2.0 Taxonomies for Probabilistic Assessment

Helping to define the role of probabilistic risk assessment (PRA), and risk assessment in general for environmental decision-making, useful taxonomic descriptions of verification, validation, sensitivity, uncertainty and variability have been presented by Beck *et al.* (1997), Saltelli *et al.* (2000), Morgan and Henrion (1990), NRC (1994), and (EPA, 1996a, 1997a), and more recently by Cullen and Frey (1999) and Vose (2000). For purposes of discussing 3MRA model evaluation procedures, several conventions and definitions in this chapter were adapted from these texts, in an attempt to better define the probabilistic context for interpretation of 3MRA model inputs and outputs.

A well-described conceptual probability framework helps to reduce linguistic uncertainty (Morgan and Henrion, 1990) associated with describing uncertainty and sensitivity structures of any model or modeling system. Such a framework is employed in the initial discussions of analysis of uncertainty and sensitivity in 3MRA Version 1.0 presented in this document. Benefits of imparting increased precision in terminology extend to interpretation of 3MRA post-processing enhancements, currently under development (3MRA Version 1.x and Version 2.0; see Figure 1-1), that provide additional capabilities for further exploration of model inputs and outputs and related uncertainty and sensitivity structures of 3MRA.

The ultimate purpose of this work is to more precisely state and interpret the underlying science of the 3MRA Framework Methodology (i.e., the Methodology of 3MRA; Marin *et al.*, 1999, 2003), and to show how it has been implemented within FRAMES 3MRA Version 1.0 (i.e., the Technology). This leads to the ultimate objective of characterizing the utility of the 3MRA modeling system output for regulatory decision-making, describing both its strengths and limitations.

2.1 Probabilistic Analysis

While not fully comprehensive, several background definitions of probability theory may be helpful, as summarized by Cullen and Frey (1999) and Morgan and Henrion (1990), and more generally by Ross (1997) and Helton and Davis (2000, 2002, 2003). Listed in Table 2-1, these definitions lay a foundation of the statistical analysis that the 3MRA Framework "Methodology" and associated "Technology" were formed upon. The concepts are explicitly needed to address uses of empirical 3MRA input data that are variable or uncertain from site to site, or within a given site. The concepts are also needed to interpret model output.

Population: A set of individuals representing the outcome of all possible events; defines the probability sample space.

Sample: A single observation or value of a random variable.

Random samples: Samples selected from a population such that each sample has an equal probability of being selected.

Probability samples: Samples selected from a population such that each sample has a known probability of being selected.

Random variable: X; An uncertain quantity whose value depends on chance. A probability model (i.e., a distribution function) defines the likelihood of selecting any particular outcome.

Stochastic process: A random process; a process not currently explained by mechanistic theory.

Frequency of an event: The number of times an individual appears within a population. Relates to knowledge of the total population size. Relative frequency is the % of time it appears.

Probability of an event: P; From the Frequentist (Empiricist) view, the frequency with which an event occurs in a long sequence of similarly conducted trials. From the Bayesian (or Subjectivst or Personalist) point of view, the belief of a person that an event will occur given all the relevant information available to that person. In the latter case, descriptions of belief are sought that are also consistent with the axioms of probability theory.

Probability distribution: f(x); The mathematical description of the function relating probabilities with specified intervals of values, for a random variable.

Fractile: The value of a random variable for a specified percentile of its population.

Probability Density Function (PDF): Graphical or tabular representation of f(x) as the relative likelihoods with which an unknown or variable quantity may obtain various values. The sum or integral of all likelihoods must equal 1 for discrete or continuous random variables, respectively.

Cumulative Distribution Function (CDF): F(x) = P(x #X) = u; Obtained by integrating the PDF. Quantitatively relates the value of a quantity and the cumulative probability (percentile) of that quantity.

Complementary Cumulative Distribution Function (CCDF): 1-u; The cumulative function that gives the probability that a random variable will take on a value greater than some specified value.

Inverse Cumulative Distribution Function (ICDF): G(F(x)) = $F^{-1}(u)$; Essentially an inverse statement of the CDF. The inverse CDF has an abscissa with the values ranging from zero to one, and an ordinate with values representing possible outcomes for the random variable X.

A PDF of the standard normal distribution N(0,1) (i.e., N(mean : = 0, variance F^2 = 1)) is shown in Figure 2-1a, a CDF in Figure 2-1b, and an ICDF in Figure 2-1c. Note the terminology used in this document typically refers to “x” as an outcome in the domain of X, “X” as a random variable, and **x** as a vector of model inputs x_i, where some x_i are random variables.

Typical graphical properties are shown in Figure 2-2 for a given set of finite data defining a probability density function (PDF) and associated cumulative distribution function (CDF). Figure 2-2 could represent the set of measured sample data used to create an input distribution for a given empirical input quantity. It could also represent the set of sampled model output, for example from a Monte Carlo simulation experiment described later. All model outputs represent a measurement of sort. If the model has random inputs, a given model run represents a data point (or set of data points) used to construct the output’s PDF (or set of PDFs). In the case of model output, the set of such points in an environmental risk assessment is interpreted to represent a set of empirically or experimentally observed measurements (Cullen and Frey, 1999).

In real environmental systems, there are never fully clear distinctions between frequency and probability interpretations regarding a specific quantity in a modeling exercise. While much of the data driving a typical fate and transport model or an exposure model is rooted in perspectives of the Frequentist, definition of model inputs and interpretation of model outputs must at some point rely upon belief systems embodying the Bayesian perspective.

2.1.1 Context for 3MRA Populations, Subpopulations, and Individuals

In 3MRA, the term “population percentile” is the percentage of the receptor population protected at a specified cancer risk level or hazard quotient, for a single waste constituent and environmental setting. For the 3MRA national assessment, the receptor “population” of interest would span a set of environmental settings. In 3MRA, an environmental setting is a specific waste management unit (WMU) at a specific site, and is defined by combining site-based information (such as unit size and unit placement) with variable environmental information, such as rainfall and hydrogeologic properties generated from regional and national data. The terms “setting” and “site” are used interchangeably here. In 3MRA, a SettingID is the combination of a WMU type and SiteID (e.g., setting LF0132456 represents a landfill at site #0132456).

The terms population and individual are applied in several ways in the 3MRA modeling system. Examples of population entities in 3MRA are presented in Table 2-2. An individual site, for example, represents a specific member of the set of all sites in the site-based 3MRA database. Through Bayesian inference and the statistical properties of the randomly sampled site-based dataset (i.e., the sampled population of sites), we extend the set of sampled sites to represent the set of all sites across the nation. In extending properties of the sampled population to this probability space, some uncertainty is imposed in characterizing the true national population. With these underpinnings, the implemented 3MRA site-based simulation methodology (Marin *et al.*, 1999, 2003) [or technology] is viewed as a probabilistic risk assessment approach [or tool] to examine the national population of all (existing and future) waste management sites. This, of course, implies the associated sampled database of current national conditions, represented by the sample of existing sites, extends to that future.

Population Protection Criteria

3MRA evaluates numerous protection criteria strategies across a full range of waste stream contaminant concentrations, specified risk levels, and radial distances from the waste management units (WMUs). 3MRA was also designed to handle various formulations of critical exposure time period. Currently in 3MRA Version 1.0, national-strategy post-processing schemes predict the % of sites protected (i.e., the protection criteria) for a given % of protection of facilities' local receptor populations, with a given probability of occurrence. Thus, the distilled output prediction can, for example, be represented as predicting 90% receptor population protection at 95% of all sites, with a 98% probability (or confidence or belief) of meeting this "dual-criteria" population protection level. Providing additional dimensionality in 3MRA, groups of individuals in the set of all receptors nationally can also be formed to describe various subpopulations or cohorts (e.g., sites within a region, aquatic biota at all sites, etc.).

Due to 3MRA's organizational scheme and preservation of outputs at various levels of detail, selected exposure pathways and contact media can also be viewed for separate risk analysis. The extensive 3MRA output dimensions for the risk analysis (Section 4.6.2) provide useful flexibility in performing national, regional, or site-specific studies of risk due to contaminant disposal in land-based WMUs. Thus, the 3MRA technology facilitates examination of risk, and uncertainty in risk, at both population and subpopulation levels, providing for identification of sensitive subpopulations and dominant exposure pathways and contact media. In doing so, 3MRA provides critical information at the tails of associated distributions of risk, allowing identification, for example, of risk levels encountered by sensitive, high-ends of the population or subpopulations exposed (e.g., 90%, 99% population percentiles).

2.1.2 Frequency Vs. Probability for Populations and Individuals

A crucial task in population-based uncertainty analysis is successfully managing the language and implications of variability and uncertainty, discussed further in Section 2.6. The discussion of 3MRA uncertainty analysis subscribes to the convention that variance expressed in model output due to all "certain" stochastic model inputs (i.e., *variable and certain*) is a valid expression of the frequency of occurrence of the response of the target population to WMU contaminant releases (i.e., describing variability of the population or subpopulation of interest). The same model output is also seen to represent an expression of probability (i.e., uncertainty) of randomly selecting a specific individual from the population or subpopulation. For example, use of a regional, random hydrogeologic variable in a site-specific (or site-based) analysis would impart (empirical) uncertainty when used to describe a specific site. There may also be other significant dimensions of uncertainty in play (e.g., random measurement error or distribution sampling error), representing (empirical) uncertainty in population response variability, or additional empirical uncertainty in risk of a randomly selected individual.

Models with "sensitive" random inputs have random outputs (in the context of language used here, the output is actually sensitive to the input). Model outputs (a collection of predictions) can be viewed for most environmental risk assessments as a sample population of empirical data, described by a "frequency" or "probability" distribution depending on one's view. When only "certain" population-based stochastic and/or deterministic input variables are significant relative to model output, (1) total populations experience certain, variable risk, (2) fractiles (e.g., percentages) of populations experience certain, specific (point-estimate) risk, and (3) randomly selected individuals experience uncertain, specific (point-estimate) risk. These roles are further setout in Figures 2-1 and 2-2 for a simple unit model of risk Y = 1*X. Point-estimates represent single values of the random variable and describe population percentile variability and individual uncertainty. In the latter case, the probability of randomly selecting an individual with a given exposure [or risk] is the same as the relative frequency of all individuals in the population subject to the given exposure [or risk] (Cullen and Frey, 1999).

2.2 Taxonomy of Model Input Quantities

Various model inputs can be classified by type. Inherently, assigning a "quantity type" to any model input is a subjective, interpretational procedure. Assignment depends on context and intent of the model and user, and can vary among persons for a given model input. Regulator and regulated, for example, may have different interpretations of what some model inputs represent, and what assignments might apply for a given model (Cullen and Frey, 1999). Each quantity type, as indicated below, takes on an assumed role in uncertainty analysis. A well-defined presentation of model evaluation methods and model results can provide for a more transparent communication and inspection of uncertainty. Said methods and results facilitate recognition and understanding of any differences held by interested parties in the interpretation of the roles of various model inputs and outputs.

The following summary of quantity types, listed in Table 2-3, was adapted from Morgan and Henrion (1990) and Cullen and Frey (1999):

Empirical: Measurable properties, at least in principle, of the real world-systems being modeled. Have a true value. Typically these are the only variables amenable to probabilistic uncertainty analysis. In modeling applications, can include both subjectively and empirically derived inputs.

Defined constants: Fundamental constants representing assumed entities in the natural system (e.g., days of the week, π, etc.). Includes some empirical quantities usually assumed to have insignificant uncertainty (e.g., Plank's constant, speed of light, etc.).

Decision variables: Policy variables under the control of the decision-maker that have no true value, (e.g., the protected population percentile chosen, permissible loading rate, acceptable level of risk, etc.). In most probabilistic analyses, typically should not be considered to be uncertain.

Value parameters: The preferences of value judgments of the decision-maker or the people they represent (e.g., cost-benefit discount rate, risk tolerance, value of life, etc.). Can be informed by related information on other's values, but is usually incorrectly applied as an empirical quantity.

Index variables: Used to identify a location or cell in the spatial or temporal model domain. Can also be used to describe a member of a set of elements (e.g., an individual of a population).

Model domain parameters: A function of the model, but not the phenomena represented. Generally used to specify the range and increments for index variables (e.g., spatial scale of land use data, surface water model time-scale, number of habitat types considered, number of C_W's, air spline technique = ON, etc.). While one may not know the most appropriate value to use, these do not have a true value.

Precision in sampling is defined here as the relative agreement among repeated measurements or predictions of the same quantity. Accuracy refers to the relative agreement of a measurement or prediction with the true expected value.

2.3 Model Evaluation

In a regulatory setting, model evaluation can be seen as an iterative procedure, integrating risk assessment and risk management phases of problem statement, research and development, and application of the model product for an intended use or purpose. As a public matter, for use in solving a complex problem statement, model evaluation is often recognized as a progression of maturation marked by periodic exchange between science, technology, and regulation. At various junctures, the problem, it self not stagnant, becomes increasingly resolved, forming a series of distinct models and decisions over time. In retrospect, the resultant archive forms a documented construction of model synthesis and model analysis, representing a trajectory of strengthening belief in the utility of the final model's predictions. 3MRA Version 1.0 is an expression of such a trajectory. The question at hand can be phrased as, "Has 3MRA matured to a reliable enough state for its intended use?". The answer will always retain subjectivity, and ultimately relies on professional judgment of decision-makers who use the model's outputs. Best judgment will underlie interpretation of the best answer captured by any modeling exercise.

This question stated is best framed in the context of first defining a practical perspective of the nature of models. Summarized by the researcher Beck and attributable to Ravetz of Beck *et al.* (1997), a model should not be viewed as a truth-generating machine which subsequent observation will reveal as indeed true or false, but rather a tool designed for a specific purpose, just as is a screwdriver or hammer. A practical focus for model evaluation thus transforms to the task of trying to quantify the "goodness" or "badness" of its design, where Beck *et al.* summarize several tasks or purposes (or functions) of a model may be to provide:

- A succinctly encoded *archive* of contemporary knowledge,
- An *instrument* of prediction supporting decision-making or policy formulation,
- A *device* for communicating scientific notions to a scientifically lay audience, or
- An explanatory *vehicle* for discovery of our ignorance.

In their paradigm, the terms *archive*, *instrument*, *device*, and *vehicle* are suggestive of a model as a tool, and model evaluation as an assessment of the tool's design. The question introduced directly above can then, alternatively, be couched in the notion, "Does 3MRA provide intrinsic value to EPA as a tool in making decisions and formulating policy regarding land-based disposal of hazardous waste?".

As a starting point to evaluate the 3MRA Version 1.0 tool, the concept of model evaluation may be framed in a three-step procedure of:

1. Internal validation (analogous to a multi-tiered verification procedure of model composition);

2. Performance validation for a specific use (i.e., a comparative measure based on either history matching or some "objective" test of relevance); and

3. Predictive uncertainty analysis, being a statement of final prediction uncertainty for the specific use.

Various procedures have been or will be undertaken for 3MRA to provide for internal validation and predictive uncertainty analysis, and various investigations of unit-level and system level performance validation. For 3MRA, due to its underlying science-design approach (Marin *et al.*, 1999, 2003) that assesses novel, future conditions, performance validation will also be undertaken via the Young-Hornberger-Spear sensitivity algorithm, as outlined by Chen and Beck (1999). An objective test of relevance, the latter technique represents a reflection of the evaluation of the external definition of the task, back onto the internal composition of the model (Beck *et al.*, 1997). To communicate EPA's overall approach for verification, validation, and uncertainty analysis of 3MRA predictions (Section 9), introductory context is first needed.

2.3.1 The Model Validation Paradox

Extending beyond a simplistic, unworkable view of retrospectively oriented model verification and validation exercises rooted in history matching, components of model evaluation for 3MRA are viewed as inextricably linked to a familiar concept of quality assurance in product (tool or technology) design (Beck *et al.*, 1997). "Use" in regulatory decision-making typically implies the final exercise of the model as a forecast of some subjectively determined protection level of human health and the environment. Only direct auditing of future attainment of the desired risk assessment objective (e.g., a certain level of protection achieved by a specific waste constituent management strategy over time) could begin to approach full illumination of the model's success, and our grasp of science involved. Even then, such a determination, if it were feasible to construct, would realistically remain, after the fact, a partially subjective conclusion for complex problem statements such as those addressed by 3MRA.

For example, it is arguably untenable that one could go about verifying 30 years from now that 30 years of past waste management practices have imparted a specific increased risk of cancer for 300 million human beings, or even 100,000. That there is inherent subjectivity in any post-audit determination becomes increasingly unimpeachable if we add to this the perspective of auditing some quantified level of protection for ecological systems from the same practices.

Our focus for the time being is placed on a more attainable, tactical challenge of evaluating the 3MRA technology for a specific use in the present. The present use is the task of predicting future system behavior under novel conditions, an unobservable future for the time being. In summary, the problem of reaching a satisfactory, empirically based measure of validation in the present is restrained by two dilemmas: (1) the future truth we seek is paradoxically unobservable in the present, and (2) subjective decision variables used in complex problems, such as exposure and risk assessments, are realistically unobservable in the present and future. Fundamentally a dilemma of extrapolation, Beck *et al.* (1997) offer a formal characterization of the general nature of this paradox and potential approaches available to deal with it. They state:

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

As Beck *et al.* (1997) indicate, despite these dilemmas, the existence of the validation paradox does not render us without an opportunity for some level of objective judgment about the uncertainty in a given decision, and the overall performance validity of a model for a specific intended use. Adopting terminology and procedure of Beck *et al.* (1997) and Chen and Beck (1999), for 3MRA, this judgment can be made in the present based upon how well we expect the model to perform its designated task reliably, with a minimum risk of an undesirable outcome, and in a maximally relevant manner.

2.3.2 Validation and Uncertainty in Prediction

The entities and procedures described in this section, as an attempt to put perspective on the problem at hand, are built upon the works of several researchers, and help define major elements of an overall model evaluation strategy for 3MRA. Of course, not all problem-specific aspects can be delineated here, and such treatment may require additional detail at several levels.

The abstraction, "model evaluation", is used here as a convenience to both avoid complete prejudice in the term "model validation", and to place into perspective our basic strategy of evaluating both validation of the 3MRA model for a specific use, and to describe uncertainty associated with its use as a forecasting tool. The term "model evaluation" is still somewhat uncharted as a well-defined notion that encompasses the entire model development (i.e., synthesis) and application process (i.e., analysis). A more robust perspective of the term "validation", as most decision-makers and stakeholders have historically used it, must reach well beyond our existing familiarity with the term (Beck *et al.*, 1997; Burns 2001).

Where used here, the terms "evaluation" (Oreskes, 1998, ASTM 1984) and "quality assurance" (Chen and Beck, 1999; Beck and Chen, 2000) should not be interpreted to represent an agreed upon convention of all authorities on the subject of evaluation of model performance for a specific use. Our adoption of these terms as useful descriptors of the confidence building process in a model's use for a specific purpose is nonetheless indebted to the continually evolving debate on this topic, traced by Beck *et al.*, (1997) through the works of Caswell (1976), Burns (1983), Burns *et al.* (1990), Konikow and Bredehoeft (1992), Oreskes *et al.* (1994), and further illuminated upon through the works of Oreskes (1998), Beck *et al.* (1997), Chen and Beck (1999), Beck and Chen (2000), Burns (2001) and others.

Summarizing evaluation of the 3MRA modeling system's "synthesis" and "analysis" stages, verification, validation and predictive uncertainty are the focus of this investigation, more precisely organized under the construct of model evaluation components described below. In this context, model evaluation, as a sum total effort, can be ultimately viewed as a characterization of uncertainty in decision-making.

2.3.3 Model Synthesis and Analysis

A key concern moving through the components of model development, evaluation, and application, is ensuring that knowledge is transferred to the scientist, developer, analyst (i.e. user), decision-maker, and peer-reviewer regarding the documented outcomes of a systematically applied quality assurance program used to:

(a) State the problem,
(b) Develop a science-based methodology for solution of the risk assessment objective,
(c) Assess computational objectives for implementing the methodology,
(d) Build the methodology into a reliable technology (i.e., create the modeling system),
(e) Concurrently develop model input needed to achieve the risk assessment objective,
(f) Compile the technology and verify (i.e., thoroughly test) the technology,
(g) Independently recompile and verify (i.e., thoroughly test) the technology,
(h) Validate the internal compositional properties of the model,
(i) Document all facets of the development process
    a. Problem statement,
    b. Methodology, and
    c. Technology,
(j) Assess final computational requirements for use of the technology,
(k) Apply the technology to the problem statement,
(l) Validate the external performance of the model's use for the purpose identified, (i.e., communicate validity (or quality assurance) of the model's use as a tool), and
(m) Communicate uncertainty in model predictions to decision-makers and stakeholders.

In developing a fully consistent problem statement, the reality of reaching a successful description of model validation for a given purpose will require not only a statement of the desired risk assessment objective, but also a description of undesirable outcomes of performance (Beck *et al.*, 1997; Burns *et al.*, 1990; Burns, 1983, 2001).

2.3.4 Pedigree of Information

Qualitative in nature, a significant element of model evaluation (or assessment of quality assurance in a model application) is communicating the pedigree of the model application's constituent information. This concept is formally introduced in the quantitative-qualitative NUSAP scheme for model evaluation of Funtowizc and Ravetz, (1990), discussed in Section 1.4.1. In steps (d-h) above, one could identify a given problem statement with an existing technology and then construct post-development strategies for quality assurance in an application. Such an approach still faces the same need to establish pedigree of the constituent information in asserting the underlying state of quality. Attention to pedigree is critical to facilitating inspection, peer-review, and reproducibility of the problem-solution domain, and ultimately in reaching a conclusion about the acceptance or rejection of a model's predictions for some intended use.

The larger or more complex the problem and model, the more iterative the technology development and application process typically becomes. This fact dramatically increases the need for: (1) more automated testing procedures for software code; (2) verification of the appropriateness of input data used; and (3) early and periodic engagement with substantial peer-review efforts. In the science and technology development of 3MRA, these three aspects were purposefully integrated into the model evaluation program as a principal activity engaged-in for all of its components.

Model size and complexity are typically positively correlated attributes of a model. For large or complex models, model validation is also more difficult to establish objectively, due in large part to an inability to conduct enough independent experimentation with the model to validate even small portions of the model's problem-solution domain. Discussed previously, observation-based validation of the 3MRA application for the national assessment of land-based hazardous waste disposal, as a history or future matching exercise, is intractable as a practical matter. This is not a unique property of 3MRA, as this describes many models in use today.

2.3.5 Components of Model Evaluation

Components of model evaluation are listed in Table 2-4. Definitions relating to model, validity, uncertainty analysis, and peer review given below are adapted in part from Beck *et al.* (1997). Definitions of model, sensitivity analysis and calibration, and, in part, verification of software codes, are themselves derived from ASTM (1984), with a practical definition for sensitivity analysis taken from Saltelli (2002a). Definitions of uncertainty, variability, total uncertainty, and in part, peer review, are further adapted from Morgan and Henrion (1990), Cullen and Frey (1999), and Vose (2000).

Model: A complex assembly of several constituent hypotheses. Alternatively, an assembly of concepts in the form of a mathematical equation that portrays understanding of a natural phenomenon.

Uncertainty (U): Lack of knowledge about: (1) the "true" value of a quantity that could be established if a perfect measuring device were available; (2) which of several alternative model representations best describes a mechanism of interest; or (3) which of several alternative PDFs best represent a quantity of interest. A property of the analyst that can be reduced in some cases.

Variability (V): Refers to temporal, spatial, or interindividual heterogeneity in the value of an input. These inputs describe populations of quantities, not quantities of particular individuals. A property of nature that cannot be reduced.

Total uncertainty (TU): The combination of uncertainty and variability (U|V). Refers to the overall indeterminacy of describing characteristics of a randomly selected individual of the target population. Historically, the term "uncertainty" has been used to describe both TU and U.

Compositional uncertainty analysis (UA_c): Evaluation of the ranges (or distributions) of values that can be assigned to the model's parameters, where said evaluation can be made, *inter alia*, on the basis of model calibration as a function of the specified sources of uncertainty associated with the data used for the test.

Performance uncertainty analysis (UA_p = UA): Evaluation of the ranges (or distributions) of values that are associated with the predictions of the model's output variables as a function, *inter alia*, of the uncertainty in the model's input values.

Sensitivity analysis (SA): The degree to which the model output is affected by changes in a selected input parameter. Pragmatically, a study of how the uncertainty in output of an analytical or numerical model can be apportioned to different sources of uncertainty in the model input.

Calibration (CAL): A test of the model with known input and output information that is used to adjust or estimate factors for which data are not available.

Code Verification (CodeVer): Activities that evaluate the model's underlying computer code such as examination of the numerical technique in the computer code to ascertain that it truly represents the conceptual model, and that there are no inherent numerical problems in obtaining a solution.

Model comparison (ModComp): A comparison of two or more models. Represents an assessment of model group precision. Based on assumptions of similar (equivalent) model inputs and phenomena handled. Where one or more of the models being compared are "validated", can represent a performance validation as a relative assessment of model accuracy.

Compositional Validity (CompVal): Internal validity or face validity. Represents a judgment of the validity of the composition of the model based on how its constituent hypotheses are assembled, that attaches to either each constituent hypothesis, or to the model as a whole, or to both. Reflects the generic properties of the model, independent of the current task (application).

Performance Validity (PerfVal): External validity. A judgment of the validity of the performance of the model in terms of being a valid instrument for undertaking a task assigned to it. Involves some comparison of data derived from the model with data (or conditions) deduced from sources of knowledge independent of the specific model under judgment.

Performance sensitivity analysis (SA_p): A specific construction of performance validation as a reflection, by way of sensitivity analysis, of the evaluation of the external definition of the current task back onto the internal composition of the model.

Model Validation (ModVal): A judgment about the overall validity of a model based on whether a model can perform its designated task reliably (i.e., at minimum risk of an undesirable outcome and in a maximally relevant manner).

Peer review: Performed by independent and objective experts, a review of and judgment on a model's underlying science, the process through which it was developed, and its overall "trustworthiness" and "reliability" for prediction. For large or complex models, a proper external assessment will typically require a multidisciplinary team, and a significant budget.

Henceforth, the term "predictive uncertainty analysis" or simply "uncertainty analysis" will be used to generally refer to the output "performance uncertainty analysis" as defined above.

As a conceptualization of roles and relationships, the components are depicted in Figure 2-3, indicating each component's role in the overall process of model evaluation and decision-making. Quality assurance in design is a result of the model synthesis and analysis effort, and is viewed as the outcome of the overall model validation effort for the specific task defined. Burns (2001) also gives a broad, foundational perspective on various model evaluation components described below, and their role in environmental exposure assessment.

In distinguishing attributes of composition versus performance, the first of two stages in developing the model application, "model synthesis", is completed by the act of code verification, at which point the model's compositional validity is established (Beck *et al.*, 1997). Proceeding from here, "model analysis" is initiated, including aspects of model input specification for the task at hand, calibration, if used, and performance validation. When calibration is performed (an initial assessment of model performance), performance can be cast in Bayesian terms as an evaluation of "prior" and "posterior" validity (Beck *et al.*, 1997; Burns, 2001). Referring to the predominantly held view, validation has historically been defined as a comparison of model results with numerical data independently derived from experience or from observations of the environment (ASTM, 1984).

Cullen and Frey (1999) provide a more informal description of (performance) uncertainty analysis as the computation of the total uncertainty induced in the output by quantified probabilistic inputs. Variability and uncertainty can be evaluated jointly within the same dimension of probabilistic analysis, or disaggregated along separate dimensions. Sensitivity is defined analogously as the computation of the effect of changes in input values or assumptions (including boundaries and model functional form) on the model output.

2.3.6 Model Use and Desired Outcome

For regulatory decision-making, model evaluation inherently seeks to identify how well one has integrated problem identification, risk assessment and risk management needs throughout model development and the exercise of the model for its intended application.

Together, uncertainty and sensitivity analyses; model verification, comparison, and validation; and peer review tasks contribute to a formal program of model evaluation. These are intended to establish a requisite level of confidence in a model's use for a given prospective task of prediction (Beck *et al.*, 1997). As stated previously, for large or complex models, overall model evaluation is best reflected upon as an iterative process of confidence building (or development of quality assurance). The challenge of solving increasingly large or complex modeling problems imparts a substantial challenge in describing the uncertainty in and quality of predictions, and in making decisions there upon. Incorporating a requisite level of quality assurance throughout design and execution becomes of the utmost importance. Meeting this challenge requires a systematic, iterative approach that cannot realistically be conducted in a vacuum external to statements of the specific purpose and implications of the decision-making.

Model evaluation is intended to eventually culminate in an integrated, qualitative and quantitative documentation of the state of quality of the model for a specific use. It is a statement of quality assurance in the model that inevitably matures over time. The NRC (1994) has previously underscored the importance and need for appropriate and balanced use of models at key stages of an implied model evolution, evaluation, and use paradigm. Decision-making ultimately ceases if made to rely upon perfection of the model development process and its forecasting quality for assessment of a given problem. Model use for a specific purpose is in the end a balancing act performed by decision-makers, who must weigh the quality of current knowledge against the detriment of inaction. There is no set threshold through which one may pass with disregard to the need to assess, in the future, the impact (or quality) of any decision.

2.3.7 Validating Predictive Exposure and Risk Assessments

Konikow and Bredehoeft (1992) were catalysts in refocusing substantial debate on model validation in water-quality sciences with their well-known proclamation that groundwater models cannot be validated (Beck *et al.*, 1997). They point out that simple use of terms like "validation" and "verification" can lead to a false impression of model capability, offering a preference for shifts in language towards terms like testing, evaluation, calibration, sensitivity analysis, benchmarking, history matching, and parameter estimation. Oreskes *et al.* (1994) expanded upon this notion, establishing that literal interpretation of validation is an impossible task. Illuminated further by Hassanizadeh and Carrera (1992), Beck *et al.*, (1997), and Burns (2001), differences between existing definitions of validation and verification are not a matter of semantics in objective language (i.e., meaning), instead they are a function of the perceptions we hold behind these terms. Burns (2001) argues the case of keeping language intact and, instead, improving upon perception; in effect a recommendation for restraint on societal tendency and predilection for renaming problems not fully understood, or which cannot be readily solved.

To continue to raise this discussion of 3MRA evaluation further out of the sea of linguistic ambiguity, the following statements attempt to capture the essential properties of the conceptual framework for model validation in environmental exposure [and risk] assessments outlined by Beck *et al.*, 1997:

- As model size and complexity increase, the degrees of freedom in assigning input conditions that match a given history also increase. Consequently, there is significant uncertainty in assigning an appropriate set of input conditions for future predictions.

- Traditional views of validation as "prior" and "posterior" performance assessments (comparisons of model predictions and observations of the system) do not inherently validate a model for other uses (e.g., forecasting; see also Burns (2001)).

- Validation in the familiar sense remains an intractable issue in environmental forecasting (i.e., prediction of futures). Familiar definitions are retrospective, depending upon past observations. The concept of predictive exposure assessment is by definition impossible to validate in this context.

- The framework is foundationally built on works of Caswell (1976), Burns (1983), and Burns *et al.* (1990), who solidify attention to the concept of "purpose". A judgment about the validity of a model cannot be made in the absence of a specific purpose. Example purposes can range from mechanism research, prediction of (future or alternate) system behavior, or regulatory compliance assessment.

- An essential point derived from Caswell (1976), Burns (1983), Burns *et al.* (1990), and Beck *et al.* (1997) is that validation is essentially a problem of design. Validation should be understood as a task of tool design for which some form of protocol for quality assurance will ultimately be needed. Quality assurance must be an integral part of design and development.

- Compositional validity is an internal measurement of the model; performance validity is an external measurement of the model made in context of its intended use.

- A model may be viewed as a complex assembly of many constituent hypotheses. In this respect, assessment of the strengths and weaknesses of each constituent hypothesis is as important as examining the validity of the model as a whole. (*For 3MRA, this statement also extends to module-level versus system-level compositional validation*).

- Quantitative descriptions of the compositional validity of a model can be tied to perspectives of model input uncertainty, and is best suited overall to a process of group peer-review, which can provide a basis for integrating expert judgment and historical experience.

- As a performance validation measure, the authors build upon the works of Young, Hornberger, and Spear in developing the notion of a model having maximum relevance to the performance of a specific task (e.g., predictive exposure assessment), broadening the discussion of model validation into one of quality assurance in environmental forecasting.

- Maximum relevancy is an assessment of key (sensitive) and redundant (insensitive) inputs, an identification process irrespective of output uncertainty, placed in context of a binary classification of "behavior" and "non-behavior" in the model output space.

A methodology for establishing performance validation of model behavior under novel (i.e., unobservable) conditions is detailed in a simplified example of a national application for multimedia modeling of hazardous waste disposal (Chen and Beck, 1999; Beck and Chen, 2000). These works serve as reference points in part of the approach to assess performance of 3MRA Version 1.x. Section 9 discusses the overall model evaluation approach to be undertaken for 3MRA. Sections 2.8 and 9.3.5 specifically discuss aspects of performance sensitivity analyses as a mechanism for quantification of performance validation under novel, future conditions.

2.3.8 Summary of Model Evaluation Concepts

To restate and capture the essence of model evaluation in lay terms, Beck *et al.* (1997) offer three "Elements of Judgment" as essential questions to be considered:

- Has the model been constructed of approved materials (i.e., approved constituent hypotheses in scientific terms)?
- Does the model's behavior approximate well our observations of the real thing?
- Does the model work (i.e., does it serve its intended purpose)?

To layout the overall approach for evaluation of the 3MRA modeling system undertaken by EPA (i.e., verification, validation, and prediction uncertainty), additional background and clarification are first needed to more precisely define uncertainty, variability, and sensitivity.

2.4 Taxonomy of Uncertainty

As delineated by Beck *et al.*, (1997), errors in predictions of a model may derive from three sources, conceptually describing predictive uncertainty analysis, defined previously:

1. The estimated initial state of the system at the start of the forecasting horizon,
2. The assumed patterns of future variations in input disturbances, and
3. The model.

Constructed analogously, as outlined by Cullen and Frey (1999), total (predictive) uncertainty, as defined by Vose (2000), can be described along lines of:

1. Input variability,
2. Input uncertainty,
3. Model error.

In resolving the perceptions among elements of the sources outlined, it is important to recognize that model error may impose an interpretational issue, depending on how it is explicitly treated in synthesis and analysis stages. To the degree that calibration is conducted, for example, some model error will be included within descriptions of input uncertainty, or analogously, within the initial state of the system represented. This convolution can be seen as an aspect of internalizing some portion of model error as compositional uncertainty.

Quantity types in Table 2-3 and Figure 2-4 indicate the typical amenability of each to probabilistic analysis of uncertainty and sensitivity. Empirical inputs typically are subjected to uncertainty and sensitivity analysis for a given model. Other quantity types in Table 2-3 can be explored for output sensitivity to assess their impacts on decision-making. As defined by Morgan and Henrion (1990), the categories listed in Table 2-5 indicate various sources of total uncertainty in empirical quantities, where, as previously discussed in Section 2.1, population variability may be treated as a type of empirical uncertainty about individuals of the population.

Built upon discussions and definitions offered by Morgan and Henrion (1990) and Cullen and Frey (1999), types of uncertainty are outlined in Table 2-6, and in Sections 2.4.1 (input variability), 2.4.2 (input uncertainty), and 2.4.3 (model error). Input uncertainty, as defined here, represents errors in empirical input quantities, henceforth referred to as “empirical uncertainty”. Attention is devoted here to descriptions of empirical and model uncertainties, and later to distinctions regarding treatment of input variability and uncertainty in predictive uncertainty analysis. Table 2-7 calls attention to various terms researchers and the public have assigned to the two “natures” or components of total uncertainty (i.e., variability and empirical uncertainty).

2.4.1 Variability

Variability, previously discussed in Section 2.1, is briefly revisited. Model outputs (a collection of predictions) can be viewed for most environmental risk assessments as a sample population of empirical data, described by a “frequency” or “probability” distribution depending on viewpoint. For a given set of conditions and output criteria (Section 4.6.2), when only “certain” population-based stochastic and/or deterministic input variables are significant relative to model output: (1) total populations experience certain, variable risk; (2) fractiles (e.g., percentages) of populations experience certain, specific (point-estimate) risk; and (3) randomly selected individuals experience uncertain, specific (point-estimate) risk.

2.4.2 Input Uncertainty

The following are adapted from definitions offered by Morgan and Henrion (1990) and Cullen and Frey (1999):

Systematic error (SE): A relative measure of accuracy. Represents constant error in a set of measurements imposed by a systematic bias in the measurement technique or experimental design. Typically refers to the difference between the true value of the quantity of interest and the mean of the population of measurements. Affects all measurements similarly. Compared to random errors, it is more difficult to identify and minimize.

Random error (RE): A measure of precision associated with imperfections in a measurement instrument or observational technique, or with processes that are random or statistically independent of each other. Random measurement error is inversely proportional to precision of the measurement instrument. Typically refers to the deviation of a measurement from the mean of the population of measurements. Affects all empirical quantities.

Input sampling error (ISE): A sub-type of random error introduced by limited sampling of the target population used to define an associated input distribution. Manifests as an error in describing distribution function parameters (e.g., mean and variance).

Output sampling error (OSE): A sub-type of random error introduced by random sampling of an output distribution, for example as represented by imprecision in finite Monte Carlo Simulation

(MCS) experiments. Can be reduced to insignificant levels by increasing the number of model runs made, if computationally feasible.

Other types of input uncertainty may include errors associated with:

- **Inherent randomness** that is irreducible in principle (e.g., Heisenberg's Principle),
- **Lack of empirical basis** in predictions about things not yet built, tested, or measured,
- **Representation** such as in sampling from non-target populations,
- **Correlation** between two or more functionally dependent empirical quantities, and
- **Disagreement** in parameterization, e.g., as differences in agreement between experts.

Total sample measurement error (SME) will always have two error components (systematic and random). Describing total measurement error, systematic and random error can be quantified separately or jointly (Vose, 2000; Cullen and Frey, 1999). In the case of correlation error, dependence between random variables can increase uncertainty in the variance of model outputs. Marin *et al.* (1999) introduces several sources of "non-sampling" empirical input uncertainty that may be grouped by representation and/or correlation errors, such as sampling from non-target populations, and errors in surrogate non-probability samples used in correlated data structures.

Notwithstanding the objective nature of empirically measured quantities, there is always some level of subjectivism (or subjective error) introduced into the statistical representation of all empirical quantities, due to a lack of absolute empirical knowledge. As such, every empirical quantity at any point in time retains, in principle, some non-zero levels of reducible error and irreducible error. Like decision quantities or value parameters, empirical quantities that are parameterized on a purely subjective basis are not amenable to statistical representation unless derived from empirically characterized analogous systems. Distribution type misspecification associated with statistical representations of random error is an example of subjective error.

2.4.3 Model Uncertainty

Uncertainty introduced by model error (ME) broadly includes model specification error and model conceptualization error, both associated with development, selection, and implementation of a model for a given purpose. Amenable to measurement through comparative study, modeler error, characterized as subjective input into the model description or conceptualization process, would also be included here. Model error can be conceptually broken down along lines of uncertainties related to model structure, boundary specification, scenario description, simplifying assumptions, resolution (i.e., choice of model domain parameter values), extrapolation beyond a model's intended application domain, and interpolation between point estimates of model output (Morgan and Henrion, 1990; Cullen and Frey, 1999).

.

Quantification of model error is an aspect of subjective judgment that is difficult to assess objectively. If segregated from empirical input errors, it can be treated similarly during uncertainty analysis. Some model error may be purposefully (e.g., calibration) or unknowingly (e.g., subjectively) incorporated during parameterization of the model, where explicit knowledge of this may be different for data collector, model developer, model user, and decision maker

(Cullen and Frey, 1999). Systematic bias can be explored through comparison of different models describing the same phenomena. Simplifying assumptions imparted during model development represent uncertainty as approximations of real-world systems.

Domain parameters are typically specified during the model development stage, representing a model's implied range of capabilities (e.g., model application scale). During model selection, the modeler attempts to assign capabilities commensurate with a given problem statement and the desired risk assessment objective. Selection of values outside the intended use of the model represents subjective extrapolation, and leads to increased uncertainty in model output, expressed as model error. Burns (2001) provides a more detailed outline of the impacts of extrapolation and interpolation for probabilistic exposure assessment.

Despite model domain parameters lack of "truth" characteristics, one may want to view impacts of these variables on the modeling approach and output through sensitivity analysis. Model domain parameters have potentially significant impact on model output, controlling the precision of the model representation and computational complexity (Morgan and Henrion, 1990). Morgan and Henrion suggest that domain parameters should not typically be interpreted as a subjective (i.e., uncertain) probability distribution, but should be subject, where feasible, to parametric sensitivity analysis to determine impacts on model output.

Reflecting on model error and associated uncertainties imposed in composite modeling systems, in the case of 3MRA, governing equations are solved in a downstream order using analytical or semi-analytical techniques. The benefits of this approach for addressing complex risk assessment questions, such as those considered by 3MRA on a national scale, include: (1) the explicit presentation and representation of underlying multidisciplinary sciences; (2) a more intuitive understanding and interpretation of model inputs and outputs; and (3) a lack of numerical instability associated with spatial and temporal grids imposed by numerical solution methods. In essence, these represent the benefits of maintaining a separation of multiple constituent, scientific hypotheses in a single, complex modeling system.

Alternatively, more complex models requiring numerical methods (i.e., finite difference or element techniques) can offer increased model accuracy given that increased resolution of input details demanded by these approaches can be acquired with sufficient quality. Here, for example, dynamic time-step linking can be employed across media accounting for feedback through solution of an integrated differential equation. Detractions of using such an approach for complex, integrated models like 3MRA are: (1) the increased needs for input data, (2) increased computational capacity to handle longer model runtimes, (3) increased complexity in dealing with uncertainty. The degree of complexity chosen for an integrated modeling system must in the end balance these issues.

2.4.4 Total Uncertainty

To be revisited in Section 2.6, the 3MRA methodology conceptually parses uncertainty analysis into individual components of variability, input sampling error (ISE), sample measurement error (SME), and model error (ME), and, further, output sampling error (OSE) associated with each of these. These acronyms represent a refinement of terminology and acronyms originally used in describing the underlying science methodology of 3MRA (Marin *et al.*, 1999, 2003). Correlations among inputs in 3MRA are handled either explicitly in mathematical formulation, in paired sampled inputs measured site-by-site (e.g., hydrogeologic model inputs), or in the form of specification of otherwise correlated joint probability distributions (e.g., hydrogeologic model inputs).

2.5 Monte Carlo Simulation

Large or complex models, such as 3MRA, are not typically amenable to model solution through analytical or transformation techniques. Such approaches are generally intractable for multimedia exposure and risk models that assess contaminant travel through many pathways, account for many biological, chemical, and physical interactions and transformations, and that predict effects in many ways, depending on receptor/cohort-specific characteristics.

Summarized in Table 2-8, Cullen and Frey (1999) provide a useful summary of the taxonomy of various solution methods and probabilistic methods for propagating distribution moments (e.g., mean, variance or spread, skewness, kurtosis, etc.) through models. The reader is also referred to Saltelli *et al.* (2000), Cullen and Frey (1999), Frey and Patil (2002), and (Vose, 2000) for a more detailed discussion of the value, attributes, and limitations of the various methods shown. The general benefit of Monte Carlo Simulation (MCS) is that it can be applied to a wide variety of difficult problems in science and mathematics without concern to many of the restrictions in assumptions underlying other techniques shown in Table 2-8.

Monte Carlo is itself a numerical solution technique analogous to numerical methods for solution of governing differential equations. The latter investigates a space or time grid non-randomly, and it is distinguished from stochastic propagation, such as MCS, that represents a similar, but random, utilization of analogous grids (Vose, 2000).

2.5.1 Historical Perspective of the Use of Monte Carlo Simulation

The beginnings of Monte Carlo Simulation (MCS) were initiated in some form at least several centuries ago. In 1768, the French mathematician Buffon conducted an experiment, repeatedly throwing needles onto a gridded field to estimate the value of π (Ross, 1976;Vose 2000). Lord Raleigh used the approach in a 1-dimensional, random walk procedure to approximate a solution of a parabolic differential equation (Rayleigh, 1899), and in 1908, the statistician Student (William Sealy Gossett) used the sampling procedure to confirm the theoretical derivation of the famous Student-t distribution (Vose, 2000).

The modern advent of use of MCS in problem solving is typically associated with the work of Ulam and von Neumann (1945) and Metropolis and Ulam (1949) in their efforts to assess mathematical equations used in predicting outcomes of nuclear fission (Burns, 2001; Vose, 2000; Rubenstien, 1981). The name for this numerical simulation technique was established at the Los Alamos National Laboratory in its use as a code word in the infamous Manhattan Project associated with development of the first atomic bomb. Monte Carlo, a Monaco city along the Mediterranean Sea, is well known for its association with games of chance (e.g., the roulette wheel, a physically based pseudo-random number generator). Many others (e.g., Kahn *et al.*, 1953) subsequently furthered the development of MCS that has, with the advent of computational technologies, led to today's ubiquitous use of the technique in the solution of many otherwise intractable mathematical equations.

Since the 1980's, due in part to access to faster computers, more complex problem statements, and an increased familiarity and comfort with probabilistic risk analysis, use of Monte Carlo Simulation has become more relied upon in environmental assessments. As one example of early regulatory use, MCS was allowed in several regulatory permit development exercises during the mid-1980's for evaluating a simple, receiving-water body dilution model. The purpose was to develop technically based permit effluent limitations for a prioritized group of industrial wastewater discharges subject to EPA's NPDES permit program. These industries were required to develop individual wastewater discharge control strategies, from which began the NPDES's toxics monitoring program. The MCS approach applied allowed an industry to essentially conduct a site-specific, probabilistic exposure assessment looking at the frequency of instream water column concentration excursions above various water quality criteria.

The basic MCS algorithm applied to a modeling solution is described in Figure 2-5. Figure 2-2 represents an example of model output construction. The purpose of the simulation experiment is presumed to be development of an adequate description of the model's output distribution derived from stochastic features of the model's inputs. Precision in OSE can be assessed in several ways, and typically requires some limited, initial output sampling to estimate the required number of runs or iterations (n_s) that would be needed.

Helton and Davis (2003) provide a useful summary of the reasons for the popularity of random Monte Carlo sampling techniques, and in particular Latin Hypercube sampling methods:

1. Computational simplicity and ease of implementation,
2. Dense stratification over the range of each sampled variable,
3. Direct provision of uncertainty analysis results without the use of surrogate models as approximations to the original model,
4. Availability of a variety of sensitivity analysis procedures, and
5. Effectiveness as a model verification procedure.

2.5.2 MCS Input Sampling Approach and Use of the Uniform Distribution

Summarized by Burns (2001) and many others (e.g., Vose, 2000; Cullen and Frey, 1999; Robert and Casella, 1999), Monte Carlo methods use computers to produce random numbers drawn from a uniform distribution that are then translated into the actual distributions of the stochastic input variables of interest. The approach was outlined by Metropolis and Ulam (1949; Burns 2001): "Once a uniformly distributed random set is available, sets with a prescribed probability distribution *f(x)* can be obtained from it by first drawing from a uniform uncorrelated distribution, and then using, instead of the number *x* which was drawn, another value *y=g(x)* where *g(x)* was computed in advance so that the values *y* possess the distribution *f(x)*." This is graphically portrayed in Figure 2-1. A random selection of the cumulative probability *u* is first made from the uniform distribution and an equivalent value *x* of the random variable X is set as the input for that quantity, via use of the inverse CDF (i.e., ICDF). This basic procedure is used in all random MCS-based sampling designs when sampling from non-uniform distributions.

Many texts summarize the various non-uniform probability distributions employed in describing a given random variable X. In some cases, intermediate distributions are first sampled (e.g., Normal, Gamma, and Geometric distributions) and their values then used to provide sample values for the target distribution (e.g., Lognormal, Beta, Binomial distributions respectively) (Vose, 2000).

2.5.3 Assessing Simulation-Induced Output Sampling Error (OSE)

Analogous to random sampling error in defining an input distribution, a similar problem is encountered in describing the output distribution of MCS experiments, referred to here as output sampling error (OSE). Because MCS, in practice, represents finite sampling of the output distribution for a given model quantity of interest, errors are introduced in describing the statistics of the true output distribution. The true distribution spoken of here is that collection of information that would result if the model were run infinitely, over and over again. Infinite sampling is, of course, supplanted at some earlier point by limited precision in machine-generated numbers.

Though the level of OSE for any given number of finite realizations is dependent upon total uncertainty in inputs, OSE is a separate aspect of uncertainty in model output that must be addressed. For a given output distribution, the accuracy of the estimates of its parameters (e.g., :, F^{2}, etc.) depends on the sample size n_s, but is independent of the actual numbers of model inputs n_x (Morgan and Henrion, 1990). The authors refer to this aspect as linear complexity in MCS: the effort (time) to run the model depends on n_x, but the sample size n_s (number of runs made) depends only on the desired level of accuracy in describing the output. It is the overall variance in inputs and their various relationships within the model that determines n_s.

In addition to simply monitoring the convergence of some desired output quantity as numerical simulation precision, for example, as a rate of change, there are two statistically-based techniques one can use in estimating the required sample size n_s for a desired level of accuracy in convergence of sampled output data. Described in the following sections, the first characterizes accuracy of the output distribution mean, and the second characterizes a given fractile of the distribution.

Confidence Interval Width for Normally Distributed Outputs

The first is straightforward and relies upon the basic outcome that large numbers of random model inputs will typically result in the output being a normally distributed random variable (Cullen and Frey, 1999). To address output sampling error, selection of sample size n_s can be established with a reasonable, initial estimate of the output distribution's true (population) variance. The estimate of the variance can be obtained by analysis of outputs for some smaller number of realizations n_{test} where $n_{test} < n_s$. From the estimated mean and variance of the interim output distribution formed by n_{test} model runs, n_s can be selected based on determining how many additional runs would be necessary to narrow the confidence level about the mean to a pre-specified interval, with some acceptable level of confidence (e.g., 95% confidence) (Cullen and Frey, 1999; Morgan and Henrion, 1990). The number of samples, n_s, is given by the following equation, derived from the confidence interval formula (Morgan and Henrion, 1990):

$$n_s > \left(\frac{2c\,\sigma}{w}\right)^2 \tag{2.1}$$

where:

- c = deviation enclosing the probability α of the standard normal distribution associated with the confidence level (1- α) of interest,
- $\sigma = \sigma_{\text{test}}$ = estimate of the standard deviation of the normal output distribution, and
- w = width of the desired confidence interval, in units of the variable of interest.

An example of the collapse of confidence intervals towards a target population percentile for increasing n_s is presented in Figure 2-6 for a normally distributed model output N(90,10) for up to 10,000 samples (i.e., 10,000 model runs). Three probability levels were evaluated, α = 0.05, 0.1, and 0.01, representing 50%, 90%, and 99% confidence intervals, respectively. Viewing N(90,10) as an example of a 3MRA "% protected sites" output value for a given C_w, a confidence limit can be used in specifying the lowest potential position of the value's true mean (e.g., for 3MRA, 99% confident that a given population percentile measure falls within ½ w units*)*. Discussed in Section 2.6.6, if the estimate of the given population percentile of a normal random output variable is subject to empirical uncertainty, this confidence would be stated about the 50th probability percentile of that estimate. Under conditions that all inputs represent "certain variability and/or certain constancy", in a national assessment, only the 50th "probability" percentile (of some given population percentile) has meaningful interpretation. Imprecise estimates of variability are attributed only to OSE. In this latter case, separation of output data values derived from groups of model runs, effectively constructing a "pseudo" 2-dimensional analysis, can be used to analogously construct a confidence interval about this estimate.

Bounded Interval For A Specific Fractile for Any Distribution Type

The second approach is based on establishing a confidence interval for whatever fractile (of a given population percentile) is of most concern in the risk assessment (Cullen and Frey, 1999; Morgan and Henrion, 1990). Here a confidence level is specified such that a specific fractile (i.e., probability percentile of a given population percentile) will be enclosed by the actual model estimates of two neighboring fractiles. A benefit is that the procedure can be employed for any distribution shape (Morgan and Henrion, 1990).

According to Morgan and Henrion (1990), and as also described by Cullen and Frey (1999), to obtain a specified confidence that the value of the pth fractile will be bounded by the ith and kth fractiles representing a common difference from the pth position, the following equations can be used to determine the needed sample size n_s:

$$i = n_s p - c\sqrt{n_s p(1-p)} \qquad \textbf{(2.2)}$$

$$k = n_s p + c\sqrt{n_s p(1-p)} \qquad \textbf{(2.3)}$$

$$n_s = p(1-p)\left(\frac{c}{\Delta p}\right)^2 \qquad \textbf{(2.4)}$$

where:

- *i* is rounded down, and *k* is rounded up to the nearest fractile
- n_s = model output sample size
- *p* = pth fractile of interest
- Δp = | %*p* - %*i* | = abs| %*p* - %*k* |

Important discussion by the authors underscores how the value of n_s that is the largest for a given level of confidence desired is for *p* = 0.5, indicating the median is the fractile most uncertain as measured by the finite sample values (i.e., *p*(1-*p*) = 0.25 is largest). This is not to say that the median is the most uncertain with respect to the value of the variable, which is often greatest at the distribution tails (Morgan and Henrion, 1990).

For 3MRA, this technique would be applied in the context of estimating impacts of OSE on CDF estimates of empirical uncertainty about a given population percentile (e.g., the 99% population percentile). This confidence interval would be stated about the p^{th} probability percentile of the estimate of the $99^{\%}$ population percentile. Following an example provided by Morgan and Henrion (1990), if we wish to be 95% confident that the true value of the 90th probability percentile, for the $99^{\%}$ population percentile, will be enclosed by the estimated values of the 89th and 91st probability percentiles, respectively, then c = 1.96 ≈ 2 (i.e., the area under the curve N(0,1) for (1- α), essentially a confidence interval on 1 sample), Δp = 0.01, p = 0.90, and n_s is calculated to be 3457 samples.

In 3MRA, 3457 "national realization" samples for a given chemical-WMU combination represents a much larger number of model runs needed. For a national assessment, 3MRA loops across all sites and C_W's, where there are currently 419 site-WMU combinations in the site-based database. To generate one exit level from one "national" cumulative distribution output curve as a measure of % sites protected across five C_W's, for all five WMU types and using all sites, it would take 2095 deterministic model runs, in essence representing a single, national sample for a single chemical. A national study then would entail 3457* (1 chemical)*(419 site-WMU combinations)*(5 C_W's) = 7,242,415 individual 3MRA model runs. Thus, if we use this procedure, we could establish the 91st probability percentile CDF as representing a conservative limit on meeting the 90th probability percentile, with 95% confidence, for a given population percentile of interest.

Again, under conditions that all inputs represent “certain variability or certain constancy”, in a national assessment, only the 50th “probability” percentile (of some given population percentile) has meaningful interpretation. In such a case, output data is continuously aggregated, eventually arriving at a precise description of the output population distribution, which embodies variability. In this latter case, separation of output data values derived from groups of model runs could be used to analogously construct a confidence interval about earlier estimates (e.g., using the actual 51st probability percentile CDF simulated to bound the true 50th probability percentile CDF).

Use of “pseudo dimensional analysis” for evaluating OSE, with or without the presence of empirical uncertainty (i.e., SME, ISE, ME), is further discussed in Section 2.6.6. Constraints in arriving at statements of confidence in any predictive uncertainty analysis include computational effort, and the total number of random numbers a computer can actually generate before it begins repeating itself (Vose, 2000).

2.5.4 Latin Hypercube Sampling

Essentially, LHS is a non-random process intended to emulate the outcome of a random process using fewer model runs, and is outlined in depth by Helton and Davis (2003).

LHS (McKay *et al.*, 1979) can be an extremely useful tool in probabilistic exposure and risk assessment since it exhibits desirable qualities for numerical solution similar to random Monte Carlo sampling. LHS uses a sampling technique known as stratified sampling without replacement, and is essentially a variance-reduction technique (Helton and Davis, 2003; Cullen and Frey, 1999). Similar to random Monte Carlo sampling, in Latin Hypercube Sampling (LHS) (Iman and Conover, 1980; Helton and Davis, 2000, 2003), each input distribution is divided into n_{LHS} intervals of equal, marginal probability $1/n_{LHS}$, where n_{LHS} is set as the number of simulations to be run, and one observation of each input quantity is made in each interval. The rank ordering of samples is typically random over the course of the simulation, where the pairing of samples between two or more random input variables is usually treated as independent (Cullen and Frey, 1999). The number of runs, n_{LHS}, is analogous to that which would be needed for random sampling (i.e., n_s) to describe a given output distribution (in a given dimension) of total uncertainty, though typically $n_{LHS << }n_s$. It can also be similarly used in sensitivity analysis.

Samples are drawn once, without replacement, from each of the n_{LHS} intervals, thus ensuring efficient representation of the entire range of the input distribution. Sampling within the intervals may be either entirely random or by explicit choice of the median of the interval, techniques known as “random-LHS” and “median LHS”, respectively. The latter is more efficient for mimicking a distribution shape with fewer runs, but under certain conditions imparts more difficulty in accepting the underlying statistical properties of the true distribution. Fundamental assumptions of LHS in attaining reliable estimates with fewer model runs are: (1) the expectation that only few inputs drive output variance; and (2) model behavior is monotonic and linear. In general, high degrees of non-linearity reduce the effectiveness of LHS (Morgan and Henrion, 1990). Meeting these assumptions of course is always a matter of degree. As a useful point of view in discussing the attributes of LHS, an underlying justification in using this approach is offered by Vose (2000):

> *"Monte Carlo sampling satisfies the purist's desire for an unadulterated random sampling method. It is useful if one is trying to get the model to imitate a random sampling from a population or for doing statistical experiments. However, the randomness of its sampling means that it will over- and under-sample from various parts of the distribution and cannot be relied upon to replicate the [joint] input distribution's shape unless a very large number of iterations is performed. For nearly all risk analysis modeling, the pure randomness of Monte Carlo sampling is not really relevant. We are almost always far more concerned that the model will reproduce the distributions that we have determined for its inputs."*

The latter part of this statement, of course, needs perspective as to what is "really relevant" in a given risk analysis objective, and a clarification of the limitations of LHS compared to random sampling. While providing unbiased estimates of means, variances, and distribution functions, random sampling can be computationally prohibitive in high order systems (Helton and Davis, 2000), and, as detailed in Section 2.5.3 above, can exhibit instability in convergence of output CDFs if n_s is small. Compared to random sampling in high order models (e.g., for a single site) that, for example, may require 1000's to 10,000's of samples (i.e., model runs) for a simulation experiment, LHS can often be completed using only 10's to 100's of samples (i.e., model runs) (Helton and Davis, 2000).

The complementary perspective is that LHS can impart bias in estimates of statistics of the output distribution, and this aspect is, as yet, not fully understood or well communicated by the risk analysis community. One reason for this is that determination of an adequate sample size to achieve reliability (in mimicking random sampling) is not directly amenable to the power provided by probability theory as that demonstrated for determining sufficient precision in random sampling. In many cases, the bias may be found to be insignificant (Helton and Davis, 2003). LHS is typically used when random sampling is computationally prohibitive and estimation of extremely high fractiles (e.g., 0.99, 0.999, 0.9999, etc.) is not required (Helton and Davis, 2000, 2003). In the latter case, a more subjective stratified sampling technique known as "importance sampling" can be employed though it can be difficult to apply in nontrivial problems (Helton and Davis, 2000, 2003).

LHS versus Random Monte Carlo Sampling

The following assertions/citations are offered to put in perspective the key issues regarding the benefits and limitations of LHS from the viewpoint of computational efficiency:

- A primary benefit of LHS is that the technique ensures each of the inputs is represented in a fully stratified manner, no matter which might turn out to be important (Campolongo *et al.*, 2000a; Helton and Davis, 2003).

- Desirable features of LHS include [comparatively] unbiased estimates of means and distribution functions [for small sample sizes] (McKay *et al.*, 1979; Helton and Davis, 2000, 2003), where it has been observed to be quite robust for small n_{LHS} (i.e., 50-200) (Helton and Davis, 2000).

- When the output is a monotonic function of its inputs, LHS is proven to be better than random sampling in describing the mean and the population distribution function (McKay *et al.*, 1979; Campolongo *et al.*, 2000a; Helton and Davis, 2003).

- Asymptotically, LHS is proven to be better than random sampling in that it provides an estimator (of the expectation of the output function) with lower variance. The closer the output function is to being additive (i.e. linear) in its input quantities, the greater is the reduction in variance. (Stein, 1987; Campolongo *et al.*, 2000a; Helton and Davis, 2003).

- Although LHS can sometimes still be more efficient, in cases dealing with non-additive and non-monotonic input functions, LHS has been shown to be equivalent to or worse than random sampling. (Stein, 1987; Campolongo *et al.*, 2000a).

- LHS is particularly practical/useful in dealing with the aspect of computational limitations in performing random sampling for long running models (e.g., single model runs of hours, days, etc.). If computational capacity is sufficient to handle random sampling, there is little reason to use LHS (Helton and Davis, 2003).

- Due to their computational complexity and expense, long-running models do not constitute convenient vehicles for comparing differences between random sampling and LHS (Helton and Davis, 2003).

In longer-term research plans for 3MRA, LHS will ultimately be employed. At the present time an LHS routine is though not functionally available in 3MRA Version 1.x, but is under development for eventual inclusion, to be based on the method of Iman and Conover (1982). It is currently functional in beta 3MRA Version 2.0 (see Figure 1-1). The lack of current capability in (potentially) taking advantage of the variance reduction capabilities of LHS for 3MRA Version 1.0 is currently offset by available computational capacity needed to reliably use random sampling. The benefits of having both eventually available include increased flexibility for more advanced 3MRA model evaluation strategies, and the ability to compare use of LHS and random sampling for complex, multimedia models like 3MRA. There are some available approaches for estimating precision of LHS analogous to discussions in Section 2.5.3 (Helton and Davis, 2003) that could also be investigated in evaluating LHS relative for 3MRA.

Of potential importance in evaluating the efficacy of LHS versus random sampling is the presence of periodic data in model inputs (Morgan and Henrion, 1990). This could, for example, be of concern in 3MRA where meteorological cycles are cyclically repeated for the period of record, over the length of the total simulation period, in lieu of actual longer-term data.

2.5.5 Dealing With Correlations Among Random Inputs

A common misperception about Monte Carlo Simulation is that it requires an assumption of independence between all random variables in the model, or stated alternatively, that all variables are uncorrelated (Cullen and Frey, 1999; Helton and Davis, 2003). Following from the discussion in Section 2.5.2, Metropolis and Ulam (1949) noted only that practical applications must, in the specification of the functions *g(x),* allow for covariance (correlations among input variables) to avoid physically impossible combinations of parameters (Burns, 2001).

Helton and Davis (2000, 2003) provide several points on aspects of correlation control:

- While it is important to address known correlations, it is equally important that variables do not appear to be correlated when they are really independent.

- It is often difficult to induce a desired correlation structure on a sample, where multivariate distributions can be incompatible with correlation patterns that are proposed for them.

- The rank correlation procedure of Iman and Conover (1982), a restricted pairing technique for controlling correlation structure in random MCS and LHS, has a number of desirable properties:

  - It is distribution free,
  - It is simple,
  - It maintains the original structure of the underlying sampling scheme, and
  - The marginal distributions remain intact.

- For many, if not most, uncertainty and sensitivity analysis problems, rank correlation (Iman and Conover, 1982) is probably a more natural measure of congruent model input behavior than is the more traditional sample correlation.

Rank correlation techniques are used in 3MRA, for example, in controlling correlation in regional hydrogeologic data.

2.6 Distinguishing Variability From Uncertainty

Uncertainty and variability have different ramifications for decision-makers in environmental risk assessments (NRC, 1994). The NRC underscored that “uncertainty forces decision-makers to judge how probable it is that risks will be over-estimated or under-estimated for every member of the exposed population, whereas variability forces them to cope with the certainty that different individuals will be subjected to risks both above and below any reference point one chooses”. Summarized by Small (EPA, 1996a), given knowledge of specific model inputs that impose appreciable output sensitivity, uncertainty may be reducible, the expense of which can be evaluated from a perspective of cost-benefit.

In assigning characteristics to an individual model input quantity in an uncertainty analysis, three general cases of total uncertainty (*constant and uncertain*, *variable and certain*, and *variable and uncertain*) can be distinguished, plus the case of a *constant and certain* input. Each case of total uncertainty presents a unique combination of attributes distinguishing a given model input's (1) variable nature, and (2) its uncertain nature, where the two natures can be viewed objectively to occupy different probability spaces (or dimensions of probability) (Helton and Davis, 2002, 2003). Figure 2-7 conceptually depicts the separation of the four cases for several inputs, where the reader is again referred to Table 2.7 for various terms used in describing variability and empirical uncertainty.

The variable nature of an input is referred to here as "variability uncertainty" or simply "variability". An input's uncertain nature is described by the term "empirical" uncertainty, or simply uncertainty; since it specifies the uncertainty we assign in quantifying a given, possibly variable, empirical input quantity. Variability is a characteristic of an input quantity that presents itself when we measure the quantity repeatedly at different points in space or time, or among individuals, and we observe that the variation in results is not reduced by carefully controlling the error in our measurement technique. Empirical uncertainty is discussed by Morgan and Henrion (1990). Type A and Type B uncertainties in Table 2-7 are defined by the International Atomic Energy Agency and are further described by the National Council on Radiation Protection and Measurements (IAEA, 1989; NCRP, 1996). Subjective uncertainty is discussed by Helton and Davis (2000), and objective and subjective uncertainty are more formally treated by Helton and Davis (2003), the latter work offering a succinct mathematical description for this comparative discussion. Other terms in Table 2-7 represent classical inferences in distinguishing variability and empirical uncertainty (Cullen and Frey, 1999).

2.6.1 Subjective Versus Objective Uncertainty

In its use to describe empirical uncertainty, the term "subjective" does not necessarily imply a completely subjective development of the input's probability distribution. For empirically derived inputs, initial analysis of original data leads to a partially objective specification of our "best guess" of the distribution describing variability. How this distribution is utilized for later describing attributes of a given individual or some (other) population of interest remains a separate matter (Section 2.1.1.). More intensive analysis can also determine, in a semi-objective manner, our best guess as to how wrong we may be in capturing the original data set as a representation of variability in the sampled population. In quantifying ISE, for example, what results in the end is a quantitative description of a family of frequency distributions that could be used to represent variability in the input, where a probability can be assigned as to how well each distribution represents the true population.

In summarizing these points, while possibly quite confusing in this linguistic landscape, the term "subjective" uncertainty is unabashedly accurate, since it is the aspect of a population (which could be a single entity; e.g., the speed of light), or the aspect of an individual in a set of entities, that we cannot currently describe with absolute certainty. Thus, "objective" uncertainty (i.e., variability) describes our best representation of the (sampled) population distribution, and "subjective" uncertainty (i.e., empirical uncertainty or uncertainty) describes the probability with which we believe this "best guess" to be actually true. Using the same "objective" description of a population, equivalently, subjective uncertainty would be implied in describing randomly selected individuals (or subsets of individuals) in that population.

Concern may arise when our best guess at the target population distribution is based on a limited number of samples in large populations, or when the variation in measured data is large (i.e., leading to greater input sampling error or ISE). Concern may also be noted when variation in data is small compared to error of the measurement device (i.e., sample measurement error or SME).

A given model or modeling system can have one or all four cases of total uncertainty present as elements x_i in the input vector **x,** where each x_i would be assigned to a specific case.

2.6.2 Sources of Variability and Uncertainty in 3MRA (Marin *et al.*, 1999)

The material following in this section was adapted directly from Appendix A of the 3MRA science methodology document of Marin *et al.* (1999) (see Appendix A of this Volume IV), and is wholly attributed to those authors. It provides an initial reference point in describing the various sources of variability and empirical uncertainty in 3MRA model inputs associated with the national assessment strategy. The presentation helps connect terminology defined in previous sections and that used by Marin *et al.* It is also useful in understanding materials presented in subsequent sections regarding techniques for separation of the influences upon model outputs derived from variability and empirical uncertainty in various model inputs.

One of the principal sources of variability in the 3MRA methodology framework is the variability of input parameters between sites. Example sources of variability include the between-site variability of the waste management characteristics such as area and volume, average spatial groundwater characteristics, climatic parameters, and number and type of receptors. Although spatial variability can also occur within sites, it is likely to be a significantly smaller contribution of the overall variability [expressed in population protection measures] than the between-site variability.

There are a number of sources that contribute to the [empirical] uncertainty in the prediction of the protective regulatory levels. These uncertainties can be generally classified as sampling (i.e., ISE) and non-sampling errors (e.g., ME and SME). Sampling errors arise because the number of samples (n) where a parameter is measured (sampled) is less than the number of sites in the population (N). The magnitude of the sampling error is a function of the variability of the parameter, the sample size n, and the population size (N). In general, the magnitude of the sampling error will be proportional to the variability and inversely proportional to the sample size. Non-sampling errors are generally independent of the sample size and are generally more difficult to estimate. Examples of non-sampling errors include measurement errors, simulation model errors (i.e. ME, not OSE), errors due to non-probability samples, improper problem statements, and errors due to sampling from non-target populations.

The input parameters for the proposed framework are used to define the modeling scenario for a facility [i.e., a site] and can be grouped into four general classes:

- *[3MRA Class 1 Inputs] - Variables that describe the characteristics of the waste management facility, including area and depth,*

- *[3MRA Class 2 Inputs] - Variables that describe the environmental conditions of the facility and its surroundings including hydrologic, hydrogeologic, meteorological, and geochemical conditions at the site,*

- *[3MRA Class 3 Inputs] - Variables that describe the (physiologic and behavioral) exposure and response characteristics of the receptors; and*

- *[3MRA Class 4 Inputs] - Variables that describe the physical, chemical, and biochemical properties of the chemical constituents.*

The first class of input parameters can exhibit variability and [empirical] uncertainty due to measurement errors and sampling errors. The second class of parameters can exhibit within and between-facility variability, and [empirical] uncertainty due to data measurement errors, sampling errors, and potentially errors due to the collection of non-probability samples. The third class of parameters can exhibit between-facility variability, between-individual-receptor variability, and [empirical] uncertainty due to sampling errors, measurement errors, and potentially errors due to the collection of non-probability samples, or non-representative samples. Finally, the fourth class of parameters is characterized by variability between batches, and [empirical] uncertainty due to sampling and measurement error.

There are also a number of prediction model error sources (i.e. ME) that would arise in the Monte Carlo simulation of the nationwide distributions of the protection measures, including: the mechanistic model prediction of the multimedia emission source terms from the WMU; the multimedia fate and transport modules that predict the media contaminant concentrations; the exposure models that predict the receptor dose; and the effect/response models that predict the receptor impacts. Additionally, there is the potential error of improperly stating the problem.

Referring to terminology used by Marin *et al.* (1999), examples of non-probability samples include samples of regional-based surrogate hydrogeologic parameters (e.g., aquifer thickness) and meteorological parameters that allow correlation structures to be established, and, similarly, for the national database, samples of surrogate environmental media characteristics (Marin *et al.*, 1999).

2.6.3 3MRA Model Input and Output for a Single National Realization

The discussions presented in following sections on the four different “cases” of total uncertainty for a given model input, introduced previously, is framed in context of the 3MRA site-based national assessment strategy. In classifying a given model input’s sources of variation in a simulation experiment, we define a scenario as a single deterministic run of the model at a specific site, described by some uniquely stated input vector $\mathbf{x} = [x_1, x_2,\ldots, x_{n_x}]$, where n_x represents the number of elements of **x** (e.g., the number of 3MRA model inputs).

The input space of any empirical quantity x_i in **x,** having either uncertain and/or variable natures in a national study, will be represented by either a single random variable (e.g., national), a set of random variables (e.g., regional or site-based), or a finite set of point measurements derived from any of the national, regional, or site-based data. In 3MRA, it is also possible to have different sites deriving data from different database levels. For a given scenario, values for 3MRA are selected from these various sets of data by the Site Definition Processor (SDP) to populate **x** for a given model run (see Section 4.3).

Initially, assume no input sampling error exists in the description of all random variables used to populate **x** for each of the 201 sites in the 3MRA site database. A single, national realization is defined by conducting a single Monte Carlo run at each site, for a specific WMU and chemical, and wastestream concentration. A general case of *variable and certain* inputs for some $x_{i,}$ and *uncertain and constant* inputs for other x_i, is established, where between-site variability in 3MRA output for this single, "national realization" would be derived from:

1. A set of regional-based constant point-estimates of x_i, each describing a group of sites
    a. (e.g., regression coefficient "a" for baseflow model).
2. A set of constant, point-estimates of x_i collected (measured) at different sites (e.g., number of water body networks across sites).
3. A set of different site-based random variables describing x_i (e.g., watershed flow length, where two or more distributions are used to describe all sites),
4. A national-based random variable x_i describing a certain property of the chemical
    a. (e.g., aerobic microbial degradation rate),
5. A national-based random variable x_i describing variability across all sites
    a. (e.g., human breathing rate), or
6. A set of regional-based random variables x_i, each describing a group of sites
    a. (e.g., spatially-averaged hydraulic conductivity).

In items 1 and 2 above, the associated set of constant values effectively represent a discrete frequency distribution of a given variable x_i, essentially a random variable describing the set of sites nationally, where correlation among all similarly measured x_i is inherently assigned. Item 3 above could represent characteristics of intra-site variability or empirical uncertainty. Item 4 would represent empirical uncertainty in determining a degradation rate (corrected for temperature) for the chemical, which would otherwise be interpreted to be constant for all site conditions modeled.

In items 5 and 6 above, the associated set of values sampled across sites effectively represents an instance of potential national variability across sampled sites, with some probability (>0 and <1) of occurrence, where individual values sampled for each site represent constant but uncertain values defined for a given site's input vector. For these items, the interpretation is referred again to Section 2.1.1, where, despite the lack of empirical uncertainty in variability represented by regional and national random variables, assigning values per site and model run imparts empirical uncertainty in describing conditions at each site (i.e., in describing the actual spatially-averaged hydraulic conductivity at site #0114001). This uncertainty would always be translated to outputs, regardless of issues relating to OSE. In

3MRA exit level processing, this is a matter of whether one correctly keeps this uncertainty separated, or incorrectly aggregates this uncertainty into the measure of variability across sites.

Interpretation of 3MRA Site-Based National Realizations

3MRA output is based on one or more deterministic runs of the modeling system. For the national assessment, a site-based analysis of 201 sites is formed from queries from the national, regional, and site-based 3MRA databases. The national assessment is constructed from repeated collections of potential outcomes across these 201 representative sites. In interpreting risk analysis results of the 3MRA national study, a cardinal rule of risk analysis modeling subscribed to here is well summarized by Vose (2000) who would imply that every 3MRA national realization must be a scenario that could physically occur. This distinction is quite important to the interpretation held here for output data generated by 3MRA for the national study. In summarizing this strategic point, we view that a single, national realization of 201 sites represents a potential outcome of future waste management conditions, nationally, with some probability of occurrence.

The aspect of national, site-based assessments, such as that discussed here for 3MRA, impose unique, practical challenges in assignment of model inputs to various cases of total uncertainty, and subsequent interpretation of modeling system output. This is because of the complexity normally imposed by site-specific studies, commingled with: (1) the aspect of rolling-up risk analyses across multiple, risk assessments of single-sites, all deriving data, in sometimes different fashion, from various scaled databases (i.e., site-based, regional, and national); and (2) the onus of evaluating how "variability" of the true national target population, is actually expressed within the site-based sampling design, model simulation design, and, ultimately, the problem statement. A fundamental aspect of interpretation of 3MRA model output is borne out of the idea that, in context of the model design and database construction, the true target population represents a collection of an infinite (or at least an extremely large) number of sites that would be needed to embody the entire potential of national variability. In reality, the decision-maker is faced with the perspective that over any time frame, only portions of this potential variability will actually be realized. It is this limited potential, as a statement of probability, that decisions of population protection is actually to be based upon.

Regional and national input distributions, despite the intuitive desire to think otherwise in a national assessment, actually represent distributions of uncertainty in assigning values to site-based studies (Section 2.1.1). Despite the fact that a "rolled-up" "national realization" of 201 sites captures a great deal of variability as uncertainty across many individuals, it is viewed here as incorrect to assert that there is 100% certainty that the "average outcome" will prevail, at all sites, in any given time frame. There is in fact only a 50% probability that this will occur.

As an example, consider to what degree the hydrogeologic conditions expressed in the sampled set of sites, if measured directly, would capture regional or national variability currently expressed in national and regional databases. To suggest that site-specific measurements at the 201 sites (or more importantly, some larger, but limited collection of sites the sampled site-based database represents) would reflect exactly the same national and regional distributions is untenable. Yet, given resources to collect all such data at the site-based scale, one would otherwise logically imply from the assessment strategy design a perspective that an "all site-based" data approach (better) reflects the true target population, short of ISE in the stratified site sample design. In essence, the sampled site-based database (or larger, but limited collection of sites it represents) itself represents a limited collection of randomly selected individuals of the true target population captured by national and regional databases.

A national perspective of the uncertainty imposed by use of the national and regional distributions for such realizable futures can be easily handled in a dimensional analysis (Section 2.6.6). This is in effect a separation of uncertainty imposed by non-target population risk analysis, laid upon site-based variability expressed in the site-based database. To separate this type of uncertainty from more familiar uses of dimensional analysis for addressing SME and ISE, we refer to the approach in this document as a pseudo 2nd-order analysis. OSE can also be readily addressed in this approach. Briefly discussed in Section 2.6.6, one may also conduct additional MCS-based experiments to further account for, and separate ISE and SME from either site-based variability, or total variability of the true target population. Such additional experiments, of course, require additional data to quantify these aspects of uncertainty in various random variable descriptions within national, regional, and site-based databases.

2.6.4 Four Cases of Uncertainty

In expanding upon the previously introduced concept, there are four possible combined "natures" (or Cases) of variance descriptions associated with any given 3MRA model input x_i, as shown in Figure 2-7. This taxonomy is discussed briefly by Marin *et al.*, (1999, 2003), and in more depth by Cullen and Frey (1999), EPA (1996a), and Vose (2000). Marin *et al.*, for example, provide example discussions dealing with the case of *variable and uncertain* data, where uncertainty is derived from ISE. The discussion here focuses on the 3MRA national assessment technology developed, and attempts to more fully characterize the various sources of input uncertainty. It provides additional context in conducting predictive uncertainty analysis for 3MRA, and arriving at an appropriate interpretation of model outputs generated by MCS.

Case 1: Constant and Certain

This case represents a condition where the input x_i is assumed to be quantified exactly the same for every 3MRA scenario, and there is no uncertainty represented in describing the value. Examples include defined constants (e.g., π), decision variables (e.g., risk protection criteria), value parameters (e.g., WMU life cycle), and certain model domain parameters (e.g., air model spline technique = ON, number of potential prey items = 20, etc.). These are represented in the example input vector **x**, as four variables labeled in Figure 2-7 as C_1 … C_4.

In the 3MRA national assessment, this characteristic is currently also associated with national-based point-estimates used for every site (i.e., constant empirical inputs represented in the national database). Most 3MRA Class 1 and 4 inputs (Section 2.6.2) would generally be included here, as currently represented in the database set. Examples include vapor pressure, showering frequency, shower water temperature, or leaf density (i.e., vapor pressure by chemical, BF = 1/day, HumRcpTemp = 43 C^{o}, rho_leaf = 770 g/L FW, respectively). At the present time in 3MRA, these values remain constant for all model runs, across all sites and all realizations.

Case 2: Constant and Uncertain

This case represents a condition where the input x_i is quantified exactly the same for every scenario, but there is uncertainty in describing the value used. In Figure 2-7, two such model inputs are shown as U_1 and U_2. These inputs have no component of variability, only uncertainty. In 3MRA national assessments, this characteristic would be generally associated with the national-based point-estimates used for every site noted in Case 1, once data were available to further quantify their uncertain component (e.g., SME). Based on additional data, assigning a *constant and uncertain* characteristic for the examples given for 3MRA in Case 1 would always be an interpretational, subjective construction. These examples would in fact likely be expected to retain components of variability as well, though their variable components are not currently quantified in 3MRA. An example of a more realistic candidate for this class in a national, site-based study with unlimited data availability would be uncertainty in the measurement of the speed of light. There is only one true value, but it cannot be exactly measured. An example of current parameterization of inputs in the 3MRA for this Case would be distributions used to characterize microbial degradation rates of organic substrates *in situ*, which are described by PDFs.

Indicated by Cullen and Frey (1999), this is a typical case assignment for probabilistic analyses, where all random variables are treated as uncertain quantities, without regard to distinctions between variability and empirical uncertainty.

Case 3: Variable and Certain

This case represents a condition where the input can be quantified differently for at least two site scenarios, but no empirical uncertainty is assumed present in the different values of x_i assigned for each site. Elements of the vector **x** could be deterministic or random variables, and all values or population distributions are presumed to be exactly known (in respect to describing statistics of the true target population). In Figure 2-7, 3 such model inputs are shown as V_1, V_2, and V_3. These inputs have no component of uncertainty, only variability. In describing the sample population of sites for 3MRA, a specific site would always have the same value assigned if derived from measured point estimates (being different for at least some sites) or where a different value is randomly assigned from either the national, regional, or site-based databases. Some examples of this Case were previously delineated in Section 2.6.3.

In reality, for random variables, empirical uncertainty (in the true target population) will be present at some level (i.e., ISE, SME, etc.). In 3MRA, the level of empirical error would, desirably, be presumed for the time being to be far less dominant than the variability otherwise expressed in the distribution or set of point estimates currently being used to describe all sites. The importance of (accepting or not accepting) this assumption on the accuracy of model output interpretation can be evaluated more objectively through sensitivity analysis. The question of interest is “How accurate is this assumption, and does it even matter (i.e., sensitivity)?”. Cullen and Frey (1999) point out this case (*variable and certain*) rarely exists in practice. For deterministic, empirical quantities varying across sites, a reasonable example of variability without uncertainty, under conditions of unlimited data availability, might be the number of water body networks at a site for a given spatial resolution of land use/cover data.

ISE aside, in a 1-dimensional analysis of total, hybrid uncertainty, all “static” distributions in the 3MRA database (i.e., all random variables or point estimates that vary from site to site in the 3MRA database) would effectively represent a basic assignment to this Case, which would impart a belief that these “frequency” distributions (of the true target population) represent a best estimate of the true variability describing the input quantity. For empirically derived distributions truly representing the target population, an uncertainty component (e.g., SME) would still obviously be associated with these variables, and as such they would, if objective data were available to separate SME, be more appropriately handled in Case 4, using a segregated 2-dimensional (or 2nd-order) analysis approach (Cullen and Frey, 1999). In the case of non-target sampled populations, the aspect of “individual” versus “population” is in play (Sections 2.1.1 and 2.6.3), and such variables (e.g., 3MRA Classes 2 & 3, Section 2.6.2) would be more appropriately assigned to Case 2, separated from sampled site-based data representing *variable and certain* quantities using a “pseudo” 2nd-order analysis. Presumed to be actually present in the latter quantities, the dimension of variability would represent a hybrid dimension of site-based variability convolved with SME.

Regarding SME, for all 3MRA empirical model inputs currently described by PDFs, effectively assigned here in a 1-dimensional analysis, uncertainty due to random error (RE) is, at the present time, convolved with variability in these descriptions. Thus, any simulation design expressing these model inputs as pure variability would ignore some level of uncertainty present in the data. For the most part, all variables currently described in 3MRA databases represent a hybrid probability space of total uncertainty (i.e., variability plus at least SME uncertainty, and, for 3MRA Class 2, some uncertainty due to sampling of non-target populations). Alternatively, the aggregated SME and variability could also be more conservatively assigned to Case 2 (in analysis of the true target population), and handled via a pseudo 2nd-order analysis, imparting an interim assumption that SME dominates the variance in these descriptions.

Case 4: Variable and Uncertain

In real systems, this is the most prevalent case for most empirical quantities in an environmental exposure and risk assessment, whether or not the actual component of uncertainty is characterized. Here we have the same description as in Case 3, however, like Case 2, uncertainty is now introduced into either the random variable or point-estimate descriptions of an empirical quantity that varies from site to site. In Figure 2-7, seven such model inputs are shown, UV_1…. UV_7. These inputs have both a component of uncertainty, for example $U(UV_6)$, and a component of variability, $V(UV_6)$. This is more properly quantified by assigning a family of frequency distributions for ISE uncertainty, and by separately characterizing SME through knowledge of precision and accuracy in measurement techniques used to collect the original data that a given, single distribution was actually developed from. These approaches would be handled by a general 2nd-order analysis.

2.6.5 Four Cases of Uncertainty for 3MRA Input Descriptions

Revisiting NRC's (1994) comment (Section 2.6), the goal is to provide the best information possible to the decision-maker, indicating that part of uncertainty about a decision that cannot be reduced, and that which can be, in principle, reduced. Of practical focus in such an effort would be the identification of key (sensitive) model inputs that have large components of reducible uncertainty. The ability to separate uncertain and variable natures of an input x_i is predicated on the existence of available data to first distinguish these components in the model input, and further, in having available a simulation capability to separate effects upon model output (e.g., 2nd-order analysis; Section 2.6.6).

Sample Measurement Error (SME)

All measured data have some random measurement error (i.e., SME) that is in many cases normally distributed, and may also be correlated with variability. The resulting data collection outcome is seen as increased variance in sample data and some offset from the true population mean due to systematic error present. In this particular instance, a 2-dimensional analysis can be conducted to assess impacts of SME on model output, in effect separating variability from uncertainty (Cullen and Frey, 1999; Vose, 2000; EPA, 1996a). Here, additional objective data (quantification of precision and accuracy of the measurement technique) can be employed in the predictive uncertainty analysis to first separate SME from variability, and then assess the SME uncertainty from variability through 2-dimensional analysis (Frey and Rhodes, 1996).

For many 3MRA inputs, measurement error is largely mitigated by accurate and precise instrumentation and the dominance of variability in the joint distribution. In the general case, it may be that: (1) the uncertainty due to SME is negligible with respect to variability; (2) SME is significant but the model output to it is insensitive; or (3) SME is significant and model output is sensitive to it. The last situation is, of course, the focus of a useful 2-dimensional analysis that would allow separation of uncertainty and variability to assess whether or not uncertainty might be reduced, and to what benefit in the risk analysis. For SME, it is paramount to understand that distributions in 3MRA (in all likelihood) include this uncertainty. It is currently not separated from variability explicitly, and the term total uncertainty is the most appropriate designation.

Input Sampling Error (ISE)

For random variables, depending on the sample size of data used to create its distribution, random input sampling error (ISE) may be significant. Here, a two-dimensional analysis can be conducted to address ISE using a family of frequency distributions to describe the input variable. An example discussion with graphical conceptualization of a single, normally distributed model input, with uncertainty in the distribution's mean value is shown in Appendix A of Marin *et al.* (1999) (see Appendix A of this Volume IV). This example generally represents an analysis of the true target population (see Section 2.6.3) and assumes sufficient sample size to minimize OSE. An additional source of ISE is associated with the ability to properly select the most appropriate distribution function type (i.e., is it normally or log-normally distributed). For ISE, distributions in 3MRA, all currently of the static form given by a single distribution, do not currently address these uncertainties, implicitly or explicitly.

The degree of ISE uncertainty is a function of the number of samples used to construct the distribution, the variability present in the population measured, and the experience of the researcher who constructed it. For inputs constructed from thousands of measurements, ISE is likely to be small compared to influences of variability on output. As Vose (2000) shows in sampling experiments with known distributions, sample sets with only 10's to 100's of measurements can impart a significant level of ISE. Various bootstrap sampling techniques can be used with the original data set to quantify an appropriate family of frequency distributions to characterize ISE (Marin *et al.*, 1999; Cullen and Frey, 1999; Vose, 2000). Given a data set of size *n* point estimates, the general approach in bootstrap simulation is to assume a distribution which describes the quantity of interest, perform *r* replications of the data set by randomly drawing, with replacement, *n* values, and then calculate r values of the statistic of interest (Cullen and Frey, 1999). It is essentially a numerical method.

Assignment of Cases for Current 3MRA Static Distributions

Empirical uncertainty in "national" variability is then an aspect of how well we have actually described a given variable x_i for the true target population. In reality, all examples in Section 2.6.4 retain some form of empirical uncertainty, which may or may not be captured in the current distribution(s) specified. At the present time, the 3MRA databases do not have available quantitative descriptions for further characterization of ISE for any given distribution (e.g., a family of frequency distributions representing ISE). Thus, each $x_{i,}$ across all sites, is currently categorized by: (*A*) a single constant; (*B*) a set of constants; (*C*) a single distribution, or (*D*) a set of distributions where each distribution in the set is uniquely assigned to some site or set of sites, where different sites can derive data from different categories *A*, *B*, *C*, or *D*.

The current set of random variables describing various input quantities in 3MRA can be viewed to embody the best available information describing variability across the sampled sites, or target populations represented by the regional and national databases. The combined "natures" of category *A* would be defined as *constant and certain*, and the remaining three categories (*B*, *C*, and *D*) would be, for current 3MRA data sets, exclusive of microbial degradation rates for chemical properties, defined as best estimates of *variable and certain* quantities. Chemical properties defined by PDFs (e.g., microbial degradation rates) would be defined as best estimates of constant and uncertain quantities (i.e., ignoring ISE).

Generally, the national variability captured by the 3MRA databases is implied at this point to be appreciably greater than sample measurement errors, or errors introduced by random sampling. Informed by sensitivity analysis, these assumptions can be reflected upon after assessing the inputs' importance in driving critical output quantities of interest (i.e., through sensitivity of the national protection measure). In the interim, even in describing the "true target population", one can also assume a more conservative position that some random variables described by "frequency" distribution functions (categories *C* and *D* directly above) actually represent uncertainty in a constant value (i.e., derived from SME). Under the condition that a single x_i as such drives model outputs, and is dominated by empirical uncertainty, such an approach, of course, becomes less conservative, and more appropriate.

2.6.6 Two-Dimensional Analysis of Uncertainty

Uncertainty and variability are described by distributions that, to all intents and purposes, look and behave exactly the same (Vose, 2000). Vose offers a basic premise held in risk analysis that variability is a fundamental basis of a risk analysis, and that uncertainty about parameter values should be overlaid onto that model of variability. Attempts to separate the characteristics of uncertainty and variability in model output are best placed in this context, where a "best estimate" of variability is distinguished, upon which additional uncertainty can be viewed.

A general approach to separation of variability and uncertainty is outlined by Cullen and Frey (1999), Bogen and Spear (1987), Bogen (1995), and Vose (2000), and is also presented in detail in the science methodology for 3MRA (Marin *et al.*, 1999, 2003). This approach can be applied to the general case where both empirical uncertainty and variability are commingled in the model output when either: (1) one or more model inputs are both variable and uncertain, as in Case 4; or (2) at least two model inputs are employed where one is variable and certain, and another is constant and uncertain. Thus, 2nd-order analysis is employed when some inputs are uncertain (which are also distinguished as being either variable or constant), and some inputs are variable or constant with no uncertainty.

As discussed in Marin *et al.* (1999, 2003), this technique of "nesting" or "double looping" (Cullen and Frey, 1999) can also be employed to separately propagate model error (ME), again relative to the reference point of the best estimate of the true model (i.e., 3MRA). In the latter case, as discussed previously, parameterization sometimes also includes model error. Thus, treatment of ME in 2nd-order analysis may be constrained by any inability to disaggregate model error from compositional uncertainty. Model error would generally be handled by imposing statistical uncertainty on model outputs, which could be done on a module-by-module basis in an integrated system such as 3MRA.

Depending on one's desire to separate various sources of uncertainty, controlled experiments can be conducted to evaluate each source of uncertainty separately, essentially expanding the dimensionality of the analysis.

General Form of 2nd-order Analysis Algorithm for Risk Analysis

Adapting terminology from Bogen and Spear (1987) and method statement from Cullen and Frey (1999), we define the risk model by $R = R(U, V) = f(U, V)$, where the estimate of risk to the population of exposed individuals is a function of variability in model inputs in **V** (with no empirical uncertainty), and uncertainty in model inputs **U**. For our purposes, elements of V, v_i, may be constant and certain or variable and certain, and elements of **U**, u_i, may be constant and uncertain or variable and uncertain. The basic premise is to separate variables into two groups, uncertain and certain, and then to conduct nested simulations in a set of experiments designed to preserve separately their effects on the model output. In so doing, probability distributions (along the dimension of uncertainty) are overlaid on the base CDF representing population variability (i.e., along the dimension of variability).

A general form of a 2nd-order analysis (i.e. in handling ISE) is given by Cullen and Frey (1999) and is stated by the following algorithm, depicted in Figures 2-8 and 2-9:

Step 1: Disaggregate model inputs into variable **V** and uncertain **U** components.

Step 2: For the M input quantities in **V**, frequency distributions are specified

Step 3: For the N uncertain quantities in **U**, probability distributions are specified.

Step 4: Using simulation (i.e., Monte Carlo, LHS, etc.), generate two sets of samples:
For each of the M variable quantities, generate *m* samples.
For each of the N variable quantities, generate *n* samples (i.e., n_s).

Step 5: Evaluate the model deterministically for each combination of sample sets.

The total number of model runs is *n* * *m*. The basic approach for managing output data, as shown in Figure 2-8, is to segregate data by realizations of *n*, generating *n* estimates of the rank, or percentile, representing a probability distribution for the rank of the individuals in the population (Cullen and Frey, 1999). Thus, each column R_{ij} for j = 1…*n* in Figure 2-8 represents an estimate of variability for a given realization of uncertainties, and each row R_{ij} for i = 1…*m* represents an estimate of uncertainty for a given member of the population. Using Figure 2-7 as an example, **V** would be constructed as **U** = [C_1, C_2, C_3, C_4, V_1, V_2, V_3], and **U** would be constructed as **U** = [U_1, U_2, UV_1, UV_2, …, UV_6, UV_7]. The model could, in principle, also be reformulated employing error terms directly as additional inputs, where **V** would be constructed as **U** = [C_1, C_2, C_3, C_4, V_1, V_2, V_3, $V(UV_1)$, …, $V(UV_7)$], and **U** would be constructed as **U** = [U_1, U_2, UV_1, UV_2, …, UV_6, UV_7, $U(UV_1)$, …, $U(UV_7)$].

From a practical standpoint in use for 3MRA, the general two-dimensional analysis could be implemented separately to address SME, ISE, and ME, employed at different levels of the model. As a general example for ISE, and as indicated by Marin *et al.* (1999, 2003) for the 3MRA methodology in Appendix A and Figure 3.6 of that document, **U** would be placed in the outer loop of deterministic model runs, where for each set of *m* samples of variability, a different estimate of a distribution's parameters would be used. Thus, *n* sets of distributions (i.e., *n* distributions for each uncertain quantity) could be used to populate the *n* samples of **U**, with each value u_i representing a single sample from a single distribution from its family.

Marin *et al.* (1999; see Appendix A) provide a useful summary and example of the impacts of addressing or not addressing ISE under the assumption of sampling from distributions representing the true target population. They cover the cases of: (1) 1-dimensional variability assessments ignoring ISE (i.e. V); (2) 2nd-order analysis incorporating ISE but aggregating dimensions of variability and uncertainty due to ISE (i.e. U+V); and (3) 2nd-order analysis separating dimensions of variability and uncertainty due to ISE (i.e. U|V which is analogous to the UV notation used in Figure 2-7 and discussions directly above). In the first case, lack of address of ISE will lead to over-estimating the true level of population protection for all population percentiles. In the second case of combining U+V in a hybrid dimension of total uncertainty, such analysis leads to over-estimating the true level of population protection for high ends of the population (e.g., population percentiles > 50%), and underestimating it for low ends.

Pseudo 2nd-order Analysis Algorithm for Risk Analysis

The term pseudo 2nd-order analysis is used here to refer to the separation of uncertainty associated with *constant and uncertain* inputs from inputs characterized as *variable and certain*. It is for all practical purposes, a 2-dimensional analysis of the general form. In the case of separating *constant and uncertain* inputs from inputs characterized as *constant and certain*, the approach would simply collapse to a standard, familiar concept of 1-dimensional uncertainty analysis in risk assessment. This technique could be used for example to assess aspects of sampling from frequency distributions representing non-target populations described in Section 2.6.3, where uncertainty is imposed in describing values for individuals of the population. It could also generically be applied concurrently in dealing with site-based, regional or national based *uncertain and constant* inputs such as those associated with random error (RE) portions of SME for input quantities (e.g., microbial degradation rates in 3MRA).

It is conceptually the same as the general form of the 2nd-order analysis technique described directly above. The basic approach for managing output data, as shown in Figure 2-10, is to segregate data by realizations of *n*, generating *n* estimates of the rank, or percentile, representing a probability distribution for the rank of the individuals in the population (Cullen and Frey, 1999). Following Figure 2-8, one is in essence taking *n* snapshots of the frequency distribution of the population, and the various percentiles of interest of that population (representing probability distributions of randomly selected individuals). In the pseudo 2nd-order analysis approach, analogously, the outer loop in Figure 2-9 simply entails the separation of output data, and forgoes any need to regenerate estimates of uncertain distribution parameters. For 3MRA, each iteration would be formed as an aggregated sample of receptor risks based on one loop through the national list of sites in the 3MRA site-based database. Described in Section 2.6.3, each iteration (i.e., realization) would be interpreted to represent a potential outcome of risk for the national population of receptors living within and around Subtitle D waste management units.

Output Sampling Error (OSE)

For the general form of 2nd-order analysis for *variable and uncertain* inputs, *m* must be large enough to sufficiently minimize OSE in the variability CDF, for each iteration in the outer loop. This case of sampling from variability distributions representing the true target population is described in Marin *et al.* (1999). To minimize OSE, Marin *et al.* (1999) imply that the bootstrap sample (*m*) used for the inner loop is sufficiently large enough to minimize OSE of the variability distribution, effectively cycling through multiple loops of the national site list. If not, one would also need to address OSE through a confidence interval on the mean (i.e., describing confidence on each instance of the dimension of variability) or some alternative 50th percentile of probability (i.e., describing confidence on the dimension of variability per iteration) (Section 2.5.3).

For the pseudo 2nd-order analysis technique in Figure 2-10, estimates of sample size needed to sufficiently reduce OSE are similarly described in Section 2.5.3. Here though, one is evaluating *n* in terms of needed sample size to converge to reliable estimates of the variability CDF, or some *p*th percentile of probability.

Discussed in Section 2.5.3, under conditions that all inputs represent "certain variability and/or certain constancy", in a national assessment, where sampling is conducted from true target populations, only the 50th "probability" percentile (of some given population percentile) has meaningful interpretation. Imprecise estimates of variability are attributed only to OSE. Note that in such a case, receptor risk data is constantly aggregated across multiple national realizations. In this case, separation of output data values derived from groups of model runs, effectively constructing a pseudo 2nd-order analysis, for example by national realization, can be used to analogously construct a confidence interval about this estimate.

2.6.7 Classification of 3MRA Inputs and Interpretation of Outputs

In predictive uncertainty analysis with 3MRA using Monte Carlo Simulation, the variance in protection measure as an output is captured succinctly as variability in a one-dimensional uncertainty analysis for Case 1 and Case 3 (Section 2.6.4), if in Case 3, the OSE due to simulation error is minimized. The need to address, and possibly separate uncertainty and variability arises in Case 2 and Case 4, is formed on several arguments: (1) the situation where a given input description in 3MRA commingles, or embeds, uncertainty and variability in the description of the input, as is the case for SME, and where we further demand an ability to describe the separate effects upon output in an eventual scheme to reduce SME; (2) the need to address the uncertainty in site-based variability due to sampling from regional and national distributions describing non-target populations; or (3) the need to better address any ISE uncertainty present (i.e., evaluating U|V, as opposed to U+V, or V).

Hybrid Dimension of Total Uncertainty

As an example, assuming infinite sampling data were available on measurements used to develop a distribution of leachate pH on a national basis, SME in the resultant distribution describing this population would incorporate both variability and SME uncertainty. Here the variance in the model input captures both features, and this variance is propagated in the model output. Without separation, it represents an expression of the hybrid total uncertainty (i.e. U+V where the vertical axis of output variation is neither uncertainty nor variability; Vose (2000)). Using data on the precision and accuracy of the measurement instrument, these features can be separated in the sample data set, and subsequently, a two-dimensional analysis can be used. If ISE associated families of frequency distributions were available, and were incorporated within a 1-dimensional analysis, a similar situation would arise where the dimension would be that of total uncertainty.

From the viewpoint of the data collector, analyst, and decision-maker, it is often unclear to what degree a random variable represents uncertainty in an otherwise constant entity, or a "natural" variability of this quantity in space or time, or both. Until the data collector experimentally separates these entities, and presents the information quantitatively, from the perspective of the decision-maker, the "variability" distribution should be viewed to represent both uncertainty and variability. We often have reason and experience to assert that our data strongly represents variability, based in large part on expert knowledge of the accuracy and precision of our measurement techniques. Nonetheless, the dimension of 1-dimensional analysis is that of total uncertainty, where either variability dominates, uncertainty dominates, or both significantly exert their natures upon the underlying empirical data set.

Consider a hypothetical example of bathroom fan airflow rates. If stated as a national random variable, it may be the case, unbeknownst to us, that all fans in every bathroom across the country were made at the same fan plant in Fan City, Idaho, and the plant only makes one type of fan. Assume the distribution was constructed from a single measurement of every fan in the country collected on the same day with the same instrument. Here, the random variable may actually represent: (1) variation in the manufacturing process of the fan; (2) uncertainty in how fan rates were measured; or (3) variable effects of aging in the fan motors over time. Only our expert knowledge, if actually known to us, that the fans "should" all be about the same allows us to reflect upon the data set as a representation of a constant and uncertain quantity. Without additional analysis of SME, we may incorrectly view this condition as an aspect of variability, guided by unchecked intuition that not all fans are made the same.

On Separating the Hybrid Dimension of Total Uncertainty

Separation of total uncertainty into its variable and uncertain components is predicated on having additional knowledge about the raw data and what it represents. Without additional data, or more extensive examination of raw data, a 2nd-order analysis is highly subjective, and represents, in effect, a type of sensitivity analysis, for which there are better ways to approach the problem. Assuming the quantity's true natures of uncertainty and variability are unknown, the analyst has little choice other than to investigate these spaces jointly, and that, in turn, imposes a certain treatment in interpreting its effect on model output. It is fair to conclude that every random variable currently presented within the 3MRA databases has some element of uncertainty and variability embedded in its distribution. Finally, unless large quantities of raw data are available, most experts would recommend against arbitrary function type assignment not consistent with typically applied function type(s) for this quantity class, in essence imparting expert opinion based on surrogate data in these cases.

Population Vs. Individual Risk

In keeping with the terminology discussed in this work, we revisit briefly discussions on the nature of individuals versus populations previously presented in Sections 2.1.1 and 2.6.3, from the view of interpreting properly constructed model outputs.

For purposes of understanding the 3MRA modeling system, we typically assign definitions of frequency and probability in describing variability and uncertainty, respectively, as they impact model outputs that describe total populations or individuals of interest. We underscore this theme again for its importance in interpreting 3MRA model outputs in the context of a national, site-based assessment methodology. Assuming random samples are selected from the population, the variance expressed in model output due to stochastic model inputs is deemed a complete expression of risk frequency in the target population (i.e., variability of the population; e.g., the % sites with 90% population protection at a given C_w) if and only if it reflects no significant influences of empirical uncertainty. For interpretation of individual risk, the same model output under these assumptions represents an expression of the probability (i.e., uncertainty) of randomly selecting a specific site having a specific risk (e.g., the probability of 90% population protection at a given C_w).

Cullen and Frey (1999) further summarize important points regarding individuals:

> *"In cases where the uncertainties are independent from one individual to another, the rank ordering of individuals is also uncertain. …… In contrast, one-dimensional simulation approaches do not capture these interactions. The resulting hybrid-frequency/probability distribution for exposure [risk] is only meaningful if interpreted to represent an individual selected at random from the population, or if only variability or uncertainty dominates. Otherwise, it is inaccurate to draw conclusions from such results regarding the rank ordering of individuals within the population, or the exposure [risk] level faced by an individual at a given fractile of the population."*

In coming to terms with our true limitations to separate the two in real systems, it should always be acknowledged in final analysis that variability and uncertainty are to some degree inextricably intertwined (Burns, 2001).

2.7 Taxonomy of Sensitivity Analysis

Restated, sensitivity analysis (SA) is the study of how the uncertainty in output of an analytical or numerical model can be apportioned to different sources of uncertainty in the model input (Saltelli, 2002a). It is also, from a quite practical point, a powerful set of tools to be used to further verify model structure, and overall compositional validation of the model. Its use as a measure of model input importance enables identification of the critical areas of lack of knowledge and data. Hopefully this leads to subsequent reduction of uncertainties in model output, through additional observations of the system under study, keyed on sensitive inputs with large uncertainties. The desired outcome of sensitivity analysis, therefore, is to support model evaluation conclusions on quality assurance, and to point decision-makers to areas of further data collection that could substantially reduce uncertainty in model outputs.

Presenting a more directed discourse on the subject of sensitivity analysis for high order modeling systems, the discussion herein outlines elements of a balanced, tiered formulation to an approach for 3MRA. Actual design and implementation of a sensitivity analysis plan for 3MRA are discussed in more detail in Section 9; the text immediately following attempts to outline the critical considerations in the use of the techniques to be employed.

The basic approach to be undertaken for 3MRA, one of global input space assessment, entails use of sampling-based correlation and regression methods, complimented by procedures employing an integrated regional and tree-structured methodology. The intent over-time is to further evaluate efficacy of simpler screening methods, as well as fully quantitative, variance-based global methods of sensitivity analysis. This activity would be constructed within the specific context of the model evaluation problem, and tasking to be undertaken to serve both short-term and long-term needs in investigating 3MRA input sensitivity.

The overview to follow outlines the basic elements possible in a general approach for 3MRA, and distinguishes their intended benefits and limitations in the model evaluation process.

2.7.1 Classification of Sensitivity Analysis Methods

Three general classes of sensitivity analysis techniques can be defined (Campolongo *et al.*, 2000a). These are graphically conceptualized in Figure 2-11:

- Factor screening methods (e.g., one-at-a-time (OAT) methods, factorial designs),
- Local methods (e.g., differential or nominal value analysis), and
- Global methods (e.g., MCS-based correlation/regression and RSA methods, FAST, Sobol's Method, etc.; see Sections 2.7.4, 2.7.5 and 2.7.6 for definitions)

Works presented within Saltelli *et al.* (2000), provide a well-rounded survey of available sensitivity analysis techniques, apportioned among the conceptual levels of screening, local, and global algorithms, and are briefly summarized in the following overview. In executing a sensitivity analysis plan for 3MRA, key surveys illuminating additional detail and applicability of these techniques are also found in recent works of Campolongo and Saltelli (1997), Campolongo *et al.* (1999), Kleijnen and Helton (1999a, 1999b), Helton and Davis (2002, 2003), Frey and Patil (2002), and Saltelli (2002a, 2002b). Together these works were relied upon to plan, and will be relied upon to execute and analyze, experimental studies to elucidate 3MRA sensitivity.

As a matter of course, the likelihood of a successful sensitivity analysis (the appropriate identification of key and redundant model inputs) is conditioned upon the use of more than one technique. A tiered analysis approach is a common theme for good practice (Helton and Davis, 2002; Kleijnen and Helton, 1999a, 1999b; Saltelli *et al.*, 2000; Campolongo *et al.*, 1999; Campolongo and Saltelli, 1997; NRC, 1994; EPA, 1996a, 1997a). This, for example, was a primary recommendation of Small and others in their development of EPA guidance on the use of MCS-based approaches for sensitivity analysis (EPA, 1996a). With a tiered strategy, a sensitive relationship missed with one approach can often be identified using another (Kleijnen and Helton, 1999a). Typically, the problem of how best to construct a tiered strategy is guided by computational cost (Campolongo *et al.*, 1999; Saltelli, 2002b; Saltelli *et al.*, 2000).

2.7.2 Overview of Screening, Local, and Global Approaches

Summarized by Campolongo *et al.* (2000a), screening methods are typically qualitative, providing only a ranking of the importance of various x_i in $\mathbf{x}$. Screening methods can usually be further characterized as retaining properties of either local or global methods, and are by design the most computationally efficient. Some screening methods allow for only univariate assessment (e.g., typical OAT – i.e., one-at-a-time methods, such as the 1st order methods of Morris or Cotter) while others allow for assessment of factor interaction (e.g., factorial designs).

Local methods tend to address only a specific point, or local region, in the input parameter space (e.g., sensitivity of $\mathbf{x}_{\mathrm{nominal}}$). Local and global methods are typically quantitative in their representation of the input-output relationship. Local methods usually embody a univariate assessment framework among x_i, based on partial derivatives. Here, all x_j are kept constant at their nominal value (i.e., central tendency) for all j <> i, while x_i is varied near its nominal value. For local methods, the interval range taken about the nominal value is characteristically small, and is usually uniform for all factors analyzed. Examples include the brute force finite-difference approach or direct methods based on differentiation. Referenced in Figure 2-11, local differential-based methods typically assume linearity between input and output variables, and are used commonly in inverse problem formulations applied during model calibration as a parameter estimation technique. Doherty (2002a, 2002b) is more recently pursuing local method research and technology development of non-linear solution capabilities for environmental fate and transport modeling problems.

Described by Campolongo *et al.* (2000a), global methods apportion the output uncertainty to the uncertainty in the input factors, typically given by probability distribution functions that cover the factors' entire ranges of existence. Global methods are far more computationally demanding and involve various methods of sampling the input factor space (e.g., LHS, random sampling, etc.) (Helton and Davis, 2002, 2003). Most global methods will be referred to as either Monte Carlo-based regression-correlation methods, or variance-based methods, and, in contrast to local methods, usually represent a multivariate assessment of model sensitivity.

Fully quantitative global methods account for four key properties (Campolongo *et al.* 2000a; Saltelli, 2002a):

1. The inclusion of the influence of scale and shape *(i.e., they incorporate range and shape of the PDFs of each factor)*,

2. Multidimensional averaging *(i.e., addresses parameter interaction where individual factors are evaluated by varying all other factors as well)*,

3. Model independence *(the method should work regardless of the additivity or linearity of the test model)*, and

4. Treatment of groups of input factors as a single factor *(A property of synthesis of analysis results essential for agility in their interpretation; in essence an ability to distill and then communicate critical sensitivity analysis results).*

A key consideration, the presence of model non-linearity or varying degrees of the magnitudes of uncertainty across input factors, can render local methods, regression-based methods, and most screening methods undesirable. Shown by many researchers, additive effects from input factors upon output represent a difficult assumption to make under such conditions.

Though computationally the most expensive, fully quantitative, variance-based global methods are robust to model non-monotonicity. Some screening-level factorial designs (e.g., Iterated FFD; see Section 2.7.3) can also perform well under such conditions (Campolongo *et al.*, 2000b).

Differential-based analysis methods used in parameter estimation and sensitivity studies are not considered here in further detail for reasons previously mentioned, plus their general inapplicability to characterize 3MRA across wide areas of the model input space needed for the national assessment strategy. Nonetheless, advancements in this area are promising, particularly in dealing with issues of non-linearity and more efficient sampling schemes. These advancements would, for example, likely provide useful inquiry into 3MRA on a site-specific basis, for example, in evaluation of future site-based, observation-based validation studies.

2.7.3 Screening Methods for Sensitivity Analysis

Screening designs can be applied as preliminary numerical experiments to isolate the most important model inputs from amongst a large number that may affect model output (Campolongo *et al.*, 1999, 2000b). Depending on analysis objectives, additional techniques would then be logically further applied, if not engaged first. As summarized by Campolongo *et al.* (2000b), screening designs are typically formed as "one-at-a-time" (OAT) experiments, evaluating each model input or factor in turn. This general class of experiments is linear with respect to the order of n_x. The approaches usually define a control experiment (i.e., nominal value of mid-point of input range) and examine two extremes over the allowable input range. Data analysis is usually formed as a residual comparison of extremes to the nominal control value.

Most methods under this class deal with only main effects of each model input considered, though factorial designs can also address parameter interactions, as well as higher-order OATs. Most rely on strict assumptions about the nature or absence of parameter interactions (Campolongo *et al.*, 2000b), and most are formed as local sensitivity analyses that require selection of a nominal value. Campolongo *et al.* (2000b) summarizes five categories offered by Daniel (1973):

1. Standard OAT designs that vary one factor at a time about a standard or nominal (local) condition,

2. Strict OAT designs that vary one factor from the condition of the last proceeding model run,

3. Paired OAT designs that produce two observations and one simple comparison at a time,

4. Free OAT designs that make each new model run under new conditions, and

5. Curved OAT designs that produce a subset of results by varying only one easy-to-vary model input.

Morris' OAT Design

Morris's method (Morris, 1991; Campolongo *et al.*, 2000a, 2000b) is classified as a global sensitivity algorithm due to its coverage of the entire input factor space, although only the second-order method attempts to estimate factor interaction. A unique aspect of Morris' method is that it does not depend on the choice of the nominal value selected. The method is preferred in complex, high-order modeling systems due to its computational proportionality to the number of model input factors considered ($n = 2*n_s*n_X$; where n = total model runs, n_s = number of realizations or number of statistical samples of the input space, and n_x represents the total number of input factors under consideration). A second-order Morris method, or extended Morris OAT (Campolongo *et al.*, 1999), is also available, as presented by Campolongo and Braddock (1997). Key aspects of the first-order Morris Method include (Campolongo *et al.*, 2000a, 2000b; Campolongo and Saltelli, 1997):

- Does not rely on simplifying assumptions between the input and output vectors.
- Representing a combination of individualized, randomized OAT designs, Morris's design attempts to determine the input factors that have:
  - Negligible effects,
  - Linear and additive effects, or
  - Nonlinear or interaction effects.
- Estimates the main effect of each factor by computing an average of several randomly selected local measures of sensitivity at different points in the input space.
- The local conditions examined are selected such that each factor is varied (discretely) over its entire input range in the model.
- The main advantage is its low computational cost, being linear, as opposed to exponential, with respect to the number of factors examined.
- A main disadvantage of the first-order approach is that parameter interaction is not addressed.

Cotter's OAT Design

The method by Cotter (1979) is described as a systematic fractional replicate design, and does not rely on prior assumptions about parameter interaction (Campolongo *et al.*, 2000a). The approach utilizes $2*n_x + 2$ model runs. Described by Campolongo *et al.* (2000b), the following model runs are constructed:

1. One initial run with all n_x inputs set at their minimum value,

2. Next, n_x runs with each factor individually set, in turn, at its maximum value while the other $n_x - 1$ inputs are set at their minimum value,

3. Next, n_x runs with each factor individually set at its minimum value while the other $n_x - 1$ inputs are set at their maximum value, and

4. One final run with all n_x inputs set at their maximum value.

Key aspects of the Cotter Method are further summarized by Campolongo *et al.* (2000a, 2000b):

- It does not rely on simplifying assumptions between the input and output vectors.

- A major disadvantage is that the method may not detect all important inputs when effects between inputs cancel.

- It is subject to increased variance, and thus, lacks the precision of other methods.

IFFD Method of Andres

Andres (Andres and Hajas, 1993) developed a factorial sampling approach known as the Iterated Fractional Factorial Design (IFFD). This approach is more ideal for higher-order models, offering increased efficiency in the number of total model runs required than the methods of Morris and Cotter. Here, $n < n_x$. As previously described by Saltelli (2002a), the IFFD is a group screening procedure where inputs are initially aggregated into groups of factors, such that an influential group must contain an influential model input (Campolongo *et al.*, 2000b). The basic approach is to repeat the analysis with different random groupings, where sensitive inputs will fall within the intersection of influential groups.

Key aspects of Andres' IFFD are summarized by Campolongo *et al.*, (2000a, 2000b):

- Requires fewer model runs (n) compared to factors (n_X), and ultimately samples three levels per factor (low, middle, and high values).

- Estimates main effects, quadratic effects, and two-factor interactions of the most influential factors.

- Ensures that sampling is balanced, where different combinations appear with equal frequency.

- It can be more efficiently coupled with stepwise regression to eliminate copycat factor effects, exploiting this redundancy.

- In contrast to Morris' method and OAT designs in general, IFFD relies on the outcome that only a few influential factors drive model output.

Efficiency in the IFFD is gained by exploiting the assumption that many higher order interactions are redundant. Recent modifications have been suggested (Andres, 1997) that combine the unit fractional factorial design (FFD) with LHS, where intervals are designated, as opposed to specific values as in the IFFD, and inputs are randomly sampled from these intervals (Campolongo *et al.*, 2000b).

The above screening-level techniques would normally be applied to the general model formulation for a typical site, and further might be evaluated separately on a chemical or WMU-specific basis. For national scale assessment strategies comprised of a set of site-specific modeling exercises, consideration would need to be given to differences in ranges of model input specification that may exist across sites, chemicals and metals, or waste management unit types.

2.7.4 Regression/Correlation Variance-Based Global Methods for SA

To evaluate the sensitivity of 3MRA, several regression/correlation-based global analysis techniques will be evaluated. These represent global sensitivity analysis techniques that will be investigated early on. They are based on random Monte Carlo sampling strategies, and rely on assumptions of near-linearity or near- monotonicity in describing relations between model inputs and outputs. Sensitivity analyses under this heading typically utilize various statistical techniques (e.g., scatter plots, regression and stepwise regression analyses, correlation and partial correlation analyses; all possibly with and without use of rank transformations).

A primary identifier of the depth of classification of a method as being global is the attribute of coverage of the entire input space in the analysis. Like higher-order, variance-based, total effect methods, methods of sensitivity analysis that fall into the category of regression and correlation explore the use of variance as a measure of the importance or sensitivity of model inputs (Chan *et al.*, 2000), exploring these relationships across the input space. Different from variance-based, total effect methods, regression/correlation-based approaches rely on a key assumption that the output and input model factors are near-linearly related, or their rank equivalents are near-monotonically related, respectively. Like all global methods discussed here, sampling-based designs are usually the mode of analysis in large or complex models.

Scatter Plots

Scatter plots are one of the most intuitive and straightforward techniques for sensitivity analysis (Campolongo *et al.*, 2000a). Extensively evaluated and succinctly presented by Helton and others, the generation of scatter plots is undoubtedly the simplest form of sensitivity analysis that can explore the full stratification of the input space (Helton and Davis, 2000; Helton, 1993; Kleijnen and Helton, 1999a, 1999b). The basic procedure is to plot pairs of data (x_{ik}, y_i) where k = 1, 2, …, n_s, for each element of x_i of $\mathbf{x}$ for i = 1, 2, …, n_x, for each output variable y_i, where n_s is the number of iterations in the sampling experiment.

In a detailed review of the procedure and its robustness by Kleijnen and Helton (1999a, 1999b), five general areas of investigation using scatter plots are detailed leading to identification of:

- Linear relationships using correlation coefficients,
- Monotonic relationships using rank correlation coefficients,
- Trends in central tendency as defined means, medians, and the Kruskal-Wallace statistic,
- Trends in variability as defined by variances and inter-quantile ranges, and
- Deviations from randomness as defined by the chi-square statistic.

Scatter plots offer a qualitative measure of input sensitivity since the relative importance or sensitivity of inputs can be estimated but not quantified (Campolongo *et al.*, 2000a).

Regression Analysis

Detailed by Helton and Davis (2000) and summarized by (Campolongo *et al.*, 2000a), multivariate regression analysis as a sensitivity analysis technique offers a more formal investigation of the mapping of the input space upon the output space. Linear regression approaches are, of course, the most accessible. Again, as with scatter plots, the data used in the analysis are normally arrived at through sampling-based simulation experiments. Typical approaches to derivation of regression models involve use of least squares to estimate model coefficients. A major feature of this approach for linear regression is the assumption of independence of the x_i's.

Standardized regression coefficients (SRC) are normally determined, and their absolute value can serve as a measure of sensitivity with respect to deviations from an input's mean value by a fixed fraction of its variance, while maintaining all other inputs at their expected value (Helton and Davis, 2000; Campolongo *et al.*, 2000a). Regression models are also typically subject to measures of the goodness of fit of the model, such as in the use of R^2 values (e.g., correlation of determination based on the Pearson Correlation Coefficient), PRESS values (prediction error sum of squares), or PRESS-R^2 values, a similar correlation-like expression based on PRESS values.

A primary limitation of this "all possible" approach, for most PC applications, is the number of variables, (e.g., a maximum of 15 inputs), which can be analyzed computationally in a single regression model (Hintze, 1997). "All possible" regression model solutions for higher numbers of model inputs quickly become computationally unfeasible. To overcome this, groups of variables (e.g., 15 at a time) can be analyzed iteratively, in essence forming a pseudo "all possible" regression analysis based on step-wise techniques.

Partial Correlation Measures

As indicated already in the discussion of evaluation of scatter plots and regression analysis, correlation measures serve as a common, useful concept to assess the relationship between inputs and outputs of a model (Helton and Davis, 2000). Similarly, a sequence of observations can also be accumulated through sampling techniques. For the case where more than one input is involved, partial correlation coefficients (PCC) are constructed on the set of observations (x_{ik}, y_i). Given the construction of a sequence of regression models, PCCs can be used to provide a measure of the linear relationship between a given output variable y_i and individual input variables x_i (Helton and Davis, 2000), where Iman *et al.*, (1985) have provided formal computational development of the procedure.

Interpretation of this approach versus standard linear regression, is summarized by Helton and Davis (2000):

- The PCC can be viewed as characterizing the effect that changing x_i by a fixed fraction of its standard deviation will have on y_i, measured relative to the standard deviation of y_i.

- The PCC characterizes the strength of the linear relationship between two variables after a correction is made for the linear effects of other variables in the analysis,

- In comparison, the SRC simply characterizes the effect on the model output from perturbing an input by a fixed fraction of its standard deviation.

- The PCC measure tends to exclude the effects of: (1) other variables; (2) the assumed distribution for the particular input; and (3) the magnitude of the impact of model input on the output.

- In comparison, the SRC is more influenced by the distributions assigned to the model inputs, and the magnitude of impact that the input has on the model output.

For the case in which the input variables are all uncorrelated, the order of variable importance determined by regression or partial correlation are exactly the same (Campolongo *et al.*, 2000a).

Step-Wise Regression Analysis

A more feasible approach for use in high-order models is that of (forward) step-wise regression. Laid-out by Helton (1993) and Helton and Davis (2000), a sequence of regression models is constructed using the following steps (Campolongo *et al.*, 2000a):

- The 1st regression model is based on the most influential model input,
- The 2nd regression model introduces the second most individually influential input, and
- The process of inclusion of increasingly influential inputs proceeds until subsequent regression models are unable to increase, meaningfully, the amount of variation in the output that can be accounted-for and described by the addition of more model inputs.

There are three primary reasons discussed by Helton and Davis (2000) for use of (forward) step-wise regression over standard regression:

1. Large numbers of variables make the regression computationally difficult and unwieldy to display results,

2. Usually only a small number of inputs are sensitive, and

3. Correlated variables result in unstable regression coefficients (i.e., coefficients whose values are sensitive to the inputs included in the regression model).

Rank Transformations

To avoid the problem of non-linearity, this non-parametric approach basically follows along similar lines as that of SRC and PCC development, except that rank ordered statistics are utilized instead. This approach is also detailed by Helton and Davis (2000). Here the data in SRC and PCC analysis is replaced by its corresponding ranks. The analysis proceeds similarly to that of the SRC and PCC, and the SRRCs and PRCCs are determined (i.e., standard rank regression coefficients and partial rank regression coefficients, respectively). Higher measures of sensitivity will typically be observed in the rank transformed approaches if non-linearity is significant, since these differences are themselves an indication of the non-linearity of the model.

Critical aspects in the use of rank transformed techniques as offered by Helton and Davis (2000) include:

- The presence of strong non-monotonic behavior renders the approach dubious,

- The rank transformation technique is essentially a construction of a different model, where the new model is more linear and more additive, and

- Care must be employed then when interpreting rank transformed results, since any conclusion drawn on the ranks does not translate easily back to the model itself.

2.7.5 Regional and Tree-Structured Global Methods for SA

In an effort to establish enhanced evaluation capabilities for 3MRA using global sensitivity analysis methods, two algorithms, Regional (or generalized) Sensitivity Analysis and Tree-Structured Density Estimation will also be explored. These algorithms, together, retain a strong potential to be used in tandem to define an integrated global analysis methodology for application to very-high order models (VHOMs). These methods, similar to the correlation/regression approaches described previously, are sampling-based where both random MCS and LHS could be employed.

As a future endeavor, the efficacy of an enhanced sampling strategy, Uniform Covering by Probabilistic Rejection (UCPR), may also be investigated as an aspect of developing an advanced integrated methodology for conducting global sensitivity analysis assessments. The latter technique, essentially a sampling scheme, would require substantial reformulation of system level processors in 3MRA to take advantage of the approach.

The envisioned, integrated, global sensitivity analysis computational approach comprises three Monte Carlo procedures – a basic univariate Regionalized Sensitivity Analysis (Spear and Hornberger, 1980; Chen and Beck, 1999; Beck and Chen, 2000; Osidele and Beck, 2001a), extended by a multivariate Tree Structured Density Estimation (Spear *et al.*, 1994; Osidele and Beck, 2001b), and augmented with a Uniform Covering by Probabilistic Rejection (Klepper and Hendrix, 1994) sampling procedure.

The following descriptive information on RSA, TSDE, and UCPR is summarized directly from the referenced works of Osidele and Beck (2001a, 2001b).

Regionalized Sensitivity Analysis (RSA)

The objective of the RSA procedure is to rank the importance of the uncertainties attributed to the model input factors, with respect to matching prescribed output behavior definitions, and on this basis, identify the key and redundant processes in the system. Monte Carlo simulation is performed with samples from a joint distribution of parameterized input factors. The model outputs are then classified as *behavior* {*B*} or *nonbehavior* {*NB*} simulations, depending on whether or not they fall within the constraints of the behavior definitions. For each input factor, a statistical test is performed on its marginal distribution in order to assess the difference between the sets of values that produced the {*B*} and {*NB*} simulations. A significant difference indicates a critical input factor, and hence a key system process. An insignificant difference suggests a redundant input factor and process. Analogously, a high significance level indicates a highly sensitive input, where the significance level of the underlying hypothesis test can form the basis for sensitivity classification of each model input. According to Osidele and Beck (2001a), the RSA best serves as a preliminary screening tool, to be supplemented by multivariate methods, such as a principal component analysis, cluster analysis, or the TSDE described below.

Kolmogorov-Smirnov Statistic

Described by Beck and Chen (2000), for each constituent model input x_i, the maximum separation distance, $d_{m,n}$, of the respective cumulative distributions of $\{x_i |B\}$ and $\{x_i |NB\}$ may be determined. The Kolmogorov-Smirnov statistic, $d_{m,n}$, is then used to discriminate between significant and insignificant separations for a chosen level of confidence. In notation here, d_{max} denotes a given model evaluation task description and simulation experiment, and $x_i |B$ infers that x_i is conditioned upon *B*. Distributions of $d_{m,n}$ can be formulated across *k* model evaluation tasks (Beck and Chen, 2000), and/or across sets of simulations assessing a single task (e.g., in the latter case assessing OSE in describing $d_{m,n}$ for a given task *k*). Osidele and Beck, (2001a) define the associated two-sample statistical test, formulated as a null hypothesis, to determine if the {*B*} and {*NB*} input values are identical, given by:

- Null hypothesis: H_o: $f_m(x_i |B) = f_n (x_i |NB)$, and
- Test statistic $d_{m,n} (x_i) = sup_x \left| F_m(x_i |B) - F_n(x_i |NB) \right|$.

Here, $F_m(x_i |B)$ and $F_n(x_i |NB)$ are the sample cumulative distribution functions for *m* behaviors and *n* nonbehaviors. The sup_x notation (i.e., the supremum over all X, X being an arbitrary continuous random variable) refers to the greatest vertical distance between two cumulative distributions.

Behavior Definitions

As an example, a logical behavior definition for 3MRA would be identification of those portions of the input space that lead to increased risk above some nominal population percentile (e.g., evaluation of site-specific scenarios that result in < 90% receptor population protection at the site, where the sensitivity analysis would operate on individual site-based scenarios across waste stream concentrations). An extremely important benefit of the approach is that behavior definitions can be established after the computational effort, provided that mapping between input and output spaces are appropriately preserved.

As discussed in Sections 2.3.7 and 2.8, the RSA will also be utilized as an approach for conducting a sensitivity-based performance validation (SA_p). The elegance of the RSA approach for the (SA_p) effort is that the same computational effort needed for the RSA sensitivity analysis procedure can also be utilized.

Tree-Structured Density Estimation (TSDE)

The aim of the TSDE procedure is to identify interactions among the input factors in the *behavior* $\{B\}$ simulations of the RSA procedure. Thus, it supports the univariate RSA procedure with a qualitative, multivariate analysis. Using a simple density estimate, the original RSA sampling domain is recursively partitioned, in a sequence of binary splits, into low- and high-density subdomains. This process is depicted by a binary tree, in which the nodes represent the subdomains and the branches (the splits) are determined by the key input factors. Tracing a high-density terminal node from the root node is equivalent to locating those regions of the input factor space with a high probability of producing $\{B\}$ simulations.

The sequence of input factors in any such trace identifies the set of factors that interact to produce a $\{B\}$ simulation. Thus, the number of high-density terminal nodes that each input factor helps define is directly related to its relative importance in the model and system. Furthermore, the combined volume of the high-density terminal nodes, in proportion to the overall sampling domain volume, is a measure of the probability of realizing the prescribed behavior definition.

According to Osidele and Beck (2001b), the TSDE addresses a critical weakness of the RSA in its use of marginal input distributions, which discount the correlations among model inputs. While a significant difference between $\{B\}$ and $\{NB\}$ values is sufficient to indicate a sensitive model input, the converse is not always true. This is due to the fact that each correlating parameter exhibits flattened marginal distributions, which does not clearly distinguish between the $\{B\}$ and $\{NB\}$ values (Osidele and Beck, 2001b).

Uniform Covering by Probabilistic Rejection (UCPR)

The main purpose of the UCPR sampling procedure (Klepper and Hendrix, 1994) is to systematically search the input factor domain for combinations of values that produce *behavior*

{B} simulations. UCPR enhances the statistical power of TSDE by augmenting the sample of behavior-producing input factors. In a series of iterations, the UCPR can progressively update the set of behavior-producing input factors by selecting trial values in close proximity (distance-wise) to the current set.

2.7.6 Other Variance-Based Global Methods for Sensitivity Analysis

In addition to methods already presented, the following regression/correlation-based and "total-effect" approaches are briefly reviewed. These represent additional, higher-level techniques that are also Monte Carlo or LHS sampling-based. They are all extremely computationally demanding, and are non-trivial to implement in large or complex models. All of these approaches are variance-based, quantitative, and adhere to several or all (e.g., Sobol's Method) of Saltelli's precepts for ideal global-based sensitivity methods.

For long-term research planning at the Office of Research and Development, it is envisioned that these additional higher order sensitivity analysis methods will eventually become functionally available within FRAMES (see Figure 1-1). Given sufficient computational power, these would provide additional insights into the structural behavior of 3MRA and other high order multimedia models and modeling systems. Incorporation and implementation of these techniques in 3MRA would be on the order of years at current staffing levels.

Evaluation of these techniques in 3MRA would be pursued, for example, after the evaluation of simpler, screening methods were complete, as described previously, along with initial development of capabilities for conducting RSA and TSDE. Two of the higher-order methods described below do not rely on regression or correlation analysis to quantify sensitivity structure. Instead, they attempt to develop model-independent, variance-based, quantitative "total-effect" measures of importance. These essentially represent the most resolved statements approachable in establishing variance relationships between model input and output quantities.

Depending on the 3MRA initial model evaluation results, and contingent upon resource constraints, additional higher-order methods investigated for 3MRA and other models may include evaluation of several or all of the following algorithms and methodologies (Chan *et al.*, 2000; Campolongo *et al.*, 2000a).

- **Correlation Ratio Analysis** (i.e., a variance-based Monte Carlo Method)
- **Sobol's Total Effect Method** (i.e., a variance-based Monte Carlo Method)
- **Fourier Amplitude Sensitivity Test** (FAST; classical 1st order indices and extended "total effect" sensitivity analysis; represents a Fourier Series decomposition of variance).

In many cases, one can employ simpler sensitivity analysis approaches, such as those previously discussed (e.g., RSA, TSDE, regression/correlation approaches, Morris' Method, IFFD, etc.) to reduce the overall number of model inputs to be investigated under these more sophisticated and computationally demanding higher-order approaches. Specifically, the application of these latter approaches becomes highly feasible after identification of 10's of key model inputs from among the hundreds possible. FAST (Cukier *et al.*, 1973, 1975, 1978) is well described in the literature, with important summaries and example applications available

(Campolongo and Saltelli, 1997; Campolongo *et al.*, 1999; Campolongo *et al.*, 2000a, Chan *et al.*, 2000; Helton and Davis 2002; Saltelli *et al.*, 2002a, 2002b; Frey and Patil, 2002). Sobol's method (Sobol, 1993; Campolongo and Saltelli, 1997; Campolongo *et al.*, 2000a; Chan *et al.*, 2000; Saltelli, 2002b) is also described with applications in the literature. Key attributes of quantitative, variance-based methods include the ability to handle non-linearity and non-monotonic behaviors, and to directly relate the "total effects" of variance in input factors on model output.

As a comparison of the required computational effort, for example, FAST and Sobol's Method are the most computationally demanding as a function of the number of model inputs, n_x (e.g., $n = n_s *2^{n_x}$). FAST requires typically ½ the number of runs as Sobol's Method, with some loss of information. More recently, Saltelli (2002b) has offered more efficient sampling schemes to reduce the "curse of dimensionality" associated with variance-based, sensitivity analysis methods. In comparison to FAST and Sobol's Methods, standard Correlation Ratio Analyses that also derive "measures of importance" retain the criticism that they lack robustness and can be unduly leveraged by outliers in input distributions (Campolongo *et al.*, 2000a; Iman and Hora, 1990).

2.8 Sensitivity-Based Performance Validation Using RSA-TSDE

Discussed in Section 2.3.1, the primary problem of reaching a satisfactory, empirically based measure of model performance validation, in the present, is restrained by two dilemmas: (1) the future truth we seek is paradoxically unobservable in the present, and (2) subjective decision variables used in complex problems, such as exposure and risk assessments, are realistically unobservable in the present and future. Despite these dilemmas, the existence of this "validation paradox" does not render us without an opportunity for some level of objective judgment about the uncertainty in a given decision, and the overall performance validity of a model for a specific intended use (Beck *et al.*, 1997).

The methodology for establishing performance validation of model behavior under novel conditions is described in a simplified example of a national application for multimedia modeling of hazardous waste disposal (Chen and Beck, 1999; Beck and Chen, 2000). Adopting their procedure for 3MRA, this will be formed, in principle, as a judgment in the present as to how well we expect the model to perform its designated task reliably, with a minimum risk of an undesirable outcome, and in a maximally relevant manner. The underlying theoretical background for this construction was first demonstrated in the seminal work of Hornberger, Spear, and Young (Young *et al.*, 1978; Hornberger and Spear, 1980; and Spear and Hornberger, 1980), with context provided on the question of model evaluation by Beck (1987) and Beck *et al.* (1997).

According to Beck *et al.* (1997), the concept of establishing the "prior performance" validity of a model is summarized in the creative freedom of an analyst in defining the "purpose" of a model, while departing from historical notions of history-matching in attempting, to locate a sample of randomly generated values for the model's inputs that:

- Enable the model outputs to match certain crude constraints on what is “defined” (not actually observed) to be an acceptable statement of “past” behavior.

- Enable the model outputs to match certain crude constraints on what is “defined” as radically different behavior of the system in the future.

- Result in exposure [or risk] above or below a given level (including “no concern”, extreme or “high-end” exposures), or within a given confidence band around a specified probability of occurrence.

The three purposes outlined above represent a reflection of the evaluation of the external definition of the task back onto the internal composition of the model (Beck *et al.*, 1997). Summarized by Beck *et al.*, (1997), the Hornberger-Spear-Young (HSY) algorithm identifies those model inputs that are crucial to discriminating a match from a mismatch of the model’s outputs with the reference behavior, and, by reflection, those parameters that are redundant in this discriminating function. Maximum relevancy is, in the formulation of these authors, then an assessment of key (sensitive) and redundant (insensitive) parameters, representing an identification process irrespective of output uncertainty, placed in context of a binary classification of behavior and non-behavior in the model output space.

2.8.1 Hornberger-Spear-Young (HSY) Algorithm

The above works serve as reference points in an approach to establishing a “prior performance validation” of 3MRA Version 1.0/1.x. As an untested extension of this concept, the TSDE procedure (Spear *et al.*, 1994; Osidele and Beck, 2001b) will also be employed in evaluation of 3MRA to better understand efficacy of this multivariate counterpart procedure in further constructing a statement of “prior performance”.

The material following in Section 2.8.2 briefly discusses specific aspects of such a performance sensitivity analysis, as a mechanism for quantification of performance validation under novel, future conditions. This material is taken directly from the EPA document authored by Chen and Beck (1999; see also Beck and Chen, 2000), and is wholly attributed to these authors. To remain brief here, and still offer the reader a thorough perspective in understanding the RSA-based approach, and the earlier research supported by EPA, the full text of this application example, serving as a blueprint of sorts for use in the more complex multimedia model 3MRA, is presented in Appendix B.

2.8.2 Summary Excerpts From Chen and Beck (1999)

Overview

The paper of Chen and Beck (1999) explores three groups of tests that might be formulated to determine model reliability. The first of these is concerned with establishing whether the uncertainties surrounding the parameterization of the model render it impotent in discriminating between which of two sites, say, gives the significantly higher predicted receptor concentration of contaminant, in conditions where this result would generally be expected. The second test is a straightforward form of regionalized sensitivity analysis designed to identify which of the model's parameters are critical to the task of predicting exceedance, or otherwise, of prescribed (regulatory) receptor-site concentrations. The third test is designed to achieve a more global form of sensitivity analysis in which the dependence of selected statistical properties of the distributions of predicted concentrations (mean, variance, and 95th-percentile) on specific model parameters can be investigated.

The latter is formulated by Chen and Beck in the context of key and redundant controls in achieving a given level of site performance. The former two approaches are summarized below.

Example Model Description

In its time-invariant form, the structure of a model may be defined by the equation $y = g\{\mathbf{x}, \mathbf{u}, \boldsymbol{\alpha}\}$ for the state vector $\mathbf{x}$, in which, in principle, x denotes the field of contaminant concentrations in the subsurface environment, $\mathbf{u}$ is a vector of inputs to the system, and $\boldsymbol{\alpha}$ is a vector of model inputs. The simpler multimedia model considered by Chen and Beck required the specification of more than 30 parameters ($\boldsymbol{\alpha}$) for its application to a Subtitle D facility. The equation is solved in the context of a Monte Carlo simulation, thus generating distributions for y as a function of the assumed uncertainty associated with $\boldsymbol{\alpha}$ and $\mathbf{u}$.

In order to be effective as a tool for determining whether contamination arising from a storage facility will be significant and, in that event, what may be done to remedy such an unacceptable situation, the multimedia model must be able to demonstrate that uncertainty about the value of y, as a result of the substantial uncertainties in $\mathbf{u}$ and $\boldsymbol{\alpha}$, does not undermine the basis of decision-making. In the extreme, for example, the outcome that more or less any value of y is equally probable under any given combination of soil, contaminant and hydrological regimes is hardly a secure basis on which to construct a decision. There are several issues to be addressed in assuring the quality of the model's predictive performance. Chen and Beck provided computational results for three such issues and indicated a fourth promising line of analysis.

Output Uncertainty as a Function of Different Site Characteristics

Suppose that the same contaminant is stored at several sites, with each site having a different underlying soil, aquifer and hydrological regime. From the perspective of making a decision relating to the performance of each such facility, interest would focus on the capacity to predict the residual contaminant concentrations y at the respective receptor sites in order to establish which facility is the most, or least effective in containing the particular contaminant. Formally, it is necessary to determine whether the model is able to separate the respective distributions of y, let us say y_A and y_B, for two sites A and B, respectively, parameterized in soil and hydrological terms through $\boldsymbol{\alpha}_A$ and $\boldsymbol{\alpha}_B$. By "separation" we mean that the probability of identical values of y being generated under the two (storage site) scenarios is less than some threshold, such as 0.01, 0.05, or 0.10 (i.e., a statement of a single evaluation task k for 2 or more models). Alternatively, it may be desirable to explore the scope of the model in discriminating (for a single site) between the residual concentrations of two (or more) contaminants with differing degradabilities, likewise parameterized through different ranges of values for $\boldsymbol{\alpha}$.

From the practical perspective of making a decision, for example, to rectify inadequate performance at site A or B, such an analysis could be used to quantify the risk of taking the (wrong) action, say, at site A, when in reality site B is the more poorly performing storage system. Here our concern is primarily with what this kind of analysis may illuminate with respect to the power of the given model to discriminate the predicted behavior of one site from that of another. Given that there are strong prior beliefs that site or contaminant characteristics ought to generate distinctly different receptor site concentrations under reasonable model parameter uncertainty, the result that this is so (or not so) is revealing of the discriminating power, or relevance, of the model in performing the stated task. Indeed, some formal manipulation of the probability of coincident, i.e., indistinct, values of y might be used as a quantitative measure of this power.

Key and Redundant Inputs in Predicting a Percentile Concentration

The latter analysis can be viewed as follows. A screening-level assessment of the risk of adverse exposure at the receptor site is concerned with knowledge of the probability that a particular contaminant concentration (ý) will be exceeded. The choice of specific values for some of the parameters in the model, within the range of values they might assume, may be key to governing whether the resulting prediction of y falls above or below ý. For other parameters, the choice of a specific value may be immaterial to such discrimination in terms of y being above or below ý. The quality of the model in performing this screening task might, therefore, be related to the relative numbers of key and redundant model parameters, $\{\boldsymbol{\alpha}^K\}$ *and* $\{\boldsymbol{\alpha}^R\}$ *respectively, that are so found.*

In general, then, our interest lies in determining which are the key parameters $\{\boldsymbol{\alpha}^K(p)\}$*, their uncertainty notwithstanding, that govern the ability of the model to discriminate the prediction of* $y \leq ý(p)$ *from the prediction of* $y > ý(p)$*, where (1-p) is the probability of y exceeding the given value . In other words, for which of the model's parameters would the best possible knowledge be required in order to determine a particular percentile of the distribution of the contaminant concentration at the receptor site? Also, do the same parameters in the model appear to be key (or redundant) in discriminating among the predictions of y in the vicinities of a range of percentiles (p), such as 99%, 95%, 90%, 80%, 50%, and so on (i.e., a statement of a multiple evaluation tasks* k_i *for a single model)? In sum, we would like to know whether the multimedia model is a good (reliable) model for predicting the entire range of exposures, or just the high-end exposures, or merely the mean exposures. If it is not judged to be reliable for fulfilling any of these tasks, we would like to know further which of its parts are the least secure.*

In order to answer these questions, Chen and Beck used the RSA algorithm of Hornberger, Spear, and Young (HSY), described in Section 2.7.5. The same form of analysis has been extended in an application to the model MMSOILS, a close relative of EPAMMM (Chen and Beck, 1999), in which the goal was to identify and explore how particular clusters (or aggregate assemblies) of parameters, as opposed to individual parameters, might be key or redundant in the above discriminating function (Spear *et al.*, 1994; i.e., TSDE). Chen and Beck used the "basic" form of the RSA, but note that any interpretations of its results will be subject to the limiting qualifications of the RSA illuminated by Spear *et al.* (1994).

This second case, predicting a percentile concentration, is introduced in the UA/SA plan in Section 9 as a formal procedure that will be undertaken in assessing the “prior performance” validity of 3MRA.

2.9 EPA Guidance for Use of Sampling Based Approaches in Risk Analysis

In closing this discussion on techniques for probabilistic risk assessment, an important backdrop has also been presented by EPA for evaluation of approaches to be used for 3MRA uncertainty and sensitivity analyses. Table 2-9, presents a summary, edited by Burns (2001), on the EPA’s established Policy for Use of Probabilistic Analysis in Risk Assessment (Hansen, 1997).

Companion text to this policy statement, much of which is captured, has also been developed as “guiding principles” for conducting Monte Carlo analysis for probabilistic risk assessment (EPA, 1996a, 1997a, 1997b). Summarized by Burns (2001), two important concepts are established by this EPA policy:

1. EPA’s current policy reaffirms the place of deterministic methods in the suite of Agency methods: “[Probabilistic] analysis should be a part of a tiered approach to risk assessment that progresses from simpler (e.g., deterministic) to more complex (e.g., probabilistic) analyses as the risk management situation requires.”

2. More importantly, the policy statement establishes a set of “conditions for acceptance” by the Agency of probabilistic analyses. These conditions, intended to encourage the ideals of transparency, reproducibility, and the use of sound methods, identify factors to be considered by Agency staff in implementing the policy.

Laid-out in the underlying framework methodology (Marin *et al.*, 1999; Appendix A), the development of 3MRA, and its associated data collection effort (see Volume II), was formulated to comply with U.S. EPA’s Guiding Principles for Monte Carlo Analysis (1997a).

Table 2-1. Probabilistic Risk Assessment Terminology.

Probabilistic Definitions
Population
Individual
Sample
Random Sample
Probability Sample
Random Variable
Stochastic Process
Frequency
Probability
Probability Distribution
Fractile
Percentile
Probability Density Function (PDF)
Cumulative Distribution Function (CDF)
Complementary CDF (CCDF)
Inverse CDF (ICDF)

Table 2-2. 3MRA Population Examples.

Population and Subpopulation Sets
Human receptors across the nation
Receptors within a given region
Adult humans within a given region
Regions in the nation
Sites in the sampled database
Sites in the nation
Sites within a given region
Receptors within a radial distance of a given site
90th percentile of population exposed to some cancer risk

Table 2-3. Types of Quantities Used in Models.

Quantities in Models	Amenable to Uncertainty Analysis	Amenable to Sensitivity Analysis
Empirical	/	/
Defined Constants		/
Decision Variables		/
Value Parameters		/
Index Variables		/
Model Domain Parameters		/

Table 2-4. Components of Model Evaluation.
(a.k.a., Verification, Validation, and Predictive Uncertainty Analysis)

Components of Model Evaluation
Uncertainty (U)
Variability (V)
Total Uncertainty (TU)
Compositional Uncertainty Analysis (UA_c)
Performance Uncertainty Analysis (UA_p = UA)
Sensitivity Analysis (SA)
Calibration (CAL)
Code Verification (CodVer)
Model Comparison (ModComp)
Compositional Validity (CompVal)
Performance Validity (PerfVal)
Model Validation (ModVal)
Peer review

Table 2-5. Sources of Uncertainty in Empirical Quantities.
(From Morgan and Henrion, 1990)

Sources of Uncertainty in Empirical Quantities
Statistical variation
Subjective judgment
Linguistic imprecision
Variability
Inherent randomness
Disagreement
Approximation

Table 2-6. Types of Uncertainty.

General Classes of Uncertainty
Variability (V)
Empirical Uncertainty (U)
Model Error (ME)
Types of Empirical Uncertainty
Random Error (RE)
Systematic Error (SE)
Sample Measurement Error (SME; see RE, SE)
Input Sampling Error (ISE; see RE)
Output Sampling Error (OSE; see RE)
Inherent randomness
Correlation
Disagreement

Table 2-7. Terms Used for Describing Variability and Empirical Uncertainty.

Variability	Empirical Uncertainty
Stochastic Uncertainty	Epistemic Uncertainty
Aleatory Uncertainty	Lack-of Knowledge Uncertainty
Objective Uncertainty	Subjective Uncertainty
Type A Uncertainty	Type B Uncertainty

Table 2-8. Sources of Uncertainty in Empirical Quantities.
(From Cullen and Frey, 1999)

Model Solution Approaches
Analytical Solutions for Methods
Central Limit Theorem
Properties of The Mean and Variance
Analytical Solutions for Distributions
Transformation of Variables
Approximation Methods for Moments
First-order Methods
Taylor Series Expansion
Numerical Methods
Monte Carlo Simulation
Latin Hypercube Cube Simulation
Importance Sampling
Fourier Amplitude Sensitivity Test
Others

Table 2-9. EPA Policy: "Conditions for Acceptance" of Probabilistic Risk Assessments.
(Hansen, 1997; From Burns, 2001)

"Conditions for Acceptance" of Probabilistic Risk Assessments
1. The purpose and scope of the assessment should be clearly articulated in a "problem formulation" section that includes a full discussion of any highly exposed or highly susceptible subpopulations [that have been] evaluated. ... The questions the assessment attempts to answer are to be discussed and the assessment endpoints are to be well defined.
2. The methods used for the analysis (including all models used, all data upon which the assessment is based, and all assumptions that have a significant impact upon the results) are to be documented and easily located in the report. This documentation is to include a discussion of the degree to which the data used are representative of the population under study. Also, this documentation is to include the names of the models and software used to generate the analysis. Sufficient information is to be provided to allow the results of the analysis to be independently reproduced.
3. The results of sensitivity analyses are to be presented and discussed in the report. Probabilistic techniques should be applied to the compounds, pathways, and factors of importance to the assessment, as determined by sensitivity analyses or other basic requirements of the assessment.
4. The presence or absence of moderate to strong correlations or dependencies between the input variables is to be discussed and accounted for in the analysis, along with the effects these have on the output distribution.
5. Information for each input and output distribution is to be provided in the report. This includes tabular and graphical representations of the distributions (e.g., probability density function and cumulative distribution function plots) that indicate the location of any point estimates of interest (e.g., mean, median, 95th percentile). The selection of distributions is to be explained and justified. For both the input and output distributions, variability and uncertainty are to be differentiated where possible.
6. The numerical stability of the central tendency and the higher end (i.e., tail) of the output distributions are to be presented and discussed.
7. Calculations of exposures and risks using deterministic (e.g., point estimate) methods are to be reported if possible. Providing these values will allow comparisons between the probabilistic analysis and past or screening level risk assessments. Further, deterministic estimates may be used to answer scenario specific questions and to facilitate risk communication. When comparisons are made, it is important to explain the similarities and differences in the underlying data, assumptions, and models.
8. Since fixed exposure assumptions (e.g., exposure duration, body weight) are sometimes embedded in the toxicity metrics (e.g., Reference Doses,...[96-hour LC50]), the exposure estimates from the probabilistic output distribution are to be aligned with the toxicity metric.

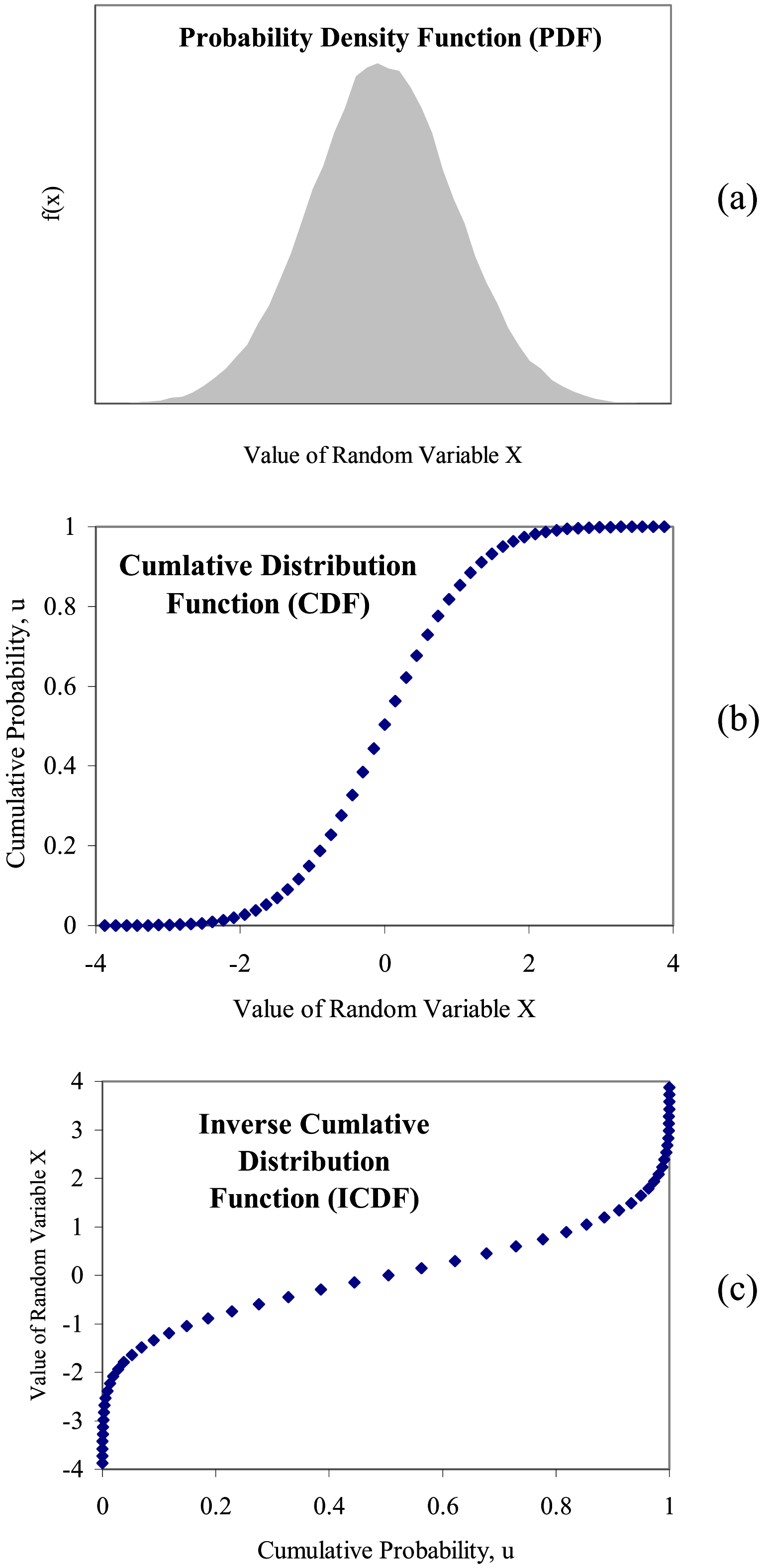
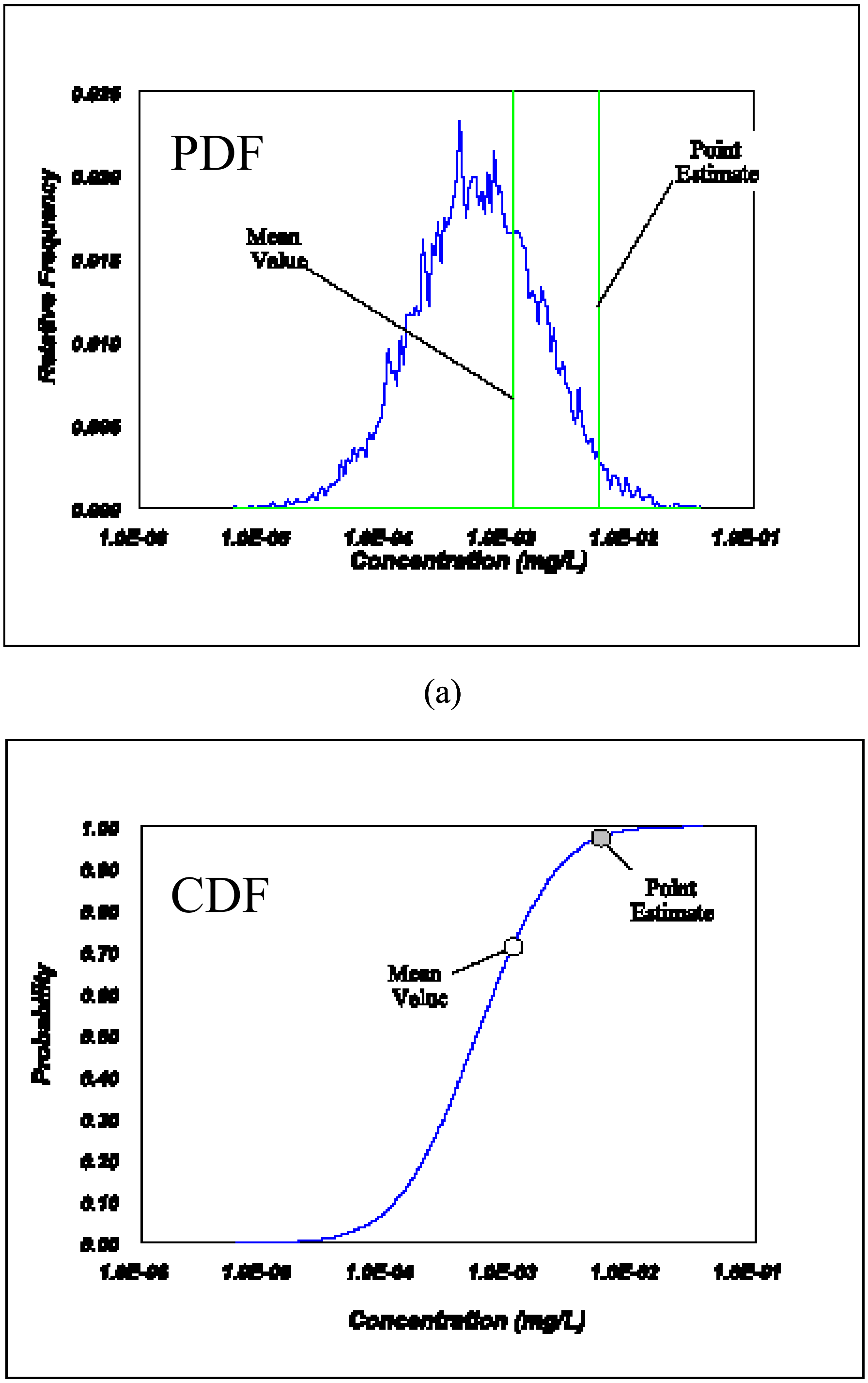


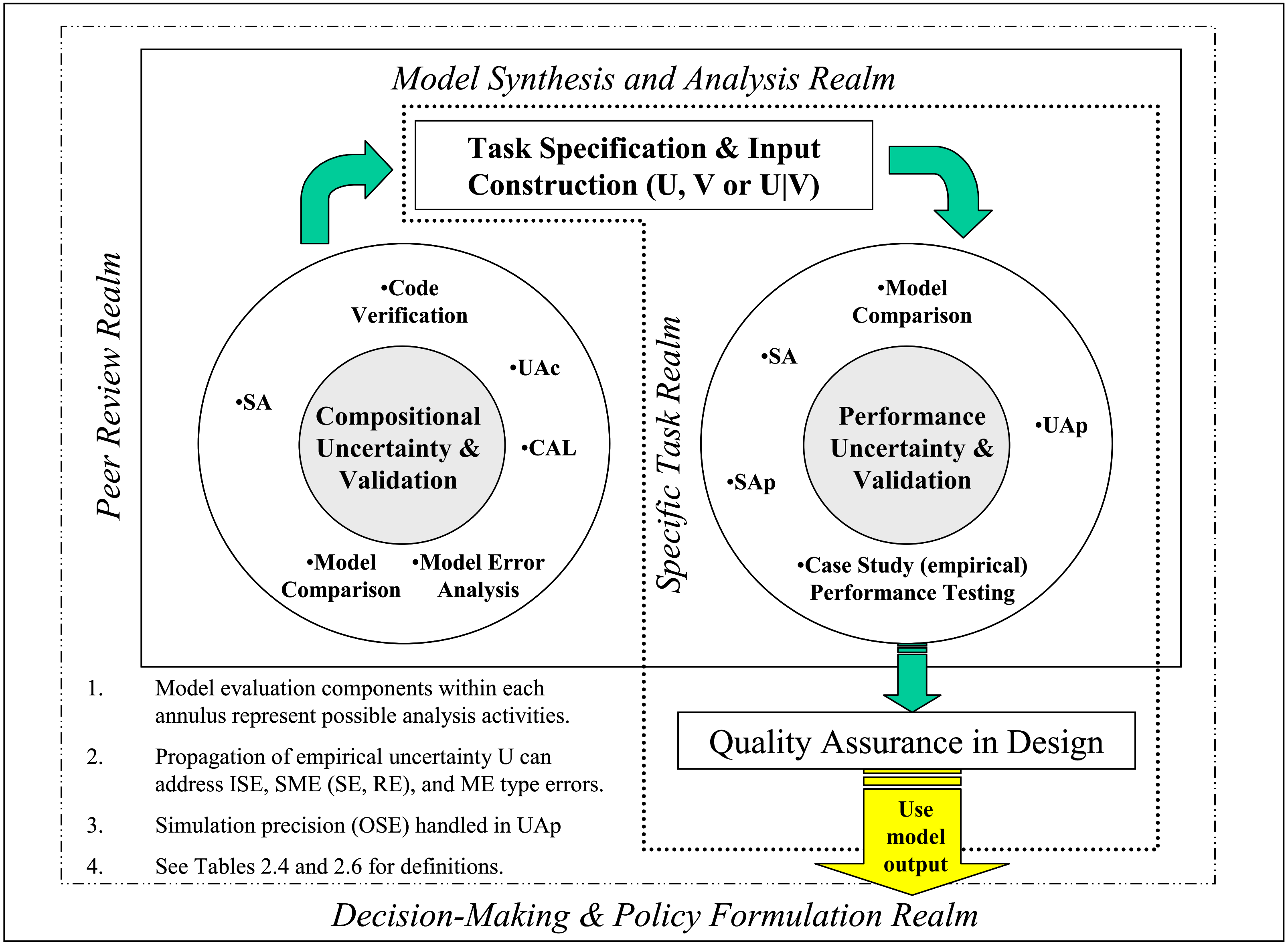
Figure 2-1. Probability Curves (PDF, CDF, and ICDF) for Random Variable X = N(0, 1).



(a)

(b)

Figure 2-2. Example of Measured Sampled Empirical Input or Estimated Model Output.
(adapted from Burns, 2001)
(a) Probability Density Function (PDF) of Model Input or Output
(b) Cumulative Distribution Function (CDF) of Model Input or Output


Model Synthesis and Analysis Realm
Task Specification & Input Construction (U, V or U|V)
Peer Review Realm
Specific Task Realm
•Code Verification
•SA
•UAc
•CAL
Compositional Uncertainty & Validation
•Model Comparison
•Model Error Analysis
•Model Comparison
•SA
•SAp
•UAp
Performance Uncertainty & Validation
•Case Study (empirical) Performance Testing
Quality Assurance in Design
Use model output
Decision-Making & Policy Formulation Realm

1. Model evaluation components within each annulus represent possible analysis activities.
2. Propagation of empirical uncertainty U can address ISE, SME (SE, RE), and ME type errors.
3. Simulation precision (OSE) handled in UAp
4. See Tables 2.4 and 2.6 for definitions.

Figure 2-3. Model Evaluation Paradigm.

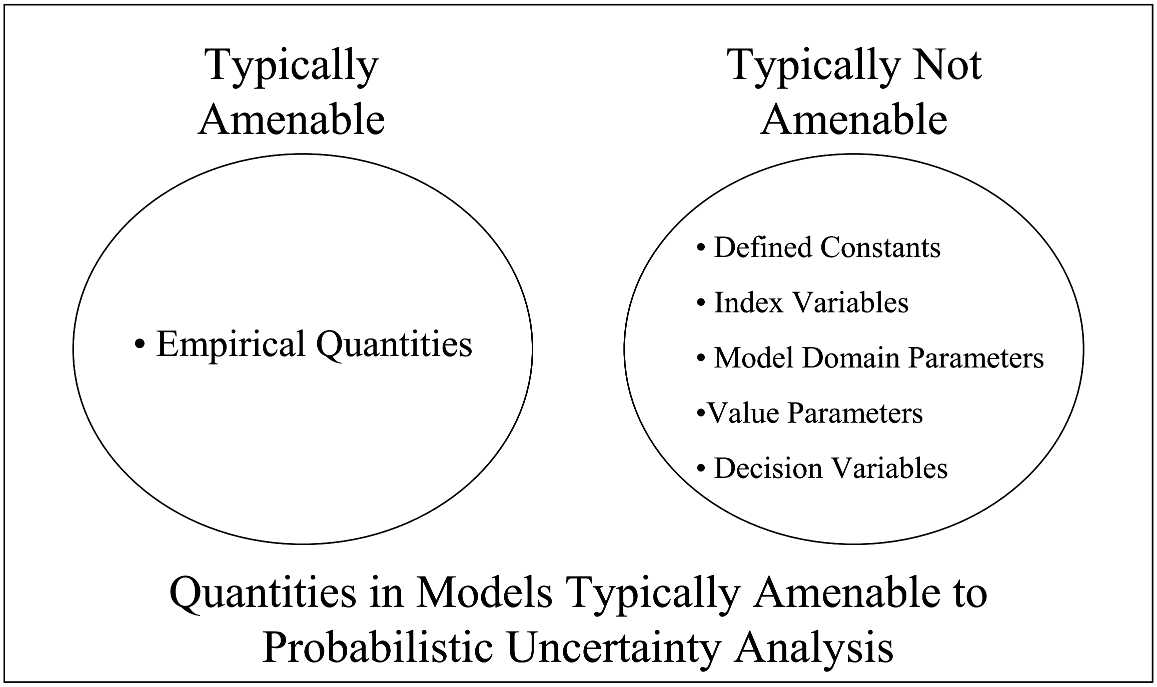


Figure 2-4. Model Quantities Subject to Probabilistic Uncertainty Analysis.

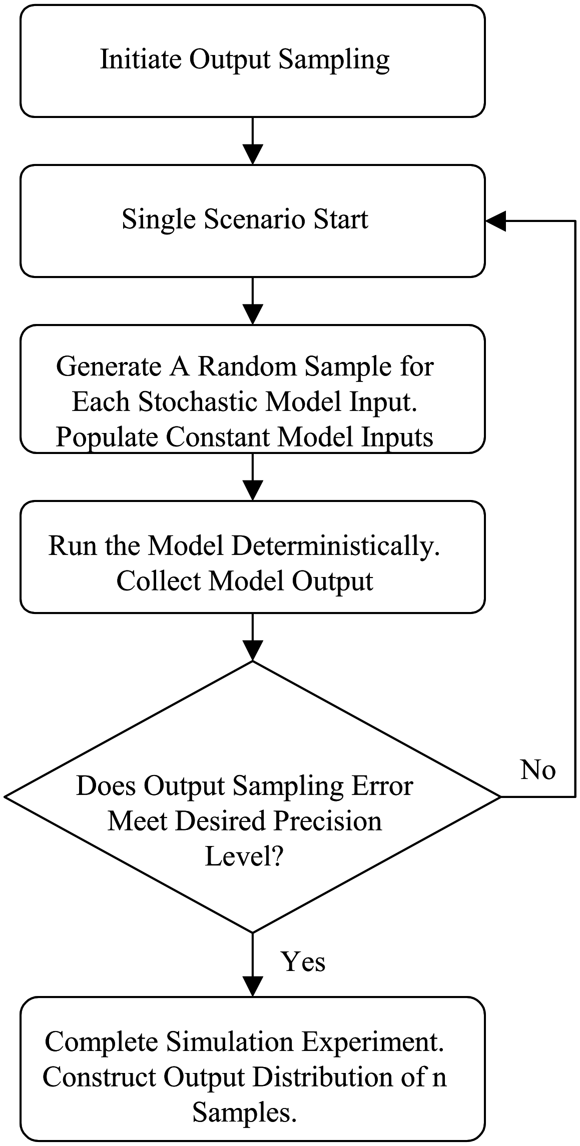

Initiate Output Sampling
Single Scenario Start
Generate A Random Sample for Each Stochastic Model Input. Populate Constant Model Inputs
Run the Model Deterministically. Collect Model Output
Does Output Sampling Error Meet Desired Precision Level?
No
Yes
Complete Simulation Experiment. Construct Output Distribution of n Samples.

Figure 2-5. Typical Monte Carlo Flowchart for Stochastic Model Solutions.

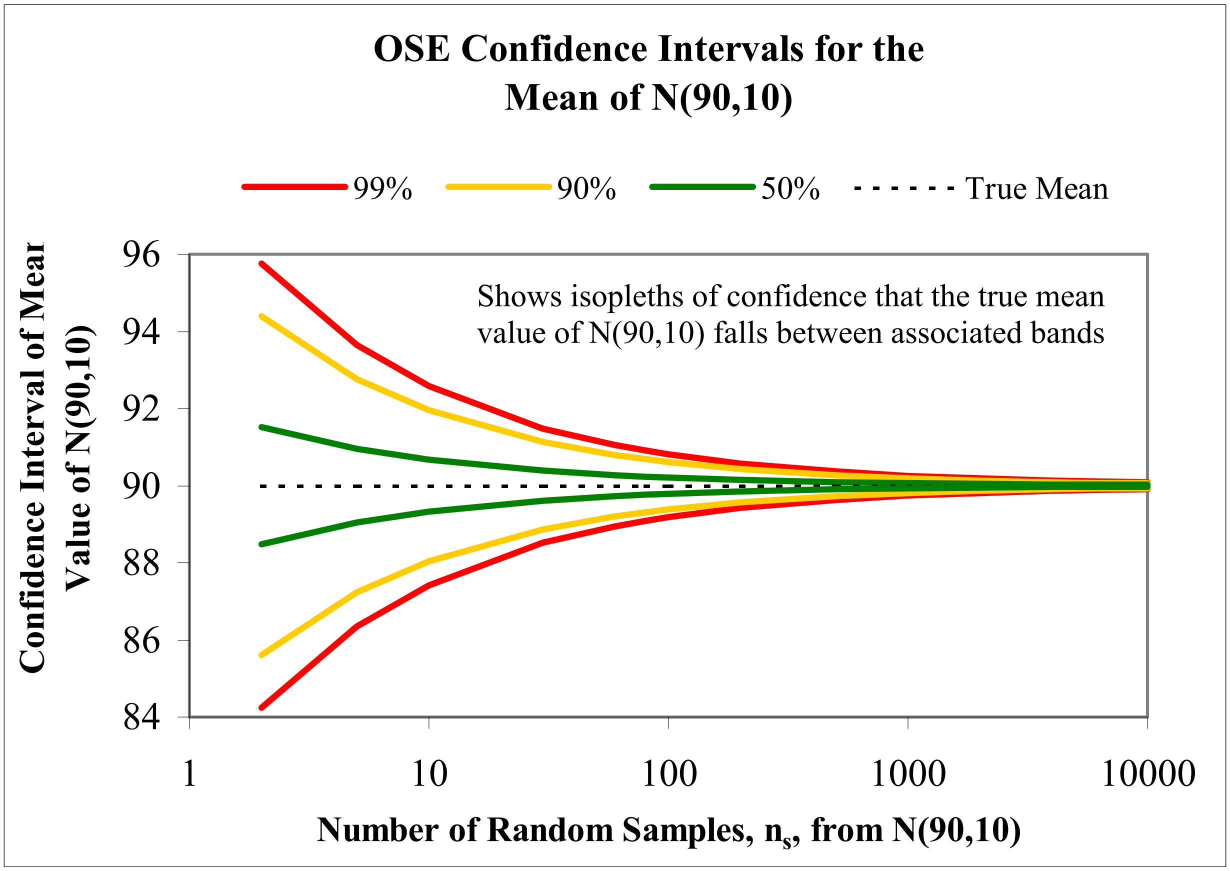


Figure 2-6. Estimating Sample Size: Error in Estimating the Normal Mean.

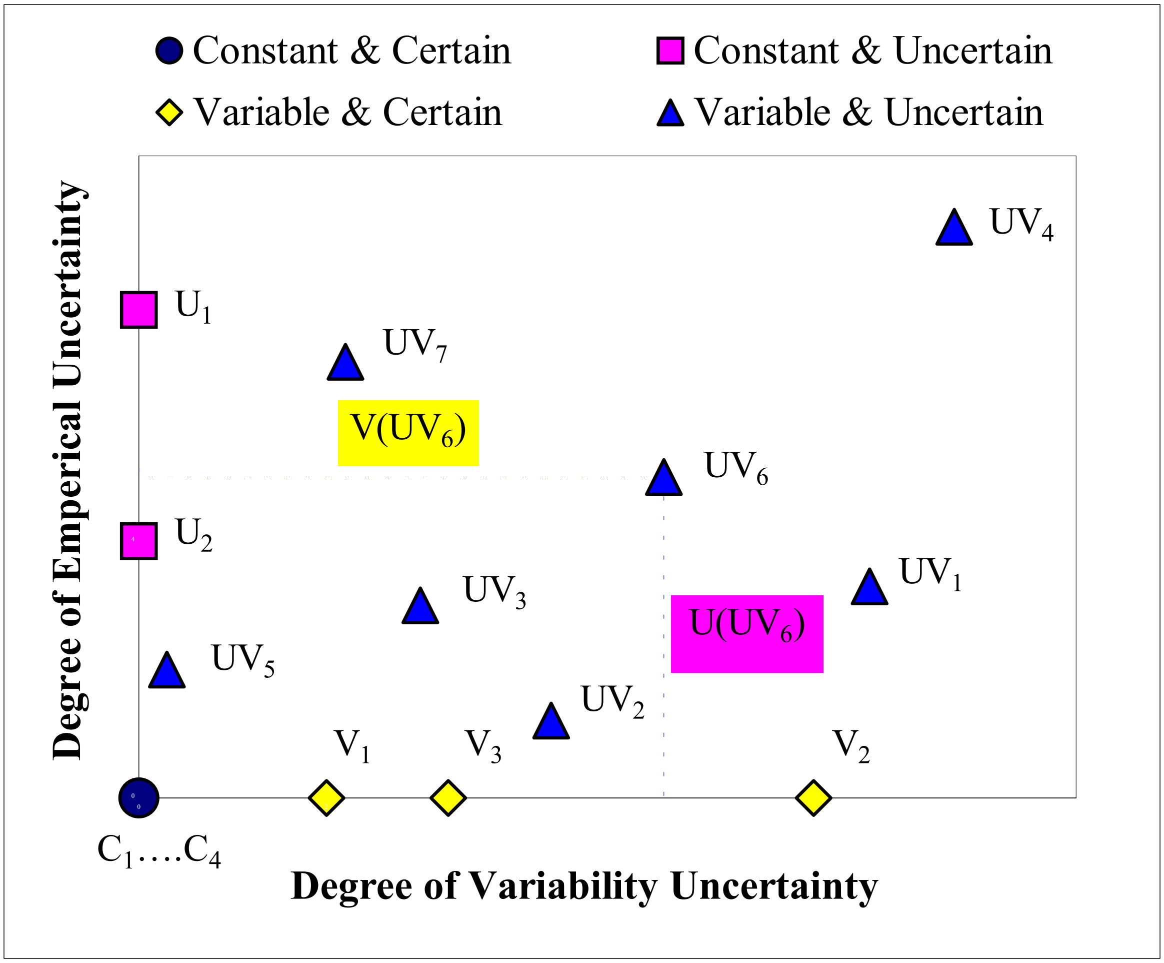


Figure 2-7. Degrees of Variability and Empirical Uncertainty in Model Inputs.

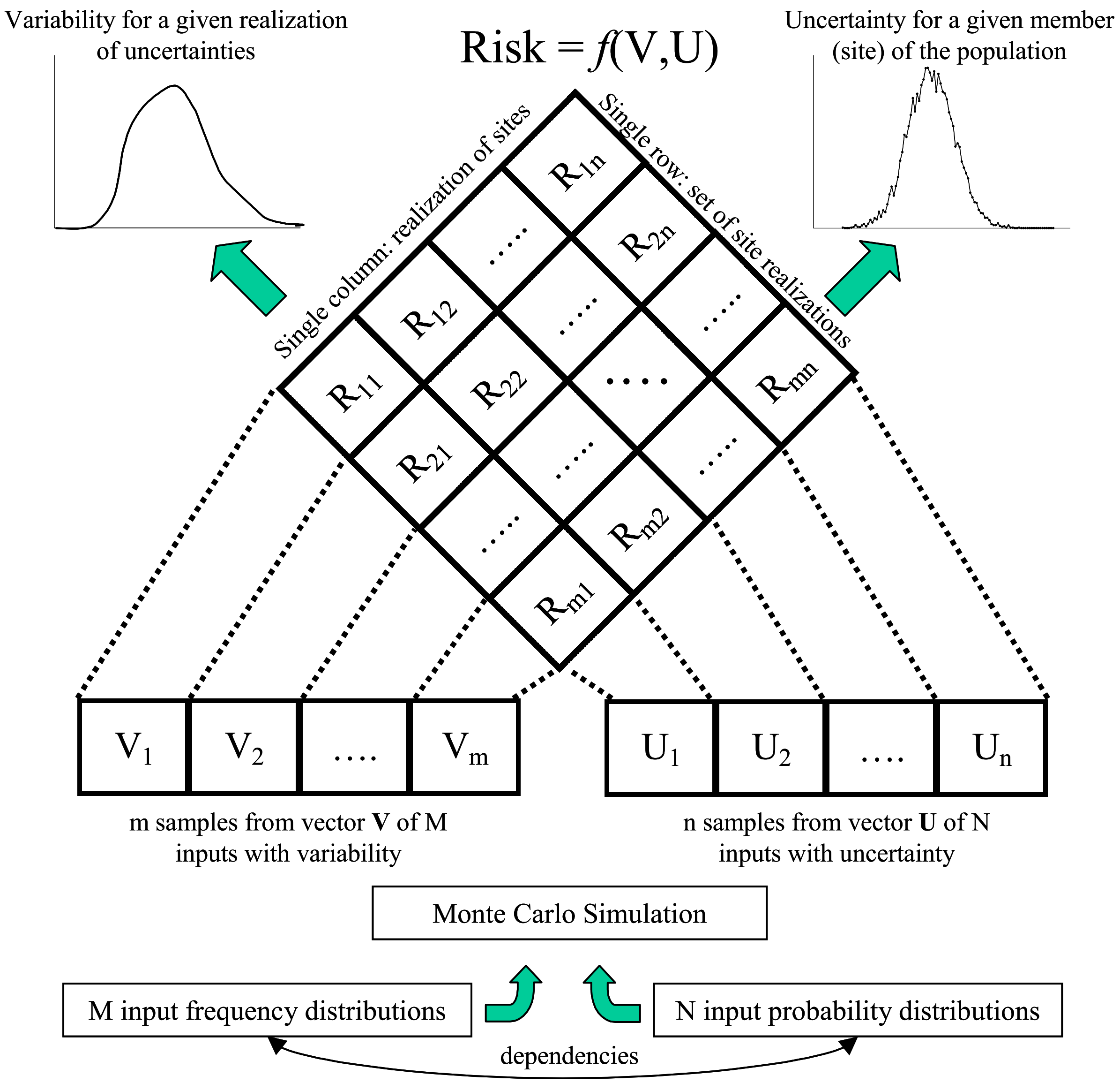

Variability for a given realization of uncertainties
Risk = *f*(V,U)
Uncertainty for a given member (site) of the population
Single column: realization of sites
Single row: set of site realizations
R_{1n}
R_{2n}
R_{12}
R_{11}
R_{22}
R_{21}
R_{mn}
R_{m2}
R_{m1}
V_1
V_2
....
V_m
U_1
U_2
....
U_n
m samples from vector **V** of M inputs with variability
n samples from vector **U** of N inputs with uncertainty
Monte Carlo Simulation
M input frequency distributions
N input probability distributions
dependencies

Figure 2-8. Two-Dimensional Monte Carlo Analysis for 3MRA.
(adapted from Frey and Rhodes, 1996; Cullen and Frey, 1999)

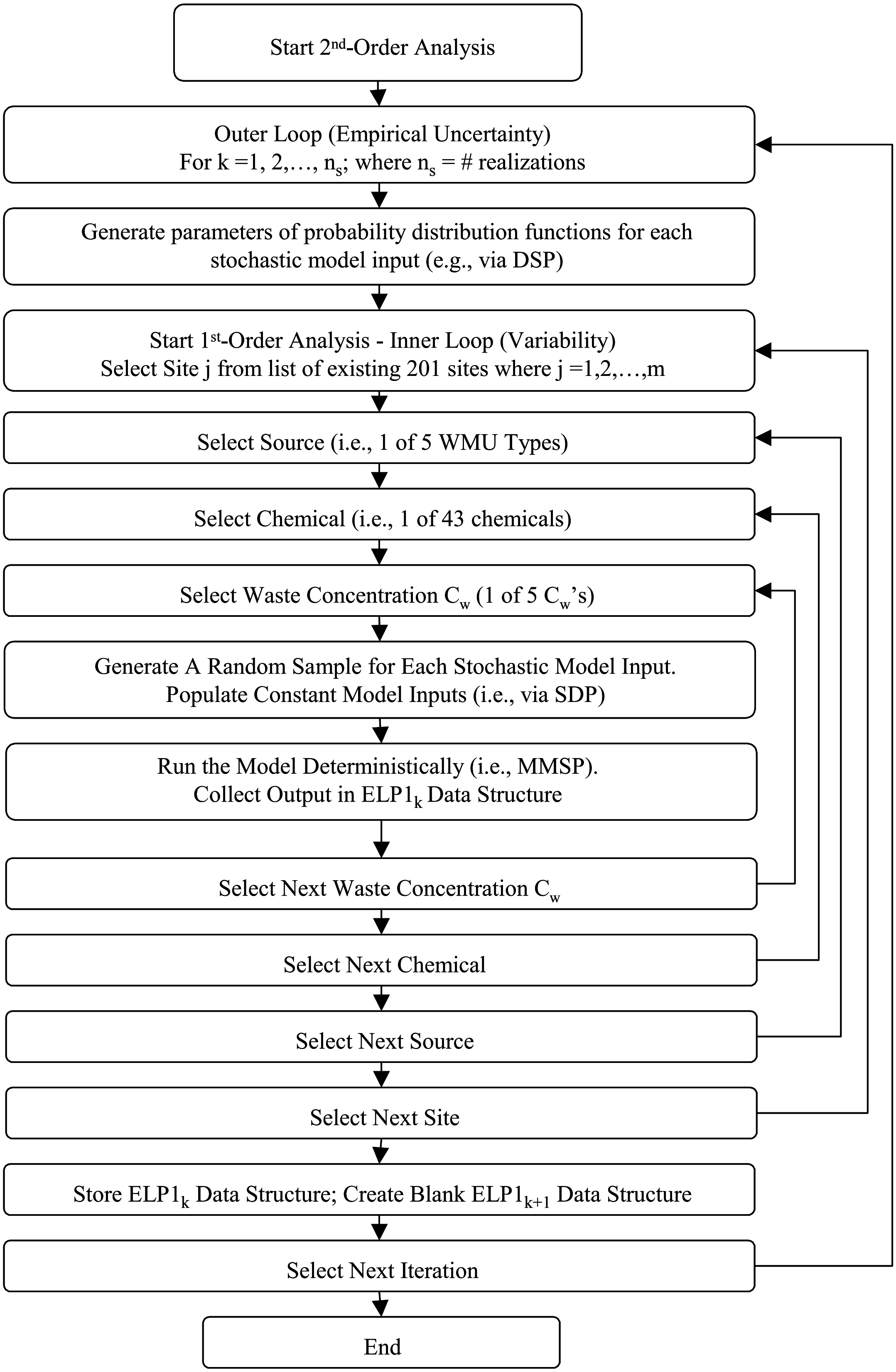
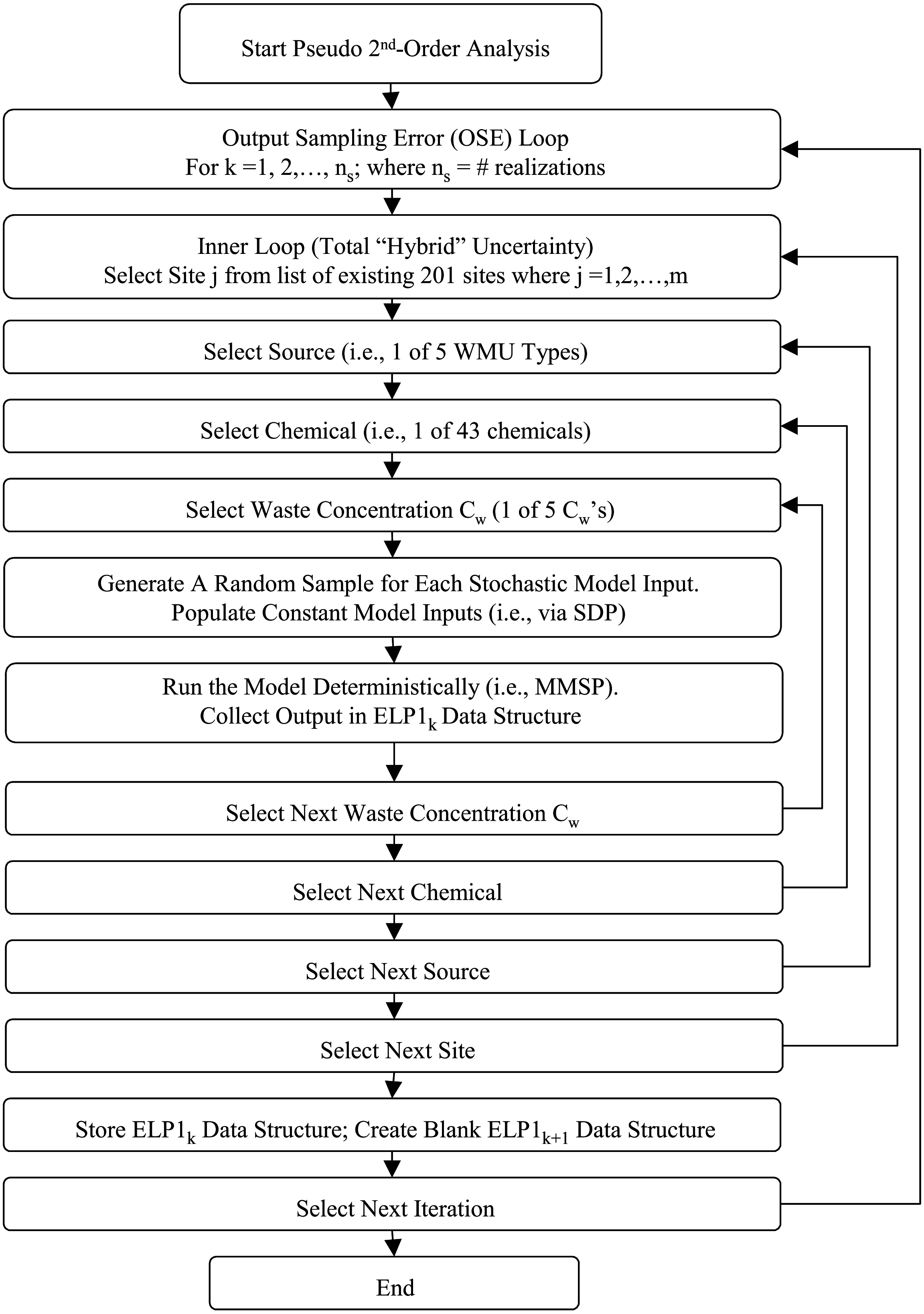
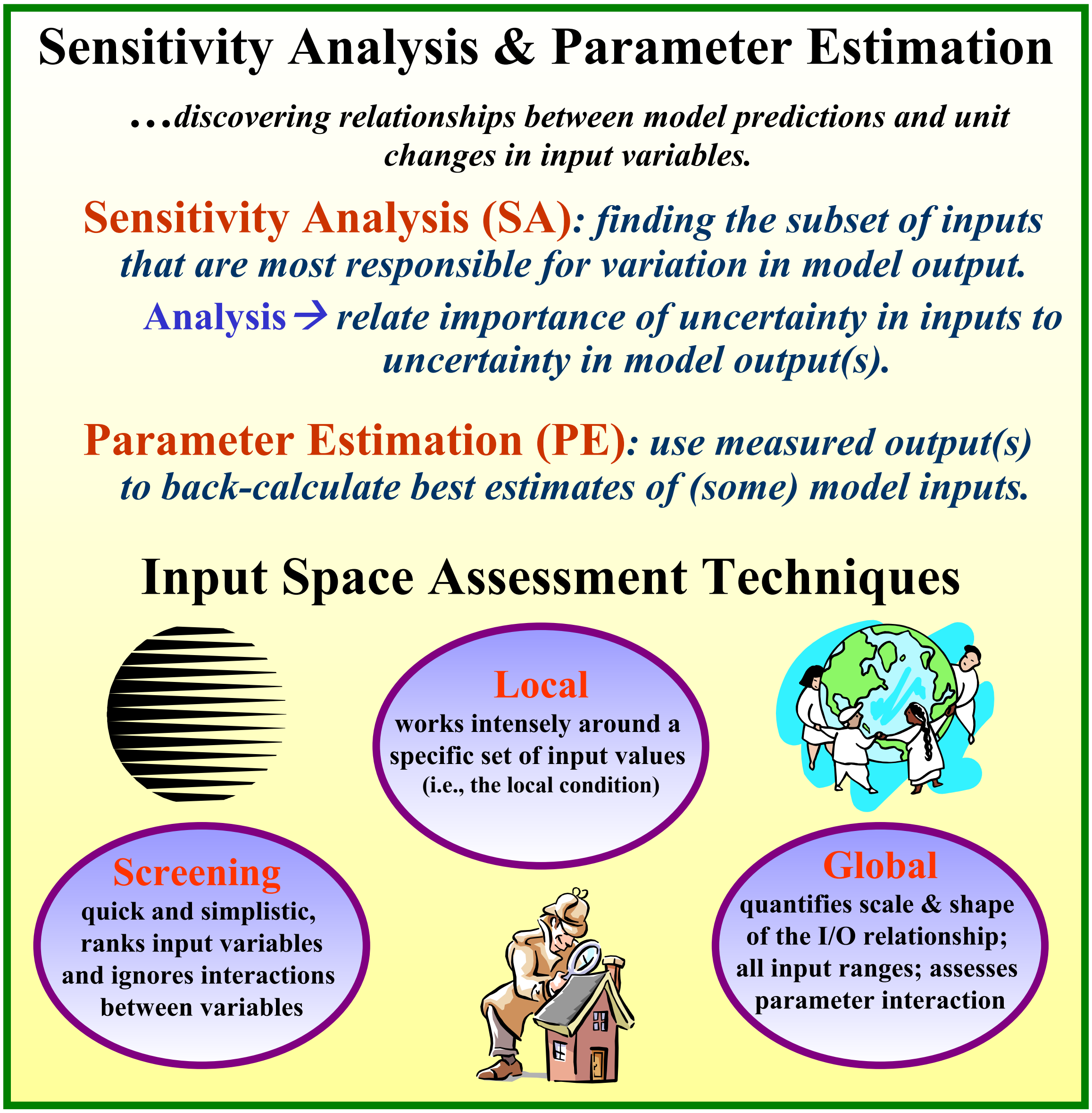

Start 2nd-Order Analysis
Outer Loop (Empirical Uncertainty)
For k =1, 2,…, n_s; where n_s = # realizations
Generate parameters of probability distribution functions for each stochastic model input (e.g., via DSP)
Start 1st-Order Analysis - Inner Loop (Variability)
Select Site j from list of existing 201 sites where j =1,2,…,m
Select Source (i.e., 1 of 5 WMU Types)
Select Chemical (i.e., 1 of 43 chemicals)
Select Waste Concentration C_w (1 of 5 C_w's)
Generate A Random Sample for Each Stochastic Model Input.
Populate Constant Model Inputs (i.e., via SDP)
Run the Model Deterministically (i.e., MMSP).
Collect Output in $ELP1_k$ Data Structure
Select Next Waste Concentration C_w
Select Next Chemical
Select Next Source
Select Next Site
Store $ELP1_k$ Data Structure; Create Blank $ELP1_{k+1}$ Data Structure
Select Next Iteration
End

Figure 2-9. Two-Dimensional Monte Carlo Analysis Flowchart for 3MRA.



Start Pseudo 2nd-Order Analysis
Output Sampling Error (OSE) Loop
For k =1, 2,…, n_s; where n_s = # realizations
Inner Loop (Total "Hybrid" Uncertainty)
Select Site j from list of existing 201 sites where j =1,2,…,m
Select Source (i.e., 1 of 5 WMU Types)
Select Chemical (i.e., 1 of 43 chemicals)
Select Waste Concentration C_w (1 of 5 C_w's)
Generate A Random Sample for Each Stochastic Model Input.
Populate Constant Model Inputs (i.e., via SDP)
Run the Model Deterministically (i.e., MMSP).
Collect Output in $ELP1_k$ Data Structure
Select Next Waste Concentration C_w
Select Next Chemical
Select Next Source
Select Next Site
Store $ELP1_k$ Data Structure; Create Blank $ELP1_{k+1}$ Data Structure
Select Next Iteration
End

Figure 2-10. Pseudo 2nd-Order Monte Carlo Analysis Flowchart for 3MRA.



Sensitivity Analysis & Parameter Estimation

…discovering relationships between model predictions and unit changes in input variables.

Sensitivity Analysis (SA)*: finding the subset of inputs that are most responsible for variation in model output.*

Analysis → ***relate importance of uncertainty in inputs to uncertainty in model output(s).***

Parameter Estimation (PE)*: use measured output(s) to back-calculate best estimates of (some) model inputs.*

Input Space Assessment Techniques

Local
works intensely around a specific set of input values (i.e., the local condition)

Screening
quick and simplistic, ranks input variables and ignores interactions between variables

Global
quantifies scale & shape of the I/O relationship; all input ranges; assesses parameter interaction

Figure 2-11. Conceptualization of Sensitivity Analysis Techniques.