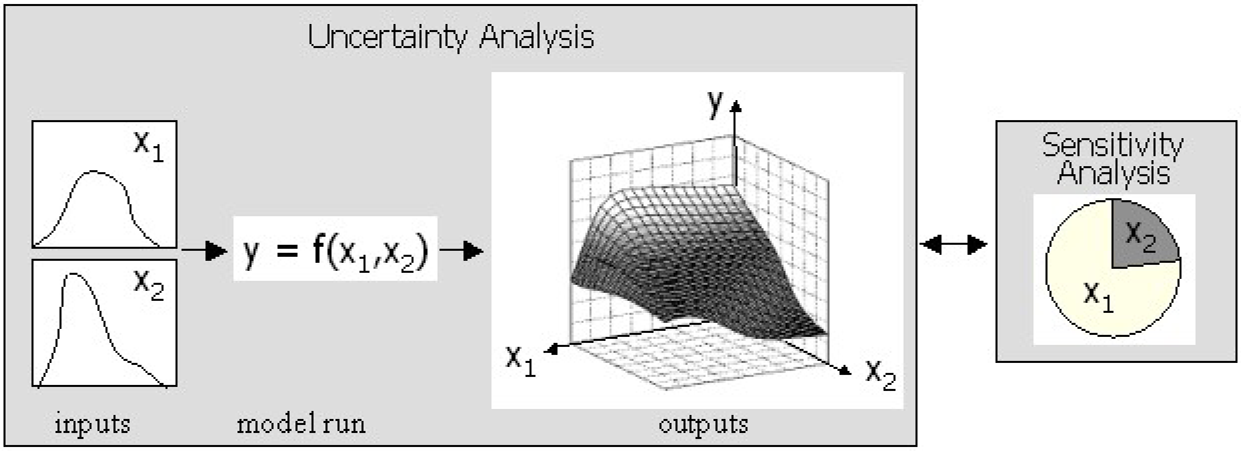


Figure C.5.1 Uncertainty and sensitivity analyses. Uncertainty analysis investigates the effects of lack of knowledge or potential errors of model inputs. Sensitivity analysis evaluates the respective contributions of inputs x_1 and x_2 to output y.

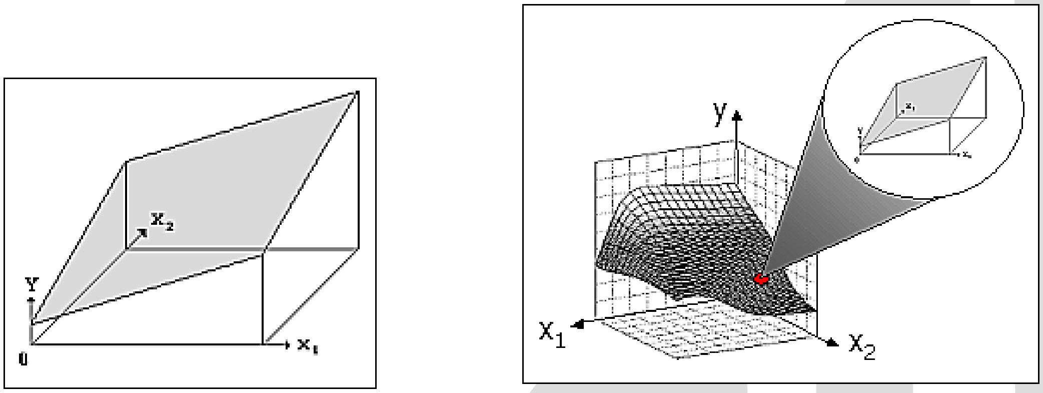
C.5.2 Sensitivity Analysis and Computational Complexity

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:

(a) is able to assume that the model's response surface is a hyperplane, as in Figure C.5.2 (1);
(b) accepts that the results apply only to specific points on the response surface and that these points are monotonic first order, as in Figure C.5.2 (2);[6] or
(c) is unconcerned about interactions among the input variables.

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



(1) (2)

Figure C.5.2. It's hyperplane and simple. (1) A model response surface that is a hyperplane can simplify sensitivity analysis computations. (2) Alternatively, these same computations can 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) based on variance methods.

C.5.3 Screening Tools

<u>C.5.3.1 Tools That Require No Model Runs</u>

Cullen and Frey [50] 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:

(a) *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.
(b) *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

[6] Related to this issue are the terms *Local* and *Global Sensitivity Analysis*. The former refers to SA conducted around a nominal point of the response surface, while the latter refers to sensitivity analysis across the entire surface.

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 [50].

C.5.3.2 Scatterplots

Cullen and Frey [50] 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 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 C.5.3.

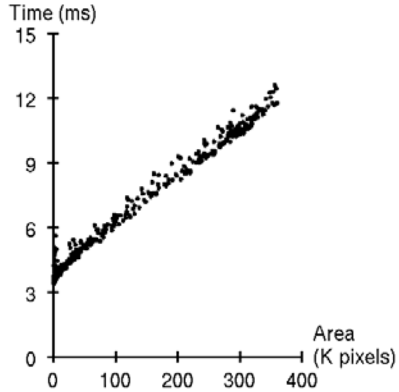


Figure C.5.3. Correlation as indication of input effect. The high correlation between the input variable area and the output variable time (holding all other variables fixed), is an indication of the possible effect of area's variation on the output.

C.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 1) that depend on the choice of base case values. To avoid this bias, Saltelli et al. [59] recommend using Morris's OAT for screening purposes because it is a *global* sensitivity analysis method in that the technique 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 C.5.4 presents the results that one would expect to obtain from applying Morris's OAT [60]. Computational methods for this technique are described in [59].

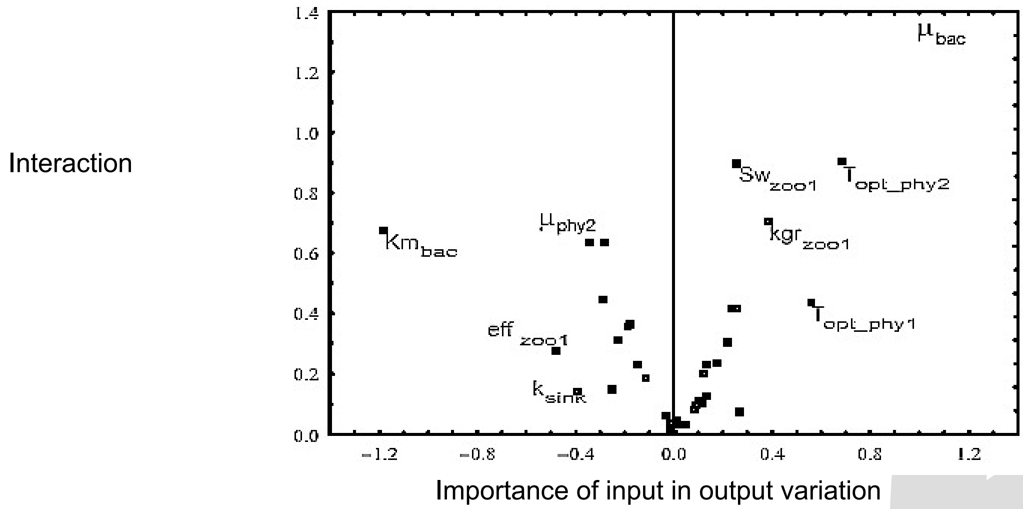


Figure C.5.4. **An application of Morris's OAT.** Cossarini et al. [60] investigated the influence of various ecological factors on energy flow through a food web. Their sensitivity analysis indicated that maximum bacteria growth and bacteria mortality (μ_{bac} and Km_{bac}, respectively) have the largest (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.

C.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 [12, 59].

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 [59]. Computational methods for this technique are described in [12].

C.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 [12, 50]. 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 [50].

EPA proposes several best principles of good practice for the conduct of Monte Carlo simulations [34]. 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: (a) is there any mechanistic basis for choosing a distributional family; (b) is the PDF likely to be dictated by physical, biological or other properties and mechanisms; (c) is the variable discrete or continuous; (d) what are the bounds of the variable; (e) 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 [61].
- *Smaller, structured trials*. The value of Monte Carlo lies not in the randomness of sampling, but rather 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 [50].
- *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. [12, 59]

C.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, then 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 [59].

C.5.7 Which Method to Use?

A panel of sensitivity analysis experts was recently assembled to conduct a review of 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 (a) deal with a model regardless of assumptions about a model's linearity and additivity; (b) consider interaction effects among input uncertainties; and (c) cope with differences in the scale and shape of input PDFs; (d) cope with differences in input spatial and temporal dimensions; and (e) evaluate the effect of an input while all other inputs are allowed to vary as well. [62, see also 63]. Of the various methods discussed above, only those based on variance (§ C.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, Table 1 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.	[50] at pp. 247-8.
Morris's One-at-a-Time	Global sensitivity analysis	Indicates, but does not quantify interactions.	[59] at 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.	[50] at pp. 186-94. [59] at 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.	[50] at pp. 196-237 [12] at 198-216.
Variance-Based	Robust and independent of model assumptions. Addresses interactions.	May be computationally difficult.	[59] at 167-97

C.6 Uncertainty Analysis

C.6.1 Model Suitability

An evaluation of model suitability to resolve application niche uncertainty (§3.1.3.1) 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.

C.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 [12]. 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 degree currently 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 [12].

Literature Cited

[1] U.S. EPA, *Agency Mission Statement*, http://www.epa.gov/history/org/origins/mission.htm, June 11, 2002.

[2] SAB, *Review of Research in Support of Extrapolation Models by EPA's Office of Research and Development*, SAB-EC-87-030, 1987, Extrapolation Models Subcommittee, Science Advisory Board: Washington, D.C.

[3] SAB, *Review of Draft Agency Guidance for Conducting External Peer Review of Environmental Regulatory Modeling*, EPA-SAB-EEC-LTR-93-008, 1993, Science Advisory Board: Washington, D.C.

[4] SAB, *Resolution on the Use of Mathematical Models by EPA for Regulatory Assessment and Decision-Making*, EPA-SAB-EEC-89-012, 1989, Science Advisory Board: Washington, D.C.

[5] U.S. EPA, *Information Quality Guidelines.* Office of Environmental Information. 2002. Washington, D.C.

[6] U.S. EPA, *Proposed Agency Strategy for the Development of Guidance on Recommended Practices in Environmental Modeling*, 2001, Model Evaluation Action Team, Council for Regulatory Environmental Modeling, U.S. Environmental Protection Agency: Washington, D.C.

[7] U.S. EPA, *Quality Assurance Project Plans for Modeling*, EPA QA/G-5M, December 2002.

[8] U.S. EPA, *Guidance for the Data Quality Objectives Process*, EPA QA/G-4, August 2000.

[9] Platt, J.R., Strong inference. *Science* 1964. **146**: 347-352.

[10] Hillborn, R. and M. Mangel, *The Ecological Detective: Confronting Models with Data*. 1997, Princeton, N.J.: Princeton University Press.

[11] Hanna S.R., Air quality model evaluation and uncertainty. *Journal of the Air Pollution Control Association* 1988, **38**: 406-442.

[12] Morgan, G. and M. Henrion, The Nature and Sources of Uncertainty. In: *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. 1990, Cambridge: Cambridge University Press. pp. 47-72.

[13] SAB, *Review of the Underground Storage Tank (UST) Release Simulation Model*, SAB-EEC-88-029, 1988, Environmental Engineering Committee, Science Advisory Board: Washington, D.C.

[14] Beck, M., L.A. Mulkey, and T.O. Barnwell, *Model Validation for Exposure Assessments - DRAFT*, 1994, United States Environmental Protection Agency: Athens, Georgia.

[15] Press. W.H., *Numerical Recipes: The Art of Scientific Computing,* 1992. Cambridge University Press, Cambridge.

[16] Booch, Grady. *Object-Oriented Analysis and Design with Applications*. 2nd ed, Benjamin/Cummings Redwood CA 1994.

[17] Kernigham, B.W. and P.J. Plaugher. *The Elements of Programming Style*. 2nd ed, June 1988.

[18] U.S. EPA, *Guidance on Choosing a Sampling Design for Environmental Data Collection for Use in Developing a Quality Assurance Plan*, EPA QA/G-5S, 2002, U.S. Environmental Protection Agency: Washington, D.C.

[19] U.S. EPA, *Guidance on Environmental Data Verification and Data Validation*, EPA QA/G-8, 2002, U.S. Environmental Protection Agency: Washington, D.C.

[20] U.S. EPA, *Guidance for Data Quality Assessment*, EPA QA/G-9, 2000, U.S. Environmental Protection Agency: Washington, D.C.

[21] U.S. EPA, *Report of the Agency Task Force on Environmental Regulatory Modeling: Guidance, Support Needs, Draft Criteria and Charter*, EPA 500-R-94-001, 1994, U.S. Environmental Protection Agency: Washington, D.C.

[22] U.S. EPA, 2003. Revision to Guideline on Air Quality Models: Adoption of a Preferred Long Range Transport Model and Other Revisions. *Federal Register*, **68** (72): 18440 – 18482.

[23] Suter, G.W.I., *Ecological Risk Assessment*. 1993, Boca Raton: Lewis Publishers. 528.

[24] Beck, M.E., Environmental Foresight and Models: A Manifesto. *Developments in Environmental Modeling*; 22. 2002, Amsterdam: Elsevier. 473.

[25] Oreskes, N.M., K. Shrader-Frechette, and K. Belitz, Verification, validation and confirmation of numerical models in the earth sciences. *Science*, 1994. **263**: p. 641-646.

[26] Beck, B., Model evaluation and performance, In: *Encyclopedia of Environmetrics*, Abdel H. El-Shaarawi and W.W. Piegorsch, Editors. 2002, John Wiley & Sons: Chichester.

[27] Sargent, R.G. Verification, Validation and Accreditation of Simulation Models, 2000. *Proceedings of the 2000 Winter Simulation Conference*, J.A. Joines *et al*. (Eds).

[28] U.S. EPA, *Science Policy Council Handbook: Peer Review*, 2nd ed, December 2000.

[29] U.S. EPA, *Review of Draft Agency Guidance for Conducting External Peer Review of Environmental Regulatory Modeling*, EPA-SAB-EEC-LTR-93-008, June 1993.

[30] U.S. EPA, *Peer Review and Peer Involvement at the U.S. Environmental Protection Agency*, June 1994.

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

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

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

[34] EPA/630/R-97/001, *Guiding Principles for Monte Carlo Analysis*, 1997, U.S. Environmental Protection Agency: Washington, D.C.

[35] Luis, S.J. and D.B. McLaughlin, A stochastic approach to model validation. *Advances in Water Resources,* 1992. **15**(1): 75-83.

[36] Levins, S., The problem of pattern and scale in ecology*. Ecology*, 1992. **73**: 1943-1967.

[37] SAB, *An SAB Report: Review of MMSoils Component of the Proposed RIA for the RCRA Corrective Action Rule*, EPA-SAB-EEC-94-002, 1993, Science Advisory Board: Washington, D.C.

[38] Reckhow, K.H., Water quality simulation modeling and uncertainty analysis for risk assessment and decision making, *Ecological Modeling* 1994, **72**: 1-20.

[39] Borsuk, M.E., C.A. Stow, and K.H. Reckhow, Predicting the frequency of water quality standard violations: A probabilistic approach for TMDL development. *Environmental Science and Technology*, 2002, **36**: 2109-2115.

[40] U.S. EPA, *Protocol for Determining the Best Performing Model.* EPA-454/R-92-025. Office of Air Quality Planning and Standards Research Triangle Park, North Carolina, December 1992.

[41] Saltelli, A., S. Tarantola, and F. Campolongo, *Sensitivity analysis as an ingredient of modeling.* Statistical Science, 2000. **15**: 377-395.

[42] NRC (National Research Council), *Assessing the TMDL Approach to Water Quality Management, Committee to Assess the Scientific Basis of the Total Maximum Daily Approach to Water Pollution Reduction*, 2001.Water Science and Technology Board, Division of Earth and Life Studies, National Research Council, National Academy Press, Washington, D.C.

[43] EPA 100-B-00-002, *Risk Characterization Handbook*, 2000, Science Policy Council, U.S. Environmental Protection Agency: Washington, D.C.

[44] EPA, *Policy and Program Requirements for the Mandatory Agency-Wide Quality System*, May 2000, EPA Order, Classification Number 5360.1 A2.

[45] NRDC v. Muszynski, *268 F.3s 91 (2d Cir., 2001).*

[46] American Iron and Steel Inst. v. EPA, *115 F.3d 979 (D.C. Cir. 1997).*

[47] U.S. EPA, *A Summary of General Assessment Factors for Evaluating the Quality of Scientific and Technical Information*, 2003, Science Policy Council, U.S. Environmental Protection Agency: Washington, DC.

[48] Small, M.J. and P.S. Fishbeck, False precision in Bayesian updating with incomplete models. *Human and Ecological Risk Assessment*, 1999. **5**(2): 291-304

[49] Shelly, A., D. Ford, and B. Beck, *Quality Assurance of Environmental Models*, 2000, NRCSE Technical Report Series.

[50] Cullen, A.C. and H.C. Frey, *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, ed. 326. 1999, New York: Plenum Press.

[51] ANSI/ANSQ E4-1994, *Specifications and Guidelines for Quality Systems for Environmental Data Collection and Technology Programs*, 1994.

[52] *Federal Managers Financial Integrity Act* of 1982, P.L. 97-255 -- (H.R. 1526), September 8, 1982.

[53] U.S. EPA, *EPA Quality Manual for Environmental Programs*, 5360 A1, May 5, 2000.

[54] American Society for Testing and Materials, *Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (D 6589)*, 2000. (Available at http://222.astm.org), 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428.

[55] Fox, D.G., Judging air quality model performance: a summary of the AMS workshop on dispersion model performance. *Bull. Amer. Meteor. Soc.,* 1981. **62**: 599-609.

[56] Weil, J.C., R.I. Sykes, and A. Venkatram, Evaluating air-quality models: review and outlook. *J. Appl. Meteor.*, 1992. **31**:1121-1145.

[57] Hanna, S.R., Uncertainties in air quality model predictions. *Boundary-Layer Met*. 1993. **62**:3-20.

[58] Cox, W.M., and J.A. Tikvart, A statistical procedure for determining the best performing air quality simulation model. *Atmos. Environ*., 1990. **24A**(9):2387-2395.

[59] Saltelli, A., K. Chan and M. Scott, eds., *Sensitivity Analysis*, 2000. New York, John Wiley and Sons.

[60] Cossarini, G., C. Solidoro, and A. Crise. 2002. http://www.iemss.org/iemss2002/proceedings/pdf/volume%20tre/285_cossarini.pdf.

[61] Babendreier, J.E. & Castleton, K.J. Investigating Uncertainty and Sensitivity in Integrated, Multimedia Environmental Models: Tools for FRAMES-3MRA. IN: Proceedings of 1st Biennial Meeting of International Environmental Modeling and Software Society, **2,** 90-95, Lugano, Switzerland.

[62] Frey, H.C., Guest Editorial: Introduction to Special Section on Sensitivity Analysis and Summary of NCSU/USDA Workshop on Sensitivity Analysis, *Risk Analysis.* 2002. **22**, 539-546.

[63] Saltelli, A., Sensitivity Analysis for Importance Assessment, *Risk Analysis.* 2002. **22**, 579-590.