SKI Report 2003:41
SSI report 2003:19

Research

Development of a Quantitative Framework for Regulatory Risk Assessments: Probabilistic Approaches

R. D. Wilmot

November 2003

ISSN 1104-1374
ISSN 0282-4434
ISRN SKI-R-03/41-SE

SKI/SSI perspective

Background

SSI has issued regulations that impose a risk criterion for radioactive waste disposal. SKI has issued corresponding regulations on long-term safety of geological disposal, including aspects and guidance on safety assessment methodology (such as time frames). Based on these regulations, SSI and SKI need to develop an attuned view of what is expected from the applicant, in terms of risk assessment in support of a license application. A previous study for this purpose (SKI ref. no. 14.9-010580/01114 and SSI ref. No. 60/2760/01, SSI P 1288.01) provided qualitative descriptions of various approaches to risk assessment by reference to assessments in other countries.

Relevance for SKI & SSI

One approach to evaluate risk in performance assessments is to use probabilistic approaches to express uncertainties. This report describes the various approaches available for undertaking such probabilistic analyses, both as a means of accounting for uncertainty in the determination of risk and more generally as a means of sensitivity and uncertainty analysis. Also the use of different outputs for presenting probabilistic results is discussed in the report, since communication of results constitutes an important part of an assessment. In addition, the report describes issues that must be considered when characterising and interpreting risk and dose as a function of time.

Results

The objective of this project has been fulfilled. The current project is intended to build upon the previous study and to provide additional support to SSI's and SKI's understanding of risk and how it can be treated in assessments. The specific objective was to review available probabilistic techniques to account for uncertainties, both regarding description of the mathematical basis for different techniques and to survey how different techniques have been used in practise.

Project information

SKI project manager: Eva Simic
Project Identification Number: 14.9-020495/02267
SSI project manager: Björn Dverstorp
Project Identification Number: 624/2265/02, SSI P 1391.03

SKI Report 2003:41
SSI report 2003:19

Research

Development of a Quantitative Framework for Regulatory Risk Assessments: Probabilistic Approaches

R. D. Wilmot

Galson Sciences LTD
5 Grosvenor House
Melton Road
Oakham
Rutland LE15 6AX
United Kingdom

November 2003

This report concerns a study which has been conducted for the Swedish Nuclear Power Inspectorate (SKI) and the Swedish Radiation Protection Authority (SSI). The conclusions and viewpoints presented in the report are those of the author/authors and do not necessarily coincide with those of the SKI and the SSI.

Executive Summary

The Swedish regulators have been active in the field of performance assessment for many years and have developed sophisticated approaches to the development of scenarios and other aspects of assessments. These assessments have generally used dose as the assessment end-point and have been based on deterministic calculations. Recently introduced Swedish regulations [SSI FS 1998:1] have introduced a risk criterion for radioactive waste disposal: the annual risk of harmful effects after closure of a disposal facility should not exceed 10^{-6} for a representative individual in the group exposed to the greatest risk.

A recent review of the overall structure of risk assessments in safety cases concluded that there are a number of decisions and assumptions in the development of a risk assessment methodology that could potentially affect the calculated results. Regulatory understanding of these issues, potentially supported by independent calculations, is important in preparing for review of a proponent's risk assessment.

One approach to evaluating risk in performance assessments is to use the concept of probability to express uncertainties, and to propagate these probabilities through the analysis. This report describes the various approaches available for undertaking such probabilistic analyses, both as a means of accounting for uncertainty in the determination of risk and more generally as a means of sensitivity and uncertainty analysis.

The report discusses the overall nature of probabilistic analyses and how they are applied to both the calculation of risk and sensitivity analyses. Several approaches are available, including differential analysis, response surface methods and simulation. Simulation is the approach most commonly used, both in assessments for radioactive waste disposal and in other subject areas, and the report describes the key stages of this approach in detail. Decisions relating to the development of input PDFs, sampling methods (including approaches to the treatment of correlation), and determining convergence may all affect calculated results. The report discusses the key issues for each stage and how these issues can be addressed in implementing probabilistic calculations and considered in reviews of such calculations.

An important element of an assessment, whichever approach is adopted for undertaking calculations, is the communication of results. The report describes the use of outputs specific to probabilistic calculations, such as probability distribution and cumulative distribution functions, and also the application of general types of output for presenting probabilistic results. Illustrating the way in which a disposal system may evolve is an important part of assessments, and the report describes the issues that must be considered in using and interpreting risk and dose versus time plots.

Contents

Executive Summary .... **iii**

1 Introduction **1**

2 Probabilistic Analysis **3**

2.1 Introduction 3

2.2 Uncertainty and Sensitivity Analysis 3

2.3 Probabilistic Approaches 5

2.3.1 Simulation 6

2.3.2 Differential analysis 7

2.3.3 Response surface methodology 8

2.3.4 Fourier amplitude sensitivity test (FAST) 9

3 Simulation **11**

3.1 Introduction 11

3.2 Developing Input PDFs 12

3.3 Sampling Methods 15

3.3.1 Monte Carlo sampling 15

3.3.2 Stratified or Importance Sampling 16

3.3.3 Latin Hypercube Sampling 17

3.4 Correlation 19

3.5 Convergence 22

3.6 Timing of Events 24

4 Outputs and their Interpretation **27**

4.1 Introduction 27

4.2 Numerical Output 27

4.3 Graphical Output 30

4.3.1 Probability Distribution Function (PDF) 31

4.3.2 Cumulative Distribution Function (CDF) 32

4.3.3 Complementary Cumulative Distribution Function (CCDF) .. 32

4.3.4 Box-whisker plots 33

4.3.5 Scatter plots 33

4.3.6 Dose / risk vs time 34

5 References....36

6 Figures....38

Appendix 1 Probabilistic Approaches....51

Development of a Quantitative Framework for Regulatory Risk Assessments: Probabilistic Approaches

1 Introduction

The responsibility for regulation of radioactive waste management and disposal in Sweden is shared between the Swedish Nuclear Power Inspectorate (SKI) and the Swedish Radiation Protection Authority (SSI). Recently introduced Swedish regulations [SSI FS 1998:1] impose a risk criterion for radioactive waste disposal: the annual risk of harmful effects after closure of a disposal facility should not exceed 10^{-6} for a representative individual in the group exposed to the greatest risk. The regulation and the accompanying guidance indicate that the regulatory authorities require a consideration of both consequences (doses) and the probability of receiving a dose to be considered in assessments.

The Swedish regulators have been active in the field of performance assessment[1] for many years and have developed sophisticated approaches to the development of scenarios and other aspects of assessments. These assessments have generally used dose as the assessment end-point. The recent introduction of a risk criterion has, therefore, required an examination of the implications of a change in end-point on the type of calculations conducted and the structure of the assessment.

As part of this examination of the implications of a risk criterion, the overall structure of risk assessments in safety cases to meet risk targets has been reviewed (Wilmot, 2002). One approach to evaluating risk in such assessments is to use the concept of probability to express uncertainties and to propagate these probabilities through the analysis. This report describes the various approaches available for undertaking such probabilistic analyses, both as a means of accounting for uncertainty in the determination of risk and more generally as a means of sensitivity and uncertainty analysis.

Following this Introduction, Section 2 of the report discusses the overall nature of probabilistic analyses and how they are applied to both the calculation of risk and sensitivity analyses. The approach most commonly used, both in assessments for radioactive waste disposal and in other subject areas, is simulation[2]. Section 3 of the report describes the stages of this approach and discusses the key issues that must be considered in implementing and reviewing this type of calculation. Whatever type of approach is adopted for undertaking analyses, the results must be communicated.

[1] The term performance assessment is used in a generic sense in this report to cover all approaches to assessing the long-term behaviour of a facility. The term risk assessment is used in a more specific sense to cover assessments that use risk as a measure of performance.

[2] The term "simulation approach" has been used elsewhere to refer specifically to the approach used by HMIP to model environmental evolution and develop time-dependent boundary conditions for the assessment model. This is a confusing use of the term and "simulation approach" is used here in its more general sense.

There are some specific aspects of probabilistic analyses to be considered in developing outputs from these calculations and these are discussed in Section 4. References are provided in Section 5 of the report and the figures are presented in Section 6.

Appendix 1 lists the main performance assessments that have used probabilistic calculations.

2 Probabilistic Analysis

2.1 Introduction

An earlier report (Wilmot, 2002) describes how the concept of probability can be used as a way of expressing uncertainty. Uncertainties can arise either because there is an objective uncertainty or randomness in the system, or because there is imperfect knowledge about the system. There can be some benefit in distinguishing between these two types in the analysis and presentation of uncertainties. However, the mathematical approaches to treating uncertainties that are expressed as probabilities are essentially the same whatever the source or characteristics of the uncertainties.

This report provides a discussion of approaches used to analyse uncertainties expressed as probabilities. Expressing uncertainties in the form of probabilities means that some form of probabilistic approach is likely to be used for *uncertainty propagation* i.e., calculating the uncertainty in model outputs induced by the uncertainties in the inputs. Similarly, once a decision has been made to express uncertainties as probabilities, then a probabilistic approach will likely be the most efficient method for *uncertainty analysis*, i.e., comparing the importance of the input uncertainties in terms of their relative contributions to uncertainty in the outputs. Finally, whatever means is used to express uncertainties, probabilistic approaches can be used for *sensitivity analysis*, i.e., assessing the effect of changes in inputs on model predictions.

Although different in detail, the probabilistic approaches used for these three steps are similar in principle. Several approaches have been developed and are applicable to systems of different complexities, although performance assessments that use probabilistic approaches predominantly use a simulation approach.

In the following section, a brief outline of the principles behind sensitivity and uncertainty analyses is presented. This is followed in Section 2.3 by more detailed descriptions of different approaches. Because the simulation approach is the most commonly used in performance assessments, an extended discussion of this approach and of the issues involved is presented in Section 3.

2.2 Uncertainty and Sensitivity Analysis

Whatever the system under analysis, the aim of any assessment or modelling study is to determine some output (y) as a function of a number of inputs (x_1, x_2, …). In its most basic form, the system can be expressed as:

$$y = f(X) \tag{1}$$

where: $X = (x_1, x_2, \ldots x_n)$

If each of the inputs has a single value, then the output will be a single value. However, if any of the inputs has an associated uncertainty, then the output will also be uncertain. If the uncertainties in the inputs are expressed in the form of probability distribution functions (PDFs), then the output will be an *m*-dimensional response surface, where m of the n inputs are uncertain. As an illustration, Figure 1 shows a response surface for the system $y = f(x_1, x_2)$, where both x_1 and x_2 have a range of values. The output can also be expressed as a PDF of values for y, from which an expectation value and other statistical descriptors can be calculated if required.

Sensitivity is the rate of change of y with respect to a change in an input x. In the simple example illustrated in Figure 1, there are two sensitivities, one with respect to x_1, and one with respect to x_2. These sensitivities can be evaluated at any point on the response surface, but in general they are evaluated at the best estimate value for each of the inputs (e.g., at the mean, median or mode of the respective PDFs). If this best estimate or nominal value is designated X^0, then the sensitivities are the partial derivatives of output y, with respect to each input at this point:

$$\left[\frac{\partial y}{\partial x_1}\right]_{X^0}, \left[\frac{\partial y}{\partial x_2}\right]_{X^0} \tag{2}$$

A disadvantage of this simple measure of sensitivity is that it is dependent on the units of the variables involved. The sensitivity to x_1 will be a thousand times greater if the units are millimetres than if they are metres. This disadvantage can be overcome by normalising the sensitivities and expressing them as relative values:

$$\left[\frac{\partial y}{\partial x_1}\right]_{X^0} \times \frac{x_1^0}{y_0}, \left[\frac{\partial y}{\partial x_2}\right]_{X^0} \times \frac{x_2^0}{y_0} \tag{3}$$

These measures of sensitivity take no account of the uncertainties in the input, although a variable with a low sensitivity but large uncertainty may be as important as a variable with a greater sensitivity but smaller uncertainty. The contribution of a variable to the overall uncertainty can be expressed as the product of its sensitivity and standard deviation:

$$\left[\frac{\partial y}{\partial x}\right]_{X^0} \times \sigma_x \tag{4}$$

The Gaussian approximation[3] can then be used to estimate the uncertainty of the output in terms of its variance:

$$\mathrm{Var}[y] \approx \left(\left[\frac{\partial y}{\partial x_1}\right]_{X^0}^2 \mathrm{Var}[x_1]\right) + \left(\left[\frac{\partial y}{\partial x_2}\right]_{X^0}^2 \mathrm{Var}[x_2]\right) \tag{5}$$

where $\mathrm{Var}[x_1] \equiv \sigma_{x_1}^2$ and $\mathrm{Var}[x_2] \equiv \sigma_{x_2}^2$.

Each of these measures can be extended to account for larger numbers of inputs. Whatever the number of input parameters involved, however, these measures all remain *local* measures of uncertainty around the nominal value of the output (y_0). To

[3] The Gaussian approximation states that the variance of the output can be reliably estimated by the sum of the squares of the contribution from each input.

take account of the behaviour of the full range of the output, it is necessary to use techniques that involve the full range of the input variables.

One approach that takes more account of parameter uncertainty than the above local measures uses a 'high' and a 'low' value for each parameter. These need not be the extremes of the distribution, but should encompass the central part of the distribution. The "nominal range sensitivity" is then calculated by varying each parameter from its high to low value, while keeping all other parameters at their nominal value. If the low and high values for the two parameters are denoted by $[x_1^-, x_1^+]$ and $[x_2^-, x_2^+]$, then the nominal range sensitivities are:

$$U(x_1, y) = f(x_1^+, x_2^0) - f(x_1^-, x_2^0) \quad \text{and} \tag{6}$$

$$U(x_2, y) = f(x_1^0, x_2^+) - f(x_1^0, x_2^-) \tag{7}$$

Although these measures are more than local, they are less than global because they hold all but one of the parameters constant. In many systems, the effects of varying one input are dependent on the values of other parameters. In these cases, a joint parametric analysis, evaluating *y* for several values of the other parameters, is useful. The case illustrated above with two parameters and two values requires four evaluations of the output. Increasing the number of values and the number of parameters quickly increases the number of evaluations required using this approach. Relationships within the resulting data can be identified by calculating correlation coefficients between different inputs and the output. However, a visual examination of scatter plots (see Section 4.3.5) is also of value, because there are many ways in which distributions with a low correlation coefficient can nevertheless have a significant relationship. For example, situations where there are thresholds or discontinuities may not be identified without a visual examination of the data (see Figure 2).

The following section describes some of the approaches that can be used to decrease the number of evaluations required for systems that are more complex than the examples presented in this section.

2.3 Probabilistic Approaches

As soon as an assessment requires more information than can be obtained from a single evaluation of the output based on a set of input parameter values, it is appropriate to consider ways in which multiple evaluations can be undertaken effectively. For simple systems, with a small number of inputs, parameter distributions defined using simple functions, no correlations between input parameters, and easy to evaluate relationships, the direct evaluation of all the required output values may be the most effective approach. It may be possible to evaluate parts of such systems analytically. As the complexity increases, with more parameters, correlations, discontinuities and complex interactions, and more complex parameter distributions, each individual evaluation requires more effort. The computational effort required to perform large numbers of evaluations for such systems may become unrealistic. This has led to the development of alternative approaches to evaluating the output values and/or to selecting the combinations of parameter values at which to undertake the calculations. A number of such approaches are described in this section.

The most familiar approach to evaluating the behaviour of system models for radioactive waste disposal in which uncertainties are expressed in terms of probability distribution functions is the simulation approach. In essence, this approach simply evaluates the output for a subset of all the possible combinations of parameter values.

There are alternatives to the simulation approach for propagating uncertainties (Helton, 1993):

- Differential analysis
- Response surface methodology
- Fourier amplitude sensitivity test (FAST)

Although these methods are not currently in use in performance assessments, it is useful to review them briefly and to highlight the reasons why the simulation approach has been preferred.

2.3.1 Simulation

All of the approaches that are used to propagate uncertainty from the uncertainty in the inputs require evaluation of the output for specific sets of input values. In some of the approaches described below, the values selected for evaluation form a regular grid or are otherwise pre-selected. This means that the calculated output values cannot necessarily be used to estimate the properties of the overall output distribution. Instead, they are used to develop an approximation or surrogate form of the output which is then used for uncertainty propagation and sensitivity analysis. The simulation approach does not necessarily have this disadvantage and can provide an estimate of the output distribution directly.

In the simplest form of the simulation approach, values are selected at random from the input distributions. The output values calculated from these randomly selected values then form a random subset of all possible output values. Because the calculated output distribution is effectively a sample drawn at random from the overall distribution, the properties of the calculated distribution can be used to estimate the properties of the overall distribution using standard statistical techniques.

There are a number of issues that must be considered in the practical implementation of a simulation approach. These include:

- The definition and use of PDFs to define the input distributions.
- Methods used for sampling.
- Accounting for correlations between input parameters
- Determining how many samples are required.

These issues are discussed in more detail in Section 3.

2.3.2 Differential analysis

Differential analysis is a method for approximating the full model using a Taylor series, and then using this series in place of the full model for sensitivity and uncertainty analyses.

The Taylor series is developed at the nominal value X^0, and expresses deviations of the output from the nominal value ($y - y^0$) in terms of deviations of the inputs from their nominal value ($x_i - x_i^0$). Successive terms of the series include higher order deviations and partial derivatives. The second-order series has the form:

$$y - y^0 = \sum_{i=1}^{n} (x_i - x_i^0) \left[\frac{\partial y}{\partial x_i} \right]_{X^0} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_i^0)(x_j - x_j^0) \left[\frac{\partial^2 y}{\partial x_i \partial x_j} \right]_{X^0} \tag{8}$$

This can be a good approximation to the model if the deviations ($x_i - x_i^0$) are relatively small and the function is relatively smooth (i.e., the higher derivatives are small).

The Taylor series can be used for uncertainty propagation, although estimates of the tails of the distribution may not be reliable. As an example, the expectation value for y from a second order series is given by:

$$\mathrm{E}(y) \approx y^0 + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}[x_i, x_j] \left[\frac{\partial^2 y}{\partial x_i \partial x_j} \right]_{X^0} \tag{9}$$

where Cov[x_i, x_j] is the covariance.

If the inputs are uncorrelated, then this reduces to:

$$\mathrm{E}(y) \approx y^0 + \frac{1}{2} \sum_{i=1}^{n} \left[\frac{\partial^2 y}{\partial x_i^2} \right] \mathrm{Var}(x_i) \tag{10}$$

and the variance can be estimated by:

$$\mathrm{Var}(y) \approx \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]^2 \mathrm{Var}(x_i) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right] \left[\frac{\partial^2 y}{\partial x_i^2} \right] \mu_3(x_i) \tag{11}$$

where $\mu_3(x_i)$ is the third central moment for x_i (see Section 4.2).

A key point from these estimates is that the expectation value for the output is not equal to the nominal output (i.e., the output calculated using the nominal values for all inputs), unless the model is linear. In the more usual case of a non-linear model, the expectation value is also a function of the variances and covariances of the inputs.

The advantages of differential analysis are that it provides a clear approach to uncertainty analysis, with the variance of the output decomposed to the sum of the contributions from each input. Also, the numerical analyses required are generally simple. The disadvantages are that the derivation of the series can be complex, particularly as higher order terms are included to achieve a satisfactory approximation to complex functions. It is also a local approach to uncertainty analysis and will not be accurate if the model has discontinuities or large uncertainties.

2.3.3 Response surface methodology

The response surface approach involves fitting an approximate response surface to a moderate number of calculated output values, and then using this approximation for uncertainty propagation and analysis. The technique is of value where the computational resources required for evaluating the full model are very high.

Key steps in using the response surface approach are identifying both the key input parameters and the combinations of values for these parameters that will be used to calculate points on the response surface. Parameters to which the response surface is less sensitive can be set to their best estimate or nominal values. A number of experimental design procedures can be used to generate efficient combinations of values. Factorial designs are based on selecting two or more values for each parameter and then using all possible combinations of these values. For a two-level design (i.e., with a ‘high’ and a ‘low’ value for each parameter) with *k* parameters, a full factorial design will involve 2^k combinations, a number that can become very large for complex systems. Fractional factorial designs, which use a given fraction (e.g. 1/10) of all possible combinations, are more efficient. Care is required in selecting a design for systems involving non-linearities and interactions between inputs so as not to reduce the number of combinations to such an extent that the response surface is not sufficiently well defined. Monte Carlo methods can also be used to define the combinations used to generate the response surface. The issues that apply to this approach are similar to the issues concerning the use of Monte Carlo methods for directly determining the output distribution (see Section 3).

Once the combinations of parameter values (design points) have been selected, the full model is used to calculate the output values at these points. A response surface is fitted to these output values using techniques such as least squares or spline functions. For relatively simple systems, linear or quadratic models may be adequate. Higher order models can be used either to provide a better fit for relatively simple systems or an adequate fit for more complex systems. Regression coefficients are a useful means of assessing how well the response surface fits the output values, but can be misleading if the number of design points is comparable with the number of degrees of freedom for the response surface. Determining how well the response surface matches output values for additional design points located at the extremes of the system is a useful way of assessing its adequacy.

Once a response surface has been established, it can be used for uncertainty propagation and sensitivity analysis. For a linear response surface of the form:

$$y = b_0 + \sum_{j=1}^{n} b_j x_j \tag{12}$$

the expectation value and variance can be estimated by:

$$\mathrm{E}(y) \approx b_0 + \sum_{j=1}^{n} b_j \, \mathrm{E}(x_j) \tag{13}$$

and

$$\mathrm{Var}(y) \approx \sum_{j=1}^{n} b_j^2 \, \mathrm{Var}(x_j) + 2\sum_{j=1}^{n} \sum_{k=j+1}^{n} b_j b_k \, \mathrm{Cov}(x_j, x_k) \tag{14}$$

More complex relationships can be used to determine these values for more complex response surfaces. The alternative approach is to use Monte Carlo simulation with the equation for the response surface. Even with a quadratic or higher order surface, the computation requirements for this simulation will be low and large numbers of samples can be used, giving a good estimate of the distribution function for the response surface.

Sensitivity studies can be based on the response surface, using either analytical methods to calculate normalised sensitivity measures for each parameter, or graphical examination of the results from simulations.

The disadvantages of the response surface approach are related to the effort required to determine the response surface and to the difficulty in defining a response surface that adequately matches the output from a complex model. Typically, the complexity of the system-level models used in performance assessments for radioactive waste disposal means that the output function is too complex to be easily approximated by a response surface. In particular, response surfaces are difficult to implement satisfactorily where there are discontinuities or thresholds in the model output.

Although not generally used for system-level modelling, a response surface approach has been used in several programmes for sub-system analyses. In the WIPP PA, for example, creep closure of the repository is accounted for in assessment calculations by changing the porosity of the waste disposal area. The SANTOS code uses finite element methods to calculate porosity as a function of gas pressure, and a series of porosity time histories is calculated based on a set of thirteen different gas generation potentials. This modelling results in a three-dimensional response surface representing changes in gas pressure and porosity over the 10,000-year simulation period. In the system calculations, the porosity corresponding to the calculated gas pressure and fluid saturations is interpolated from this response surface.

2.3.4 Fourier amplitude sensitivity test (FAST)

The Fourier amplitude approach to sensitivity and uncertainty analysis is based on a transformation of the multi-dimensional integral over all model inputs to a one-dimensional integral which is easier to evaluate. This transformation is achieved by defining a curve in model space, and a Fourier series representation of this curve is then used to estimate the contribution of each input variable to the variance of the model output.

The general form of the one-dimensional integral used to estimate the expectation value for the general model $y = f(X)$ is:

$$\mathrm{E}(y) \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} f\left[G_1(\sin(\omega_1 s), G_2(\sin(\omega_2 s), \ldots G_n(\sin(\omega_n s)\right]\mathrm{d}s \tag{15}$$

where G_1, …, G_n and $\varpi_1, \ldots, \omega_n$ are series of functions and integers respectively. The variance can be similarly estimated:

$$\mathrm{Var}(y) \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} f\left[G_1(\sin(\omega_1 s), G_2(\sin(\omega_2 s), \ldots G_n(\sin(\omega_n s)\right]\mathrm{d}s - \mathrm{E}^2(y) \tag{16}$$

The advantages of the FAST approach are that it is global, covering the full range of input distributions, and that a surrogate model is not required. The principal disadvantages are the difficulties in deriving the curve in parameter space, the inability to take account of correlations between inputs, and the lack of information about discontinuities or thresholds in the output.

There are no examples of the use of this approach in performance assessments for waste disposal facilities, and its disadvantages mean that it is unlikely to be used for system analyses. It may, however, have a role in some sub-system analyses.

3 Simulation

3.1 Introduction

Simulation is the most common method used for probabilistic calculations, not only in performance assessments for radioactive waste disposal, but also in other sectors. A principal reason for this is that the concept is easy to understand and does not involve any complex mathematics. There is sometimes resistance to adopting a probabilistic approach because of the perceived difficulties of expressing uncertainty as a probability, but the concept of repeating a deterministic calculation many times with different parameter values is a simple one.

Although simple in principal, the application of the simulation approach to a complex problem comprises several stages and requires a number of decisions to be made. The principal stages are:

- Identification of key parameters and uncertainties
- Identification of correlations between parameters
- Model development (conceptual, mathematical and computational)
- Definition of probability distribution functions (PDFs)
- Sampling
- Calculation, and repetition of the calculation a sufficient number of times
- Presentation and analysis of results

During the development, licensing and operation of a disposal facility, a number of assessments will be undertaken. There will be iterations of these principal stages both within an assessment and between successive assessments. In particular, the analysis of results will help to identify the key parameters and where there would be most benefit in reducing uncertainty in subsequent iterations.

The scope of this report does not include a detailed description of all the stages of an assessment. Instead, it focuses on those aspects that are unique to probabilistic calculations. Model development is not discussed, because in principle the same models can be used for deterministic and probabilistic calculations. In practice, different models may be required, both to reduce the computational burden of running a complex model several hundred times, and also because a model used for probabilistic calculations must be robust over a wider range of conditions than a deterministic model applicable to a single set of conditions.

The key areas discussed in this section are the definition of PDFs, sampling, correlation and convergence. The definition of scenarios to be analysed, and specifically the use of sampling to define these, is also discussed. Section 4 describes the presentation and analysis of results.

3.2 Developing Input PDFs

Identifying parameters

The aim of probabilistic calculations is to take account of uncertainties. In the complex systems that are assessed in performance assessments, there will be uncertainties about many aspects of the system and system behaviour. Whether all of these uncertainties should be addressed is a strategic decision that is dependent upon the resources available and the purpose of the assessment (the assessment context). An assessment solely for design optimisation may, for example, account explicitly for a different set of uncertainties than an assessment aimed at developing a license application.

Any level of uncertainty in parameter values will have some effect on the overall calculated result, but in practise there will be a restricted set of uncertainties that will dominate the overall level of uncertainty. These key uncertainties may result from large uncertainties in the input values, or from sensitivities in the models that make the overall result sensitive to particular parameters. Where there are both large data uncertainties and significant system sensitivities, the final result may be dominated by a very few parameters.

Significant resources can be expended on developing PDFs if processes such as expert elicitation are required or additional site characterisation or experimental programmes are undertaken. There may be other reasons for undertaking additional data collection, such as confidence-building, but it is sensible to have an understanding of the key sensitivities before undue effort is expended on defining PDFs that do not have a significant effect on the calculated result. Paradoxically, the most appropriate tool for determining sensitivities is a probabilistic calculation. A key issue is therefore how to identify sensitivities prior to undertaking the calculations that would help to identify them.

Model sensitivity is not, however, the only criterion for determining the effort or resources that should be expended on defining parameter PDFs. A risk assessment model may be sensitive to parameters such as radionuclide half-life or inventory, for example, but these are known or well constrained and would not normally be input as a PDF. The issue, therefore, is to identify those parameters to which the model is sensitive and for which there are large uncertainties.

The resolution of this paradox lies in the iterative nature of the assessment process. This means that early assessments can use approximations to define levels of uncertainty without significant resources, and the results from these can be used to identify potentially important parameters. Further studies will then allow a better definition of the uncertainties for these parameters. Later assessments can focus on the key uncertainties, and work programmes to reduce the level of uncertainty can be effectively targeted.

Selecting distribution types

There is a large literature on identifying the most appropriate distribution to use in defining PDFs, and on assessing or optimising the "goodness-of-fit". In studies where the intention of the simulation is to accurately reproduce conditions for which there are a large number of observations and measurements, obtaining the closest possible match is appropriate. However, in the case of elicited data, or where there are comparatively few observations to define variability or uncertainty, simple distributions that have the same broad characteristics as the data are likely to be adequate, at least in the initial stages of the assessment process. Simple distributions are intuitive and their form can be described in terms familiar to non-statisticians. Figure 3 shows, for example, how a triangular distribution (defined by maximum, minimum and mode) can be substituted for a normal distribution (defined by mean and standard deviation).

There are three main distribution types used in performance assessment programmes, together with their log-transformed equivalents[4]:

- Uniform and log-uniform
- Triangular and log-triangular
- Normal and log-normal

Most assessments also have provision for defining a PDF using data pairs rather than a function. This cumulative or empirical distribution allows any data to be used and is useful in cases such as bi-modal data or sparse data that cannot easily be fitted by a function.

In addition to this basic set of distribution types, other distribution types are used for specific purposes. The Beta distribution has been used or proposed in several assessments. It has the benefit of being extremely flexible (Figure 4), but has the disadvantage of not being intuitively related to the data. The WIPP CCA (Appendix PAR) uses a Delta distribution to sample between different conditions or alternative models. The recent SKB report on developing PDFs (Mishra, 2002) uses the Poisson and Weibull distributions to account for data on canister failures.

SYVAC, the assessment code used for the Canadian performance assessment, also allows parameter values to be determined from an explicit equation and a sampled residual error parameter. For example, sorption coefficients for minerals are calculated from an equation fitted to observational data, multiplied by a random error factor to represent the uncertainty in the fitted equation. The error factor is sampled from a log-normal PDF with geometric mean 1.0, geometric standard deviation $\sqrt[3]{10}$, and truncated at 0.1 and 10 ($\pm 3\sigma$).

The effects of changing the type of distribution used to characterise uncertainty can be determined mathematically. Calculating the significance of such effects and whether they would significantly affect calculated doses and risks is more difficult because of the interactions and non-linearities present in models of complex systems.

Defining PDFs

There is an overlap between the task of determining which distribution type to use for a particular PDF and defining the values that describe the distribution (maximum,

[4] A logarithmic distribution is generally more appropriate for parameters where the difference between the minimum and maximum values is greater than an order of magnitude.

minimum, etc.). As already noted, initial estimates of the maximum, minimum and mean or best-estimate values for a parameter can be made with significantly less resources than required for a comprehensive data review and/or elicitation exercise to define a PDF. Using such initial estimates, probabilistic calculations can be undertaken and sensitivity studies used to identify those parameters for which detailed studies aimed at better defining uncertainties are justified.

Methods for performing sensitivity studies are described in Section 2.2. In the context of deciding which parameters should be examined in detail, it should be noted that two types of sensitivity may be important. Results may be sensitive to parameters that have a linear relationship with the output but which have a wide variation from minimum to maximum. Results may also be sensitive to parameters where there is a much smaller range of uncertainty but which have a non-linear relationship with the output. In these cases, the key sensitivity will probably be to the selected maximum value (or minimum if the relationship is an inverse one). Parameters that display this type of sensitivity to the exact form of the distribution are particularly important to identify because confidence in the overall result may depend on the justification presented for the selected distribution.

In the initial stages of an assessment, the principal method for defining PDFs comprises a provisional analysis of available information and expert judgement. As discussed in Wilmot and Galson (2000), it is important that these judgements are acknowledged and documented even in the early stages because they may otherwise be “lost” if they are incorporated into the later stages of the assessment. Once key parameters have been selected and more formal methods for defining PDFs are adopted, there are two methods available. The first involves a detailed review of the available data, expert judgement as to the reliability and relevance of this information, and fitting a distribution to the data by varying the characteristics of the distribution. The second approach involves expert elicitation to define the shape of the PDF and then fitting a distribution (Wilmot et al., 2000; Hora and Jensen, 2002). In both cases, empirical distributions can be used if the distribution is complex.

A key concern when eliciting distributions is the phenomenon known as anchoring, whereby experts focus on a narrow range of values and underestimate the uncertainty. Facilitators therefore encourage experts to think carefully about circumstances that may give rise to larger or smaller values than their initial estimates. A similar underestimate of uncertainty can arise if experimental data used to develop PDFs is too restrictive and does not correspond, for example, to a wide enough range of physical and chemical conditions. Underestimating parameter uncertainty by defining PDFs that are too “narrow” will lead to an underestimate of uncertainty in the overall performance measure (dose or risk).

Because there is an acknowledged tendency toward anchoring and underestimating uncertainty, there may be a tendency for analysts to extend the range of PDFs in order to compensate. However, if this is done on an *ad hoc* basis, rather than being justified by documented information or reasoning, it can contribute to a phenomenon known as risk dilution. This is the paradoxical situation in which an increase in uncertainty about contributing factors, which should lead to caution, leads to a decrease in calculated risk and hence makes the system appear “safer”.

Significant risk dilution arising from an over-estimate of parameter uncertainty could occur, but it requires that the additional uncertainty is applied only to the tail of the distribution that lowers the calculated risk. An increase in the other tail, or a symmetrical increase, could in fact lead to "risk amplification" and over-estimate calculated risks. This could be an issue if it is added to other conservatisms in the analysis and suggested that the risks were above regulatory constraints.

3.3 Sampling Methods

Once the parameters to be sampled have been determined and appropriate PDFs defined, probabilistic simulations can be undertaken. Each simulation corresponds to a deterministic calculation, but with parameter values sampled from PDFs rather than being defined *a priori*. There are three methods that can be used for sampling from PDFs:

- Monte Carlo or random sampling
- Stratified or Importance Sampling
- Latin Hypercube Sampling (LHS)

3.3.1 Monte Carlo sampling

Although random samples can be generated directly from some types of distribution, the computational algorithms used to generate random (or more correctly, pseudo-random) numbers generally give numbers in the range 0 - 1. These can then be mapped to the complementary distribution function for a particular parameter to yield a random value for that parameter. This method is illustrated in Figure 5.

If sufficient random samples are made, then the resulting distribution of parameter values will approach the sampled PDF. However, the number of samples required to ensure that all regions of the PDF are adequately represented can become large if there are low-probability regions in the distribution. Because these low-probability regions generally correspond with the extreme values (tails) of the distribution, it can be important that they are sampled if the full range of system behaviour is to be explored.

In order to be 99% certain that at least one sample lies above the 95th percentile of a distribution, the number of samples required, N, is such that:

$1 - 0.95^{N} > 0.99$

which gives a value of 90 samples. To provide the same confidence that at least one sample lies above the 99th percentile, some 459 samples would be required.

As the number of sampled parameters increases, there are two approaches to determining the number of samples required. The first approach is to use the same calculation as above to determine how many samples are required to provide confidence that one or more calculated values exceed the required percentile. On this basis, and provided that the calculated values are independent, the number of samples required is independent of the number of sampled parameters or their distributions.

This independence of the number of samples required from the number of parameters sampled arises because this calculation assesses the distribution of calculated values and not the distributions of sampled values. For example, consider a distribution calculated by multiplying sampled values from two uniform distributions (Figure 6). The 95th percentile of this distribution would correspond to samples lying at the 78th percentiles of the sampled distributions:

$(1 - 0.78) * (1 - 0.78) = 0.048$

This calculation will vary according to the types of distribution involved and how parameters are used in the calculations. However, in general, as the number of parameters increases, the more likely it becomes that the maximum calculated value will lie above a selected percentile without necessarily sampling the extremes of the parameter distributions. As noted above, this could mean that parts of the distributions that could lead to high consequences are not sampled.

The second approach to determining the required number of samples is based on ensuring that the full range of system behaviour is sampled. Again, the exact number of samples required will depend on the types of distribution and the calculation, but the approach can be illustrated using the same example as above. To provide 99% confidence that a pair of sampled values come from above the 95th percentile of their respective distributions, the number of samples required, N, is such that:

$1 - [1 - 0.05^{2}]^{N} > 0.99$

which gives a value of 1840 samples. Extending this calculation to 5 sampled parameters:

$1 - [1 - 0.05^{5}]^{N} > 0.99$

shows that some 14.8 million sets of sampled values would be required to provide the same confidence that one set represented the combination of the upper 5% from each sampled distribution.

These illustrative examples of the numbers of random samples required show why methods for determining when sufficient simulations have been run (convergence) may be more effective than an *a priori* determination of the number of samples required. Convergence is discussed in Section 3.5. More efficient approaches to sampling have also been developed. Some of these are discussed in the following sections.

3.3.2 Stratified or Importance Sampling

The number of samples required to fully explore model space using random sampling is large because there is no assurance that particular parts of this space will be sampled. Importance or stratified sampling overcomes this problem by dividing the model space into regions and then sampling from within each region. Generally, the number of samples corresponds with the number of strata, but the size of the strata can be uniform (equal probability, Figure 7) or variable (unequal probability, Figure 8). In each case, when the calculated values are used to define an output distribution or determine an expectation value, they are weighted according to the probability of the strata.

Equal strata probabilities (stratified sampling) are generally easier to implement, but may require a large number of strata to ensure that low-probability, high-consequence regions are adequately explored. Using unequal probability strata (importance sampling) ensures that such regions can be selectively sampled without an excessive increase in the number of samples (i.e., fewer samples are taken from the high-probability, low-consequence regions), without compromising the validity of the probabilistic approach.

The disadvantage of either approach to stratified sampling is that it requires knowledge of the model space so that the strata and their probabilities can be defined. This is relatively easy when there are only a few parameters and simple relationships between them. However, as the number of sampled parameters increases and the relationships become more complex, it becomes increasingly difficult to determine the characteristics of the model space and hence to define the strata and their probabilities.

Importance sampling was investigated in the HMIP programme of work prior to Dry Run 3. The methodology used was to run pilot simulations to identify parameters and parameter interactions leading to high dose (those that contribute to 95% of the risk estimate) and also times of maximum risk. Importance sampling functions were then defined by fitting beta or log-beta distributions to the cumulative risk curves at the times of maximum risk. Initial studies defined the distributions manually, but algorithms were later developed to automate the process. Sampling efficiencies[5] of more than 100 were demonstrated. However, the level of processing required to generate the importance sampling distributions reduced the overall efficiency.

Later HMIP models used methods for environmental simulation, introducing significant variations in the dose-time curves between simulations. This prevents the definition of a generally applicable importance sampling distribution. In Dry Run 3, an importance sampling case was specified in an attempt to improve convergence of a full climate simulation model. This resulted in improved convergence at the time for which importance sampling was specified, but considerably poorer convergence at other times.

3.3.3 Latin Hypercube Sampling

Latin Hypercube Sampling (LHS) has some of the advantages of both random and stratified sampling, but without the requirement to have *a priori* knowledge of how the parameter distributions and relationships interact to generate the model space. Instead, LHS divides each *parameter* range into intervals of equal probability and selects one value from each interval (Figure 9). When there are two or more sampled parameters, sample sets are formed by combining values at random from each

[5] Sampling efficiency here is the ratio of the number of random samples to the number of importance samples to reach the same degree of convergence.

parameter set. This combination is done without replacement so that each selected value is used once only in the analysis.

Values can be selected from each interval by random sampling within that interval. An alternative is to use the median value of each interval. Such Median Latin Hypercube Sampling (MLHS) provides more evenly distributed samples than random LHS. For a parameter that is defined as a single continuous distribution, the PDF generated using MLHS will usually look fairly smooth, even with a small sample size (such as 20), whereas the result using random LHS may look noisy. As sample size increases, the distinction between output from the two approaches becomes less.

MLHS requires slightly less computational effort than random LHS, although the effort saved is unlikely to be significant in comparison with the overall computational effort required for an assessment model. MLHS also yields the same set of sampled values each time a distribution is "sampled", if the number of samples is constant. This may be an advantage when system behaviour is being studied, although the order in which values are combined to form sample sets may lead to differences in the calculated output. MLHS can generate a distribution that is not representative of the true parameter distribution, specifically when the true distribution has a periodic function with a period similar to the size of the equiprobable intervals. However, parameters used in assessment models do not typically vary according to a periodic function of this kind.

The principal advantage of LHS is that it ensures that the entire range of each parameter is sampled, which in turn means that the distribution of sample sets will be more uniformly distributed across model space than with the same number of random samples. Helton and Davis (2001) show that, above a certain sample size, LHS results in calculated outputs with lower variance than those generated by random sampling.

Although it is not necessarily the case that the most important effects associated with a particular parameter only occur at the tails of the distribution, these tails, and the interactions between them, are commonly of interest. This interest arises both because these interactions can help in understanding system behaviour, and also because low-probability, high-consequence combinations can significantly affect the expectation value. With random sampling, there is no guarantee that the tails of the input distributions will be sampled. With LHS, the tails will be sampled but, because the combination of samples into sample sets is random, there is no guarantee that a sample from the tail of one distribution will be combined with a sample from the tail of a second distribution. Statistical assurance that particular parts of model space are represented in the output distribution can only be provided by increasing the number of samples, and thereby increasing the probability that samples from the tails of the distributions are combined.

The number of samples required to give a specified assurance that the maximum value in the output distribution exceeds the 99th or other percentile can be calculated in the same manner as for random sampling. For example, 299 samples of each parameter will give a 95% confidence that at least one value exceeds the 99th percentile of the calculated distribution. This approach was used to justify the selection of 300 samples to demonstrate compliance with the regulatory criterion in 40 CFR §194.55(d) for the WIPP. The number of samples needed to provide assurance that particular parts of model space are represented can also be calculated in a similar manner as for random sampling. In this respect, LHS is more efficient than random sampling, but large numbers of samples are still required if the sample sets include more than a few parameters.

3.4 Correlation

A criticism that can be levelled at poorly designed probabilistic calculations is that some simulations evaluate conditions that are not physically realistic. This can occur if parameters that are in reality correlated are defined and sampled using independent PDFs. For example, whereas in reality high values of one parameter would be associated with high values of a second parameter, and low values with low values, sampling from independent PDFs could combine high values of one parameter with low values of the other. If such combinations are not physically realistic, then any doses or other end-point calculated as a result will not be meaningful. In some circumstances, such erroneously calculated doses may lie within the central part of the calculated distribution, and therefore have no significant effect on the overall result. In other circumstances, however, unrealistic conditions can lead to calculated doses in the tails of the output distribution and thus significantly affect the calculated expectation value.

A specific element of the modelling methodology used in Her Majesty's Inspectorate of Pollution's (HMIP) Dry Run 3 was the re-examination of simulations that contributed significantly to risk. This step was to allow for the elimination of any simulations that modelled conditions that were physically unreasonable. Although a close examination of the results should be a part of the analysis of all assessment calculations, determining whether sampled conditions are realistic or unrealistic is a subjective process. Eliminating high dose cases on the basis of such judgements, and thereby reducing the overall expectation value of dose or risk, may not be transparent and could introduce an unintentional bias to the results. With large numbers of simulations, it is not feasible to apply the same level of scrutiny to all sets of sampled conditions. This means that unrealistic conditions that result in low doses are left in the analysis results, lowering the overall calculated dose or risk. This is another example of how risk dilution can arise if an assessment is poorly planned.

The most appropriate method for reducing the potential for simulations to represent unrealistic conditions is to account for correlations within the definitions of parameters and in the sampling stages of the analysis. Correlations between parameters can be accounted for in assessments in several ways:

- Explicit relationships in the equations of the computational models. This approach means that values for one parameter are calculated from the sampled values for one or more other parameters.

- A sampled parameter is used to select between different PDFs for one or more other parameters (Figure 10).

- A limited number of calculations using detailed models is used to define a response surface relating two or more parameters. This response surface is sampled and the corresponding parameter values used in assessment calculations.

- A normally distributed parameter can be correlated with another normally distributed parameter by adjusting the sampled values by a factor including the correlation coefficient and a normally distributed random number. Groups of parameters can be correlated by use of a dummy parameter.

- A sampling protocol is used to ensure that sampled values for two parameters reproduce the observed correlation between the parameters.

The first of these approaches is prescriptive and does not allow for any uncertainty concerning the extent of the correlation. It may nevertheless be of value where alternatives are not available and where independent sampling could cause significant errors. The second approach is more flexible, but the first parameter is effectively restricted to only a few possible values. For example, the selection of redox conditions by sampling could be used to determine the distributions to be sampled for radionuclide solubility, but only about four sets of redox conditions and associated PDFs for solubility could realistically be established.

The response surface technique is useful under certain specific circumstances. When the relationship between two or more parameters is relatively simple, other approaches are probably more efficient. However, when there are more complex relationships between parameters, which cannot be incorporated into a system model without excessively increasing run-times, a response surface can be calculated using a stand-alone detailed model. Sampling from this allows consistent sets of parameter values to be used in the simplified system model and reduces the potential for generating unrealistic conditions.

The method for generating a correlated parameter using the correlation coefficient and a random number is presented in Box 1, and an example is shown in Figure 11. This is a useful technique but is limited in its application because it is restricted to normally distributed parameters.

Box 1 - Generating correlated distributions

A normally distributed parameter x with a specified correlation to a normally distributed parameter y can be generated from the relationship:

$$x = \mu_X + (y - \mu_Y) C \sigma_X / \sigma_Y + \sqrt{(1 - C^2)} \sigma_X z$$

where:

x sampled value of X adjusted for correlation to Y

μ_X, μ_Y specified means for X and Y

σ_X, σ_Y specified standard deviations for X and Y

y sampled value of Y

C specified correlation between X and Y

z a random number sampled from a unit Normal distribution

The final approach listed above for introducing correlations is the most flexible, in that it can be applied to any type of distribution (and the correlated parameters need not have the same type of distribution), and more than two parameters can be correlated. This approach does, however, require additional steps in the modelling process and has also only been implemented in conjunction with LHS. A full description of the method is given in Iman and Conover (1982) and the mathematics are summarised in Helton and Davis (2001). An outline of the method is presented below.

The implementation of correlation into the LHS process uses the rank correlation coefficient (RCC) between the parameters rather than a correlation coefficient describing the relationship between parameter values. The RCC is determined by ranking the values for each parameter (the smallest value is ranked 1 and the largest is ranked n, where n is the number of observations) and then calculating the correlation between the ranks for each pair of values. The RCC is not only easier to implement in a sampling scheme, but it also allows for different types of parameter distribution to be sampled without changing other aspects of the model.

Table 1 illustrates a small set of observations and the calculated relationships and correlations.

The initial step for incorporating correlation into a LHS scheme is to sample the parameter PDFs. For each parameter, this gives n samples, each representing an equally probable part of the input distribution. Each set of sampled values is then rank ordered, and, instead of randomly selecting pairs of values, the pairs are selected so as to match the RCC. For example, if there was perfect correlation (RCC = 1), then the samples would be paired in rank order (r_{x1}, r_{y1} ; … ; r_{xn}, r_{yn}), and if there was a perfect inverse relationship (RCC= -1), then the samples would be paired in reverse order of their ranks (r_{x1}, r_{yn} ; … ; r_{xn}, r_{y1}).

	Observations		Ranks	
	x	y	r_x	r_y
1	10	2	1	1
2	15	6	2	4
3	22	4	3	2
4	30	5	4	3
5	45	8	5	5
	y = 0.13x + 1.84 Correlation coefficient = 0.8		Rank correlation coefficient = 0.7	

Table 1. Calculation of the rank correlation coefficient for a set of observations.

The method for pairing samples at other values of the RCC is based on a target correlation matrix that reflects the required correlation structure between parameters[6]. The sample matrix cannot be manipulated directly, and instead an independent matrix of scores is manipulated using a factorisation of the desired correlation matrix. The selection of the independent matrix is key to the validity of the approach as it determines how well the correlation of the manipulated matrix matches the desired correlation matrix. The scores used by Iman and Conover (1982) are approximations to Normal scores[7]. These scores were found to be generally effective and have been incorporated into the computer program developed to implement this approach (Wyss and Jorgenson, 1998).

Having manipulated the independent matrix, and checked that the resulting rank correlation matrix closely matches the target matrix, the sample matrix can then be ordered in the same way so as to give the same RCC.

3.5 Convergence

The model space that represents the interactions between all of the input parameters (and across the full range of these parameters) cannot, in general, be defined analytically. Probabilistic approaches are a means of overcoming this drawback by calculating the value of the output parameter(s) at various points across the model space. The model space for real systems is likely to be complex and not represented by a simple surface (see the discussion in Section 2.3). Conducting only a few

[6] More than two parameters can be correlated using this approach.

[7] Obtained by evaluating the inverse cumulative Normal distribution function (Φ^{-1}) at the values of the ranks scaled into the interval (0,1) using the van der Waerden transformation. The score s_i for observation x_i with rank r_i is given by $s_i = \Phi^{-1} (r_i / (n + 1))$.

calculations is unlikely to adequately represent this surface, and the expectation value calculated from a few calculations may differ significantly from the expectation value for the model surface. An infinite number of calculations could fully reproduce the model surface and give an exact estimate of the expectation value. This is clearly impractical, and the aim of a well-designed probabilistic analysis is to perform sufficient calculations to provide a reasonable representation of the model surface and estimate of the expectation value.

Two approaches can be taken to determining the number of calculations required. The first relies on general statistical relationships to determine the required number of samples *a priori*. The second approach is based on repeated examination of the calculated results to determine whether they adequately represent the system or if more calculations are required.

The *a priori* determination of the number of calculations required is based on the premise that, if samples are drawn at random from the input distributions, then the output distribution can be regarded as a random sample of the output population (model space). However, determining how many samples are required to ensure that this random sample is an adequate representation requires further assumptions about the form of the model surface. Assuming, for example, that a "95% confidence of at least one value being above the 99th percentile"[8] is a reasonable measure of adequacy has an implicit assumption about the form of the output. If these implicit assumptions are not met, for example if the distribution is highly skewed, then the number of calculations may not be sufficient to provide a reasonable estimate of the expectation value.

Conceptually, the second approach does not require any assumptions about the form of the model surface. Instead, a comparison is made between the calculated output distribution and the model surface. As more and more calculations are performed, these two distributions will converge. A measure of how similar they are at any stage can be compared to an established criterion and a decision made as to whether more calculations are required.

The drawback of this approach to assessing convergence is that the form of the model surface is unknown, so that a direct comparison cannot be made. Various surrogate measures can be used, but these are based on assumptions about the form of the distributions and some caution is still required in their interpretation. The most common approach is based on the standard error of the mean (SEM).

If a sample set is taken from a population, then the mean of the sample set provides an estimate of the mean of the population. If a number of independent sample sets are taken, then the means will form a distribution. For a large sample, the distribution of the sample means is approximately a normal distribution, even if the population from which the samples were drawn is not a normal distribution. The central limit theorem states that the standard deviation of this distribution (termed the SEM) is equal to the standard deviation of the population divided by the square root of the sample size:

[8] This is the criterion for determining the number of samples required in the regulations applicable to the WIPP [40 CFR §194.34(d)].

$$\text{SEM} = \frac{\sigma_P}{\sqrt{n}} \tag{17}$$

Because the standard deviation of the population is not known, it is common practice (valid if the sample is large) to use the sample standard deviation instead:

$$\text{SEM} \approx \frac{\sigma_s}{\sqrt{n}} \tag{18}$$

One test of convergence is that sufficient calculations have been undertaken when the SEM is less than 1% of the sample mean. Alternative tests are based on the extent to which the SEM, or similar measures, change as additional sets of calculations are performed.

These convergence tests, because they are applied to the actual calculated results, are likely to be seen as more justifiable than setting limits independently of the calculations, even if there is statistical validity to both approaches.

3.6 Timing of Events

Introduction

The derivation of parameter PDFs discussed in earlier sections has focused on parameters representing physical quantities that can be measured or calculated through detailed models. Parameters for which PDFs need to be elicited because there are no physical data are also included as the techniques used to select distributions and sample from them are similar even if the definition of values differs. There is a class of this latter type of parameter that requires special consideration, however, both because different distribution types and sampling techniques may be required, and also because they can have a significant effect on calculated dose / risk and the potential for risk dilution.

The parameters in question are those relating to the timing of events. Such events might be the onset of different climate conditions, the initiation of faulting in the geosphere, canister failure, or intrusion into the repository or a contaminant plume. All of these types of event have been considered in different probabilistic assessments using various methods.

In Dry Run 3, changes in climate were simulated using a Markov Chain model, which assigns probabilities to the various possible transitions from one climate state to another. Successive sampling of these probabilities leads to sequences of climate states from which the overall climate evolution sequence is generated by sampling PDFs of the duration of the different states. The WIPP assessment accounted for the effects of climate change by sampling a Climate Index from a bi-modal distribution developed to account for two possible future climate patterns.

These, and other, approaches to modelling climate change are based on the extensive knowledge about past changes; reasonable assumptions can be made to extrapolate these changes into the future, thus allowing PDFs to be developed for the duration of climate states. The other potential events do not share this attribute that allows an extrapolation of past events, and the most reasonable assumption that can be made is that future events will take place at random times.

Random events

An example of events treated as random in time (and space) is the treatment of human intrusion (drilling) at the WIPP site. This approach is prescribed in the regulations [40 CFR §194.33(b)(2)], which state that drilling should be assumed to occur in the region at random intervals in time and space. Events that are random in time can be described as following a Poisson process. For a single event, the probability, P, that the event will take place in a specific time interval, $[E(\Delta t)]$, is given by:

$$\mathrm{P}[E(\Delta t)] = \lambda \, \Delta t \, e^{-\lambda \Delta t} \tag{19}$$

where λ is a rate constant with units of events per time. This can be extended to consider the probability that *n* events occur in a specific time interval:

$$\mathrm{P}[E_n(\Delta t)] = \frac{[\lambda \, (\Delta t)]^n}{n!} \, e^{-\lambda \Delta t} \tag{20}$$

In a simulation model, the period of interest is divided into successive intervals of length Δt, and a random number is generated for each interval. If this random number is less than the probabilities given by the equations above, then one or more events are assumed to occur in that interval. For example, if the rate constant is 0.001 per year and the time interval is 1000 years, then the probability of a single event occurring is 0.368, the probability of two events is 0.184, and the probability of three events is 0.061. If the random number generated for a particular period is 0.3, then one event is assumed to occur, if it is 0.1 then two events are assumed, and so on.

This approach to sampling the occurrence of events has no "memory" in that events, or lack of them, in one time interval have no influence on the occurrence of events in the successive intervals. Conditional assumptions can be introduced to the model to account for this type of effect, or to limit the overall number of events simulated.

Risk dilution

Simulating events at random in the manner described above means that each simulation will have a different history of events. If the rate constant is very low, and particularly if the time interval is short, then the probability of an event occurring in the same time interval in more than one simulation becomes very small. The overall significance of this in terms of the expectation value for dose or risk is dependent on how the event affects the overall performance and the time over which the effect persists.

If the event being simulated is one that initiates a new set of boundary conditions or has some other effect that persists for a long period, then the effect of different histories may be small. For example, the propagation of a new fault that provides an alternative pathway for radionuclides may be modelled this way. The time at which the fault is formed may be different in each simulation, but there is a cumulative effect so that after, say 100 time-steps, faulting will have been simulated in the majority of the simulations. The expectation value will change steadily over the period in which more and more simulations have the new conditions. This change may properly reflect the uncertainty in the time at which a fault might occur.

If the event being modelled initiates a change that is short-lived (i.e., the effect persists for only a few time-steps), then the effect of different time histories may be much more significant. In such a case, there is no cumulative effect, so that the effect of the event does not appear in more and more simulations at later time-steps. Instead, the expectation value at any particular time is derived from many simulations not affected by the event concerned and a few simulations with the effect. The same applies at subsequent time-steps, except that the simulations with the effect of the event will be different.

As an example of this issue, consider the following hypothetical case. Suppose the conditional risk from the normal processes acting on a repository is $1x10^{-7}$ per year and the conditional risk from a particular event is $1x10^{-5}$ per year for about 1,000 years. The assessment model considers a 100,000 year interval using 100-year time-steps. If the timing of the event is fixed, then the expectation value for risk will be $1x10^{-5}$ per year for around 10 time-steps, and $1x10^{-7}$ per year for the remainder. However, if the event occurs randomly in time, then for any one time-step the effects of the event will be apparent in only 1% of the simulations. This means that the expectation value of risk will be about $2x10^{-7}$ per year throughout the period considered.

This example illustrates the potential for an *increase* in uncertainty (i.e., assuming that the event occurs randomly rather than at a specific time) to nevertheless result in a *decrease* in the expectation value of risk. It is a further example of how risk dilution could occur in assessment calculations.

The example above is illustrative and not based on a real system or assessment. Nevertheless, it suggests that there is potential for calculated results to be affected significantly by the assumptions on how random events are treated in an assessment. A similar conclusion was reached by the International Peer Review for the Yucca Mountain performance assessment (NEA/IAEA, 2002).

4 Outputs and their Interpretation

4.1 Introduction

All of the stages of a probabilistic approach are important in developing a robust analysis in an efficient manner. However, the presentation of results could be regarded as the most important, as it is at this stage that the approach must be communicated to others. If the results cannot be clearly communicated, then the analysis may not provide as much benefit to decision-makers as would otherwise be the case.

There are two principal ways of presenting the output from an analysis - numerical and graphical. Any graphical presentation will, of course, be based on numerical data, but the intent is to detect or illustrate trends and patterns rather than to present precise information. There are also hybrid approaches, such as colour-coding of tables of data to highlight patterns (Table 2). However, the discussion in this section is focused on the numerical information that can be derived from probabilistic calculations and used for decision-making, and the ways in which graphical techniques have been used in uncertainty and sensitivity analyses.

0.5526	0.3471	0.3825	0.9109	0.1038	0.0331	0.5915	0.6516	0.8211	0.6531
1.0609	0.8803	0.5116	0.9276	0.1333	0.2057	1.0530	1.0603	0.9668	1.1201
1.2335	1.5121	0.6501	1.1109	1.1278	0.2197	1.1643	2.0136	1.6826	1.2666
1.2656	1.8408	0.8511	1.8501	2.2900	0.3139	1.8189	2.0820	2.4957	1.3544
2.0041	2.7446	1.4610	2.0042	2.6928	0.6747	2.0240	2.5036	2.5618	1.4775
2.0174	3.0495	1.4921	2.6198	2.8053	0.7460	2.4679	3.6941	2.6669	2.0711
2.3149	3.2754	2.2312	3.0380	3.1315	2.6079	2.8106	3.6943	3.0448	2.0909
2.4285	3.6949	2.5905	3.4272	3.4456	3.8542	3.1319	5.4333	3.3782	3.3441
2.9058	4.0295	3.7212	3.8396	3.7904	4.3086	3.9789	5.5370	3.5686	4.2372
3.2730	4.8105	4.1960	4.7072	4.2854	4.6696	5.4086	5.5407	4.8293	4.9531
4.4734	4.8896	4.2361	6.5708	4.5736	5.1941	5.9030	6.2676	4.9406	5.2798
5.6215	4.9267	4.3246	7.3122	4.7350	5.1952	6.5953	6.4495	5.5498	5.3415
6.1329	5.7180	4.8336	7.9063	4.7660	5.4590	7.8134	6.6181	6.2700	6.0349
6.9818	6.1786	5.3610	8.1818	5.1223	5.5945	8.0188	6.7046	6.2890	6.3901
7.0329	6.2123	5.5884	8.3761	5.1813	5.7987	8.0978	6.8117	6.6663	6.5010
7.6002	6.4958	7.7034	8.3918	5.3152	7.0289	8.1527	7.3331	8.3839	7.5030
7.9174	6.8036	7.8678	8.5415	6.7569	7.6059	8.7782	7.3618	8.3854	7.9412
8.4221	8.2906	7.9327	8.8241	7.2416	8.6593	9.1775	9.3260	8.6064	8.7436
9.5261	8.3006	8.3322	9.0407	7.8613	8.7560	9.5425	9.4841	9.3601	8.7696
9.8977	9.0174	9.9137	9.3241	8.2946	8.9574	9.5554	9.5292	9.7489	9.6992

Table 2 Illustration of colour-coding numerical data to highlight trends.

4.2 Numerical Output

Probabilistic calculations for complex systems can generate large amounts of numerical data. As a minimum, each simulation will generate a set of input values and an output value. However, most calculations undertaken for performance assessments are concerned with the behaviour of the system over long periods, and so there will be output values for all the times considered. In addition, it is often of value to have more than one output parameter. For example, in addition to the overall dose, the doses arising from individual radionuclides are useful outputs. Intermediate outputs, such as radionuclide concentrations in different media and compartments, fluxes across sub-system interfaces, and the inventory remaining in the repository, are also of value in developing a system understanding and communicating the assessment. When this amount of output is multiplied by the number of simulations performed, which may be several thousand, difficulties can arise in both the management and interpretation of the data.

One approach to data management is to not store the output values but to calculate, and re-calculate, them as required. Providing that the sampled input values are retained (or the algorithm for generating "random" numbers is repeatable), any output required can be generated by re-running the model(s). The factors that affect the balance between retaining and retrieving data on the one hand, and re-running the model as required on the other, include the maturity of the assessment as well as the relative costs of different computing tasks. These costs are too dependent on the computing environment concerned, and also vary too much with time, for any recommendations to be developed, but some general observations can be made.

In the early stages of an assessment, users developing a system understanding may want to access a wide range of outputs, but storing data without knowing which data will be useful could lead to redundancy, so re-calculation as new requirements are identified would be favoured. Model changes and changes to input distributions would also favour re-calculation. As the assessment programme matures, there will be fewer changes in the input distributions, models will become finalised, and the outputs that are most useful will have been identified. All of these factors would favour storage. Stored data would also allow other stakeholders to examine assessment results without the difficulties of making models and codes widely available. Within the assessment programme itself, quality assurance may be easier to demonstrate for stored data than for the configuration management necessary to re-calculate this data.

A hybrid approach to reducing the computational requirements of re-calculating rather than storing large amounts of data from probabilistic assessments is to use a response surface. A relatively small number of simulations using the full model(s) are used to determine a number of output data points, which are then used to define an approximate output response surface. Additional output points are then generated by interpolation on the response surface rather than by using the full model(s). A prototype of such an approach, known as DeskTop PA, was developed for the WIPP (Crawford *et al*., 1998).

Whether data are stored or re-calculated as required, tabulations of large amounts of "raw" data are difficult to interpret. To overcome this, either the data must be presented or plotted in a more easily interpretable manner (see the following section), or aggregated or summarised so that only a few data values are presented.

The simplest means of summarising large amounts of data is to use the average value. There are several ways to express the average, depending on the distribution of the data. However, the most familiar average is the arithmetic mean, and this is commonly used even when an alternative, such as the geometric mean, would be more appropriate from a mathematical perspective.

Apart from the average, there are other measures of central tendency, such as the median or mode. These can have specific uses - for example the mode or modes are useful where the data fall into two or more groups and the mean might correspond to a value that does not occur in the data-set (Figure 12). Such examples are difficult to identify *a priori*, and will likely only be found if the data are examined graphically. An attribute of the median and mode that can be important is that they are defined even if there are missing data values, which would invalidate the mean. The mode and median are also more readily determined than the mean when assessing results from plotted data rather than directly from the numerical data.

The mode and median provide information about the central part of the distribution and are not influenced by wide tails to the distribution as is the case with the mean (see Figure 12).

In addition to the various measures of central tendency, there are numerical descriptors for the shape of distributions. These are largely based on the moments of the distribution about either the origin or the mean. The mean itself can be defined in this way, as the first moment about the origin:

$$\mu = \sum_i (x_i - 0)^1 / n \tag{21}$$

The second moment about the mean, or variance:

$$\mu_2 = \sum_i (x_i - \mu)^2 / n \tag{22}$$

is a measure of the dispersion of the data. The other common measure of dispersion, the standard deviation (σ), is the positive square root of the variance ($\mu_2 = \sigma^2$).

The variance and the third moment about the mean:

$$\mu_3 = \sum_i (x_i - \mu)^3 / n \tag{23}$$

are used to define the skewness (γ_1) or extent to which the distribution is asymmetric:

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \tag{24}$$

A comparison of the mean and the median or other percentiles can also be used as an indication of how skewed a distribution is, and there are several measures in use, including:

$$\text{Pearson mode skewness} = \frac{[\text{mean}] - [\text{mode}]}{\sigma}, \text{ and} \tag{25}$$

$$\text{Quartile skewness coefficient } = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1}. \tag{26}$$

Positive skewness indicates a tail towards the maximum side of the distribution.

Although all of the numerical descriptors can be calculated for any set of values, some care is needed in interpreting and using the values in statistical tests if the underlying distributions differ significantly from normal distributions.

A key type of output from performance assessment calculations is the time-dependent behaviour of a parameter (see Section 4.3.6). Average values for such distributions (with respect to time) are not generally meaningful. However, the maximum values, for example peak dose or maximum concentration, may be useful measures of system performance.

All of the derived values discussed above relate to individual parameters. There are in addition a large number of statistical techniques for assessing the relationships between different sets of values, and these can be used to examine correlations between input parameters and output parameters or between different output parameters. Multivariate techniques are available for examining relationships between more than two parameters and these may be useful for identifying key input parameters. The majority of sensitivity and uncertainty analyses (see Section 2.2) reported for performance assessments, however, rely on the examination of scatter plots (see Section 4.3.5) and the calculation of regression coefficients.

The linear correlation coefficient between parameters x and y is given by:

$$r = \frac{\mathrm{Cov}_{xy}}{\sigma_x \sigma_y} \tag{27}$$

where Cov_{xy} is the co-variance:

$$\mathrm{Cov}_{xy} = \frac{1}{n}\sum_i (x_i - \mu_x)(y_i - \mu_y) \tag{28}$$

and σ_x and σ_y are the standard deviations for x and y. The value of r varies from +1 (perfect correlation) through 0 (no correlation) to -1 (inverse correlation). The same calculation can be performed for the ranks of the parameter values (see Section 3.4). A high correlation coefficient is a reliable indicator of a linear relationship between the distributions, although it may not indicate a causal relationship. A low correlation coefficient, however, can result from the presence of outliers, threshold effects where the relationship changes across the parameter ranges, or a non-linear relationship. The examination of scatter plots (section 4.3.5), in addition to reviewing correlation coefficients, will help to identify these situations.

4.3 Graphical Output

There is a wide variety of methods by which the data from performance assessment models can be plotted and displayed. These methods can be classified in various ways, depending on whether they are intended to display all of the data concerned or to summarise information, whether they show data for one parameter or for several, and whether they explicitly show parameter values as a function of time. The overall purpose of performance assessments is to assess the evolution of a disposal facility and the associated barriers, so that some output involving time is almost always required. The other forms of output are useful for understanding system behaviour and may also be the principal type of output if the applicable regulations require particular performance measures.

For example, the regulations governing the WIPP require the determination of "cumulative releases of radionuclides to the accessible environment for 10,000 years after disposal from all significant processes and events that may affect the disposal system" [40 CFR §191.13(a)]. This performance measure is therefore not a function of time, although understanding system evolution remains a key part of the assessment process. These requirements are reflected in the way in which data are plotted and presented.

The types of graphical output described below are:

- Methods for illustrating behaviour of individual parameters
  - Probability distribution functions
  - Cumulative distribution functions
  - Complementary distribution functions
  - Box-whisker plots
- Methods for showing the relationship between two parameters
  - Box-whisker plots
  - Scatter plots
- Methods for showing time-dependent behaviour
  - Dose / risk vs time plots

4.3.1 Probability Distribution Function (PDF)

Mathematically, PDFs for output parameters have the same meaning as those for input parameters, and show the probability of the parameter having a particular value or falling within a particular range. There is a difference, however, in how these two types of PDF are typically defined. PDFs for input parameters are generally defined as mathematical functions, requiring just 2 or 3 values to define the characteristics (e.g., maximum, minimum and mode, or mean and standard deviation). In contrast, output PDFs are typically defined as histograms or pseudo-continuous distributions (Figure 13).

The number of data points used to define output PDFs is not the same as the number of output values from the calculations. The overall range of output values is divided into a number of sub-ranges and it is the number of output values in each of these sub-ranges that defines the PDF. In the case of a uniformly-distributed output parameter, the limit on the number of sub-ranges is such that there is a single output value in each sub-range. As the distribution becomes more peaked, however, more and more output values will lie within a few sub-ranges around the mode. If the same large number of sub-ranges is maintained, then this will mean that more and more of the sub-ranges near the tails of the distribution, and even some near the centre, will have no output values. Reducing the number of sub-ranges too much, however, will obscure subtleties within the distribution. The optimal number of sub-ranges is too dependent upon the form of the distribution for it to be prescribed, although a starting point for N values is k sub-ranges such that $2^k > N$. For the example in Figure 13, this would suggest 7 sub-ranges.

PDFs are the most useful method for illustrating the overall form of the distribution, particularly the position of the mode and the extent of the tails. However, it may be more difficult to extract numerical information (e.g., the value of the median or the 95th percentile) from such plots.

4.3.2 Cumulative Distribution Function (CDF)

The CDF for output parameters is analogous to the PDF in the sense that it is defined by a set of data values defining a histogram or pseudo-continuous distribution rather than as a mathematical function. There are, however, two ways of defining the CDF from a set of output values. The first approach is the similar to the approach for the PDF and involves dividing the overall range into a set of sub-ranges or bands, determining the frequency of values falling within these bands and then plotting the cumulative frequency from the minimum value to the maximum. Limiting the number of sub-ranges is not as important as with the PDF, as sub-ranges with no output values have less effect on the plotted distribution. In fact, increasing the number of sub-ranges can be beneficial in that it will decrease the vertical (frequency) step size on a CDF plot, although the horizontal step size will increase.

The alternative approach to plotting a CDF involves plotting all the output values rather than grouped data. In this case, it is the frequency scale which is divided into intervals (of size 100/N, where N is the number of output values). The output values are ranked and plotted against the frequency intervals.

The choice between the two approaches to plotting CDFs depends in part on the intended use of the plots and in part on the amount of data involved. For summarising information, particularly where there are large numbers of data points, the histogram-type plot may be most appropriate. Where further detailed analysis of the data is intended, and particularly where there is a comparatively small amount of data, plotting the output values directly is likely to be most effective, The use of interactive analysis tools may allow specific values or ranges of values to be selected and assessed for other characteristics or relationships. Plots such as scatter plots (see below), which already show at least one relationship may, however, be more suitable for this type of data analysis and inspection.

CDFs are particularly useful for deriving specific values (e.g., the median or 95th percentile), and also illustrate the extent of the tails of the distribution. It can be difficult to visualise complex distributions (e.g., bi-modal) from this type of plot.

4.3.3 Complementary Cumulative Distribution Function (CCDF)

The CCDF has a particular place in the presentation of results from performance assessments as it is the form of output required by the regulations that apply to the WIPP site [40 CFR §194.34(a)]. Whereas the CDF plots the proportion of the distribution less than a particular value, the CCDF plots the proportion of the distribution greater than the value. In other words, the CDF ranges from 0 to 100%, whereas the CCDF ranges from 100 to 0% (Figure 14). Except where required by regulation, the choice between using the CDF or the CCDF of a distribution depends simply on how questions about the data are likely to be asked. If the question is of the form "What is the probability that a particular value is exceeded?", then the CCDF is the most appropriate. Otherwise, the CCDF has the same advantages and disadvantages as the CDF.

4.3.4 Box-whisker plots

A key purpose of using graphical output is to allow for easy comparison of distributions arising from different models, assumptions or input data. Superimposing PDFs or CDFs allows such comparisons, but can become confusing and difficult to interpret if there are more than a few curves. An alternative approach is to plot the principal features of the distributions and to plot these side-by-side so that similarities and trends can be readily made.

A common form of this type of presentation is the box-whisker plot. There are a number of variants of this plot, but in the form illustrated in Figure 15 the "box" is defined by the quartiles, with a line marking the median, and the whiskers extending to the 10th and 90th percentiles. The dots mark the maximum and minimum values. Variants of this plot use different criteria for the whiskers. For example, they may mark the maximum and minimum values, or define a range of values that are less than Q_1 - 1.5IQ, or more than Q_3 + 1.5IQ (where IQ is the inter-quartile range).

Box-whisker plots provide a useful summary of a distribution, but cannot be used to determine anything other than the specific values used to construct them (median, quartiles, maximum and minimum). It can be difficult to visualise the overall form of the distribution, particularly of complex distributions, from this type of plot.

4.3.5 Scatter plots

All of the graphical outputs described above display results for a single parameter, although stacked plots (such as the box-whisker plot) may be used to show variability in one parameter for different values of a second parameter. The scatter plot is the simplest form of output for showing the relationship between two parameters. The regression coefficient can be used to determine whether there is a strong relationship between two parameters, but the scatter plot allows a more detailed examination of the relationship and is of particular value in developing an understanding of system (and model) behaviour. Scatter plots may show, for example, where there are thresholds in a relationship (Figure 2) or where an otherwise strong relationship is affected by a few outliers.

Scatter plots can be used to examine the relationship between input parameters and output parameters, or to examine how two different output parameters are related. The first type of these comparisons is the basis for sensitivity and uncertainty analysis and will identify those input parameters on which the output is particularly dependent. The second type of comparison is of value in developing an understanding of system behaviour. An example of a detailed analysis based on scatter plots is provided by Kleijnen and Helton (1999).

It would often be of value to compare relationships between more than two parameters, but three-dimensional plots tend to require individual manipulation of the scales and perspective to allow interpretation of the relationships. Statistical analysis and display software can help the analyst examine more complex data structures through the use of colour, filters and interactive data queries, and customised plots can be generated to illustrate key relationships. However, for the routine communication of results to a wider audience, the basic two-dimensional scatter plot is likely to be the most efficient approach. These basic plots may also provide useful support to customised plots by allowing different audiences to draw their own conclusions from the data; customised plots may give the perception that the data have been manipulated to show a particular interpretation.

4.3.6 Dose / risk vs time

The key calculation within a performance assessment of a radioactive waste disposal facility is the calculation of the doses received by members of the public, or the risks to which they are exposed. During the operational phase of a facility, the individual or group of individuals concerned (the critical group) can be identified and various scenarios or sets of events that could lead to a dose can be defined. This means that the principal output of the analysis can be defined as a single parameter[9]. If probabilistic methods are used in the calculation, then the output can be illustrated using one of methods described above – PDF, CDF or CCDF.

For assessments of post-closure performance, however, a single output parameter is not sufficient because the doses received will vary with time. It therefore becomes necessary to calculate a set of parameter values – the dose or risk at a series of time-steps. If a probabilistic approach is used, then these calculations are repeated using different sets of input values. Illustrating this type of time-dependent behaviour can be done using stacked PDFs or similar plots of dose or risk at a set of specific times, or using a pseudo-continuous dose or risk vs time plot.

The number of time-steps at which doses are actually calculated is a function of the rate of transport of radionuclides through the system and various modelling factors. Too few time-steps may lead to inaccurate results, but there are a resource constraints (run-times and data storage) on increasing the number too much. In any event, for a typical performance assessment model, the number of time-steps will be sufficiently large that plotting individual PDFs for *all* time-steps would be likely to obscure rather than illustrate system behaviour. Reducing the number of plots to a manageable number would require selecting particular time-steps to be illustrated and this could be difficult to do *a priori* when the significant times are unknown. Using a series of box-whisker plots side-by-side could, however, increase the number of time-steps at which the output can be explicitly plotted (Figure 16).

The next step in increasing the number of time-steps explicitly illustrated is to plot individual values at each time-step and to join these together to form a pseudo-continuous plot of dose or risk changing with time. For probabilistic calculations, the

[9] This single parameter is aggregated across all the scenarios, pathways and radionuclides considered, but the same principles apply to the display of results if parameters are calculated for particular sub-sets of these factors.

curves for each set of input values can be superimposed to show the range of potential behaviours (Figure 17). Alternatively, a single curve can be plotted based on a value derived from the distribution at each time-step (Figure 18). The derived value usually used is the mean, leading to a mean dose / risk vs time plot. Other derived values can be used, and several curves, such as the mean, median and 95th percentile, can be plotted to summarise the overall behaviour.

It is important to note the mean dose or risk vs time curve is *not* the same as the result of using the mean values for each of the input parameters. This curve may not, in fact, correspond to any of the individual dose or risk vs time curves. At some point in the analysis, *all* of the individual curves should be examined to ensure that the derivation of a single value has not obscured features of the underlying calculations. The potential for the mean dose or risk vs time results to obscure some aspects of system behaviour is discussed further in Section 3.6.

5 References

Crawford, M.B., Wilmot, R.D., Galson, D.A., Swift, P.N. and Fewell, M.E., 1998. PASS: A component of DeskTopPA for the WIPP. In *Proc. Eighth Int. High Level Radioactive Waste Management Conf.* (Las Vegas, 10-14 May 1998), American Nuclear Society, La Grange Park, IL and American Society of Civil Engineers, New York, NY, 1998.

DOE (U.S. Department of Energy), 1996. Title 40 CFR Part 191 Compliance Certification Application for the Waste Isolation Pilot Plant, October 1996. DOE/CAO-96-2184.

DOE (U.S. Department of Energy), 1998. Viability assessment of a repository at Yucca Mountain, DOE/RW-0508, December 1998.

Goodwin, B.W., McConnell, D.B., Andres, T.H., Hajas, W.C., LeNeveu, D.m., Melnyk, T.W., Sherman, G.R., Stephens, M.E., Szekely, J.G., Bera, P.C., Cosgrove, C.M., Dougan, K.D., Keeling, S.B., Kitson, C.I., Kummen, B.C., Oliver, S.E., Witzke, K., Wojciechowski, L. and Wikjord, A.G. 1994. The disposal of Canada's nuclear fuel waste: Postclosure assessment of a reference system. AECL-10717, COG-93-7. AECL, Pinawa, Manitoba.

Helton, J.C., 1993. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. *Reliability Engineering and System Safety*, **42**(2-3), 327-367.

Helton, J.C. and Davis, F.J., 2001. Latin Hypercube Sampling and the propagation of uncertainty in analyses of complex systems. SAND2001-0417. Albuquerque, NM: Sandia National Laboratories.

Hora, S. and Jensen, M., 2002. Expert judgement elicitation. SSI Report 2002:19. Swedish Radiation Protection Authority, Stockholm.

Iman, R.L. and Conover, W.J., 1982. A distribution-free approach to inducing rank correlation among input variables. Communications in Statistics: Simulation and Computation. **B11**(3), 311-334.

Kleijnen, J.P.C. and Helton, J.C., 1999. Statistical analyses of scatterplots to identify important factors in large-scale simulations. SAND1998-2202. Albuquerque, NM: Sandia National Laboratories.

Mishra, S., 2002. Assigning probability distributions to input parameters of performance assessment models. SKB Technical Report TR-02-11. SKB, Stockholm.

NEA/IAEA, 2002. An International Peer Review of the Yucca Mountain Project TSPA-SR. OECD, Paris.

Norris, S., Bailey, L.E.F., Askarieh, M.M., and Hickford, G.E., 1997. Nirex 97: An Assessment of the Post-closure Performance of a Deep Waste Repository at Sellafield. Overview. UK Nirex Report S/97/012. Nirex, Didcot.

SKB (Swedish Nuclear Fuel and Waste Management Company), 1999. SR 97 Main Report Volumes I and II. SKB Technical Report TR-99-06. SKB, Stockholm.

Sumerling, T.J. (Editor), 1992. Dry Run 3 - A trial assessment of underground disposal of radioactive wastes based on probabilistic risk analysis - Volume 10: Overview. UKDOE/ HMIP Report DoE/HMIP/RR/92.039.

Wilmot, R.D., 2002. Formulation and presentation of risk assessments to address risk targets for radioactive waste disposal. SKI Technical Report 02:21. SKI, Stockholm.

Wilmot, R.D. and Galson, D.A., 2000. Expert Judgement in Performance Assessment SKI Technical Report 00:47. SKI, Stockholm.

Wilmot, R.D. Galson, D.A. and Hora, S.C., 2000. Expert Judgements in Performance Assessments. Report of an SKI/SSI Seminar. SKI Technical Report 00:35. SKI, Stockholm.

Wyss, G.D. & Jorgensen, Kelly H., 1998. A User’s Guide to LHS: Sandia’s Latin Hypercube Sampling Software. SAND98-0210. Albuquerque, NM: Sandia National Laboratories.

6 Figures

Figure 1 **Illustration of a response surface for output parameter *y* as a function of two input parameters x_1 and x_2.**

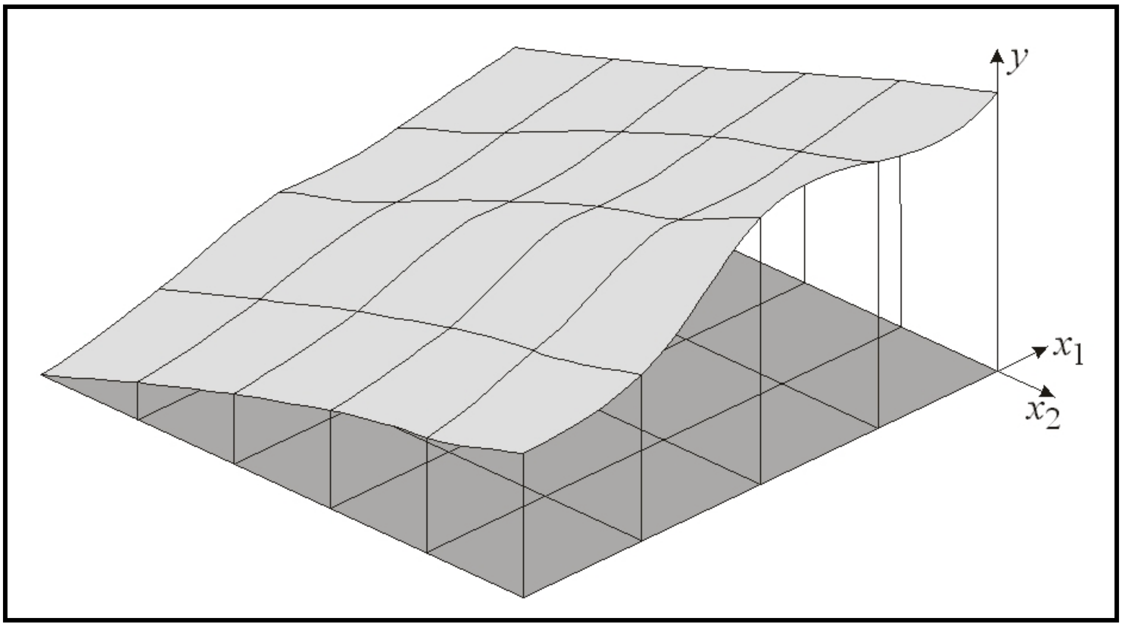


Figure 2 **Illustrative scatter plot showing how a low correlation coefficient (r=-0.5) can mask a threshold effect.**

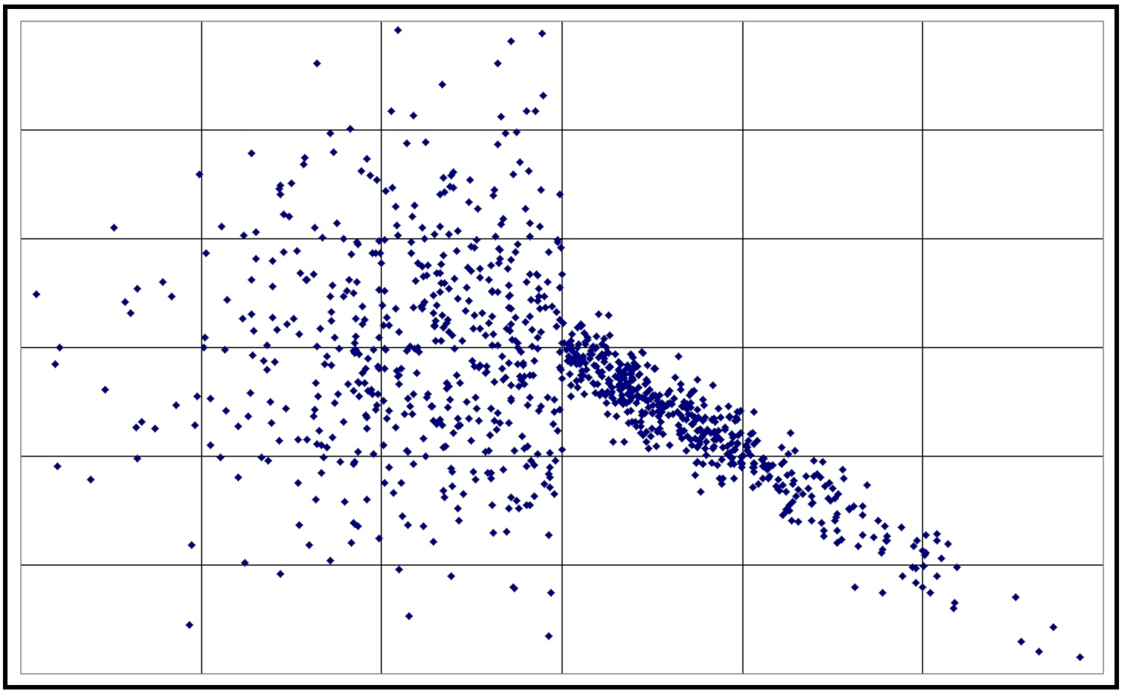


Figure 3 **Comparison of cumulative distribution functions for a triangular distribution (-3,0,3) and a normal distribution (0,1).**

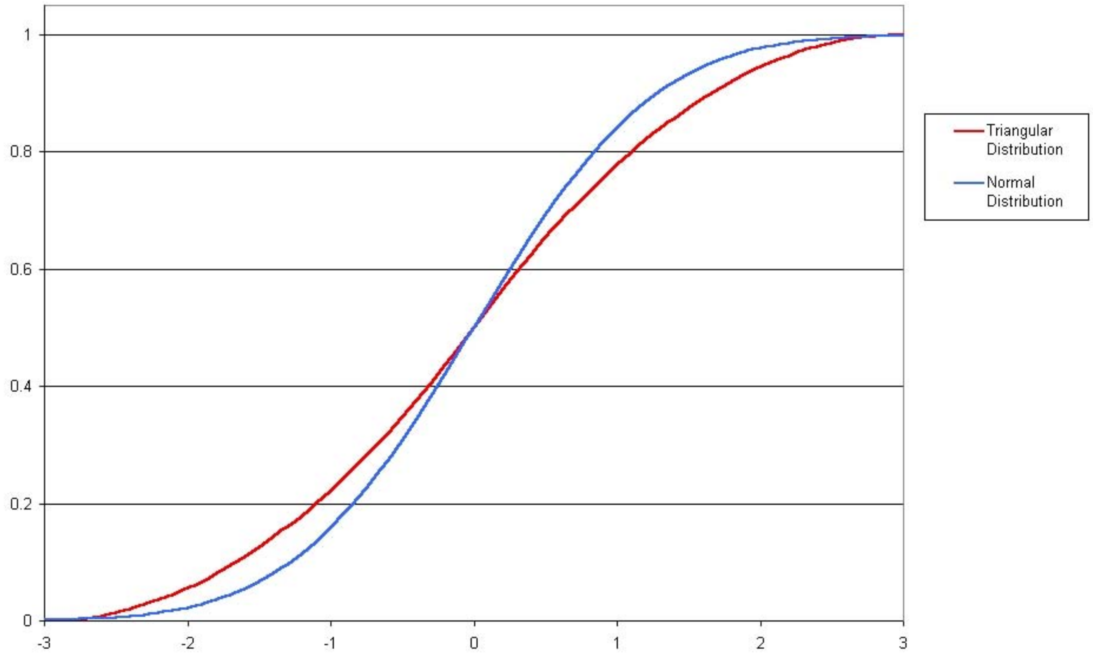


Figure 4 **Examples of the wide range of distributions that can be generated from a beta function.**

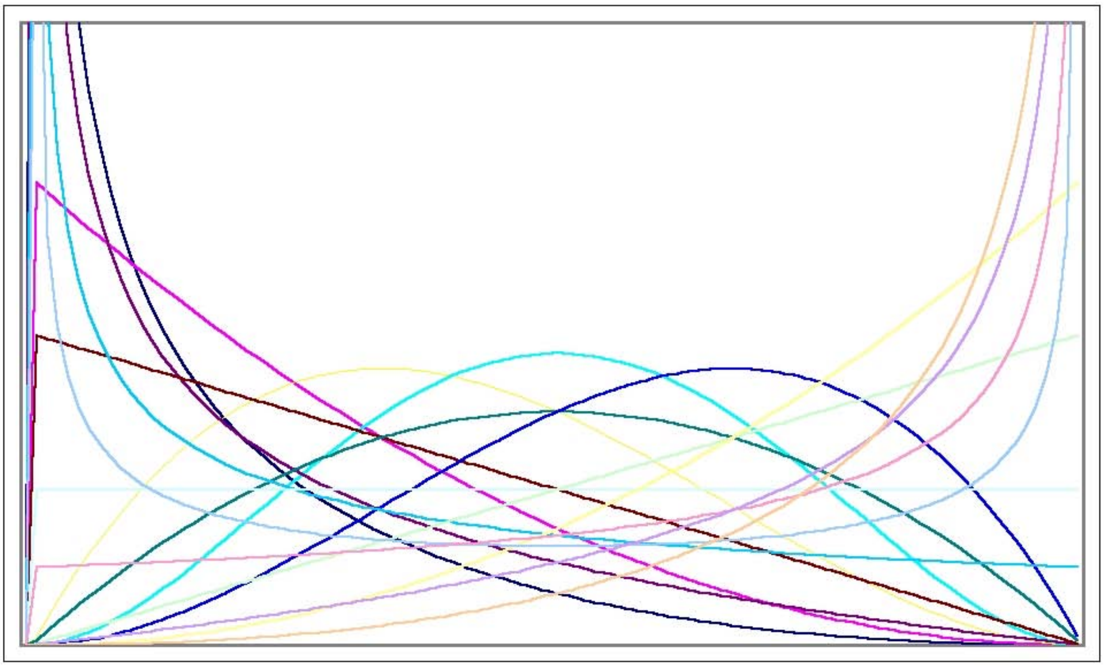


Figure 5 **Generating random samples from a triangular distribution by mapping uniformly distributed samples to the cumulative distribution function.**

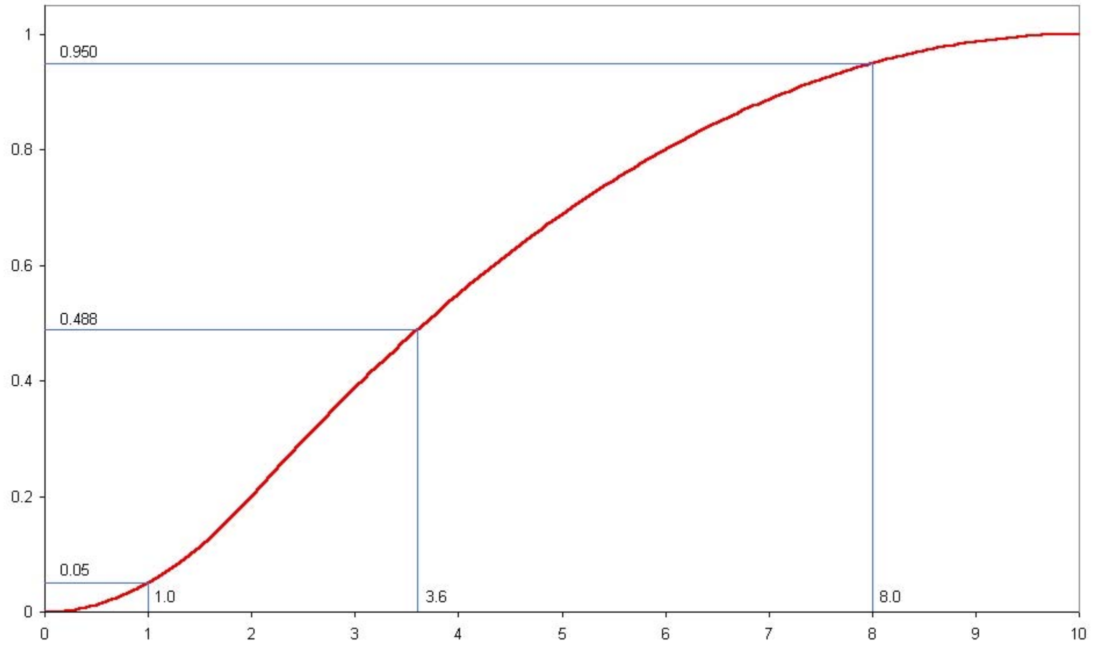


Figure 6 Comparison of the regions of model space defined by the 95th percentiles of the input and output parameter distributions.

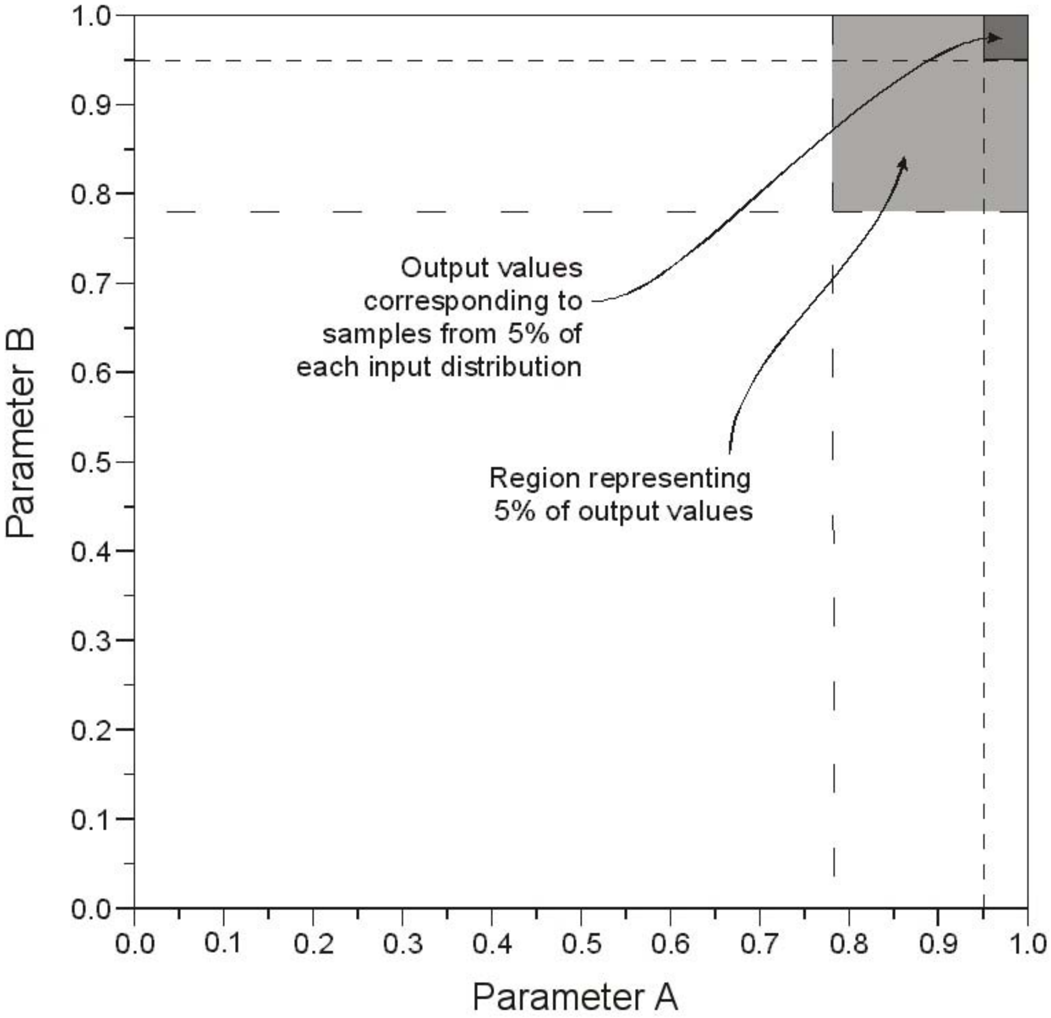


Figure 7 Illustration of stratified sampling, showing model space divided into uniform regions with one sample from each region.

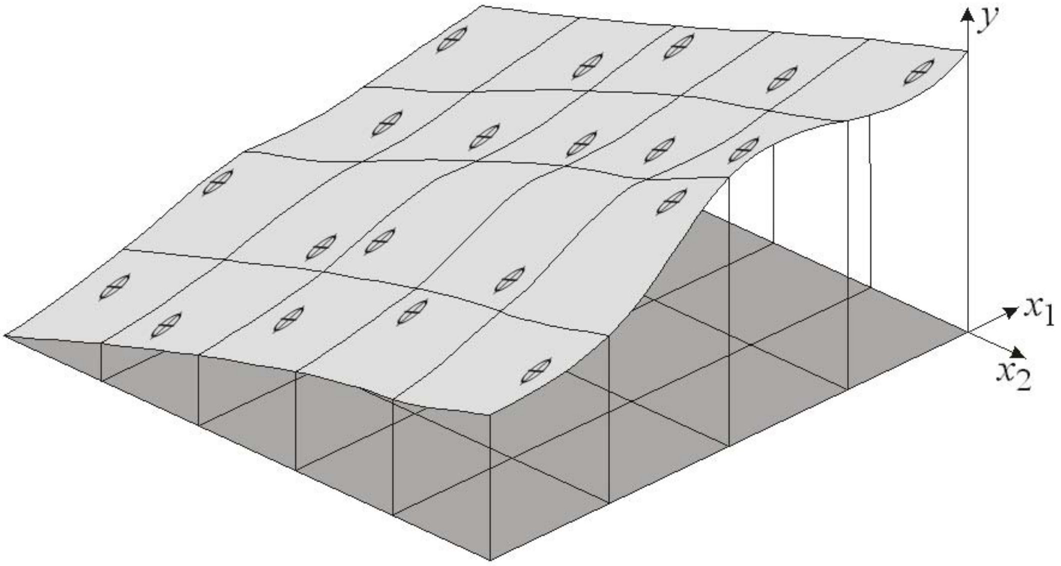


Figure 8 **Illustration of Importance Sampling. Model space (represented by contours) is divided into unequal regions that reflect importance to the expectation value.**

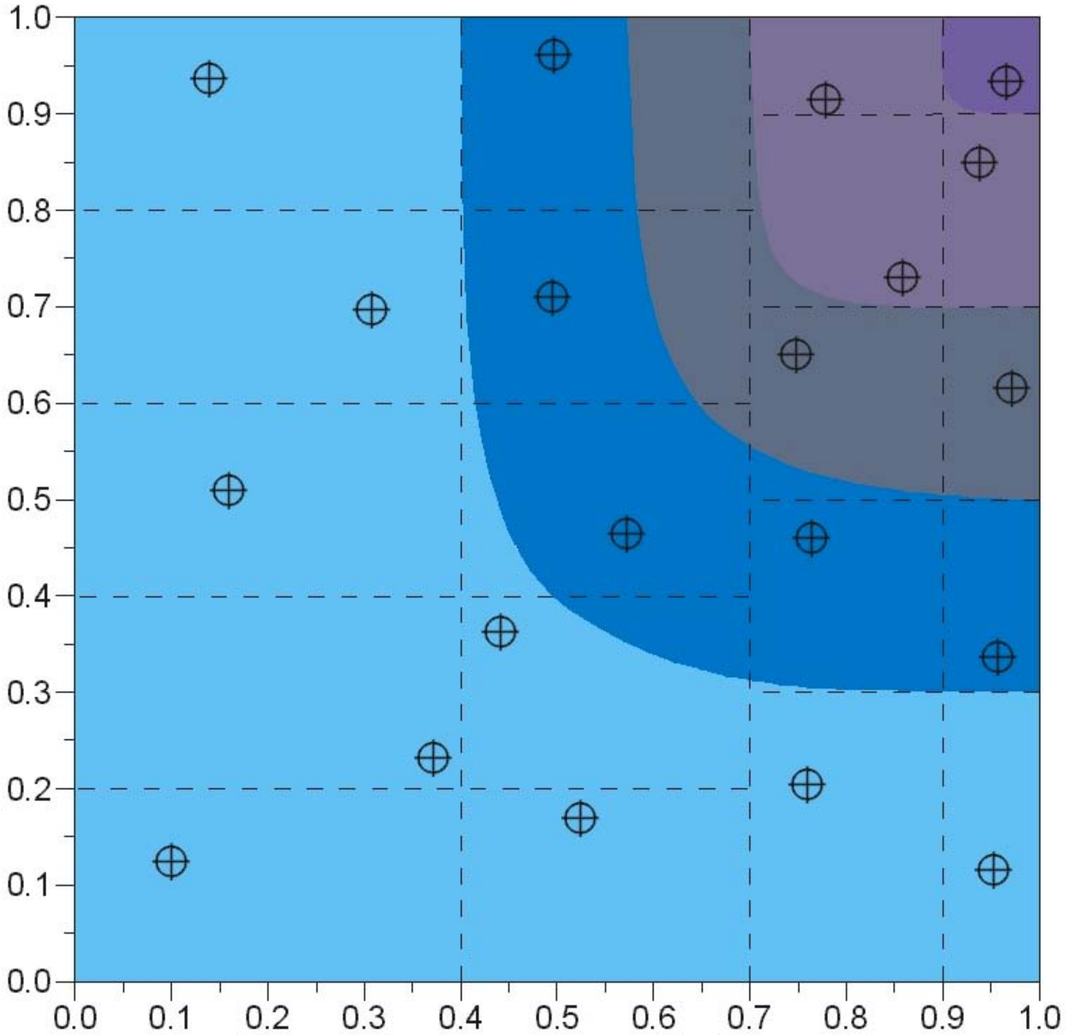


Figure 9 Illustration of Latin Hypercube Sampling, showing one sample from each equi-probable region.

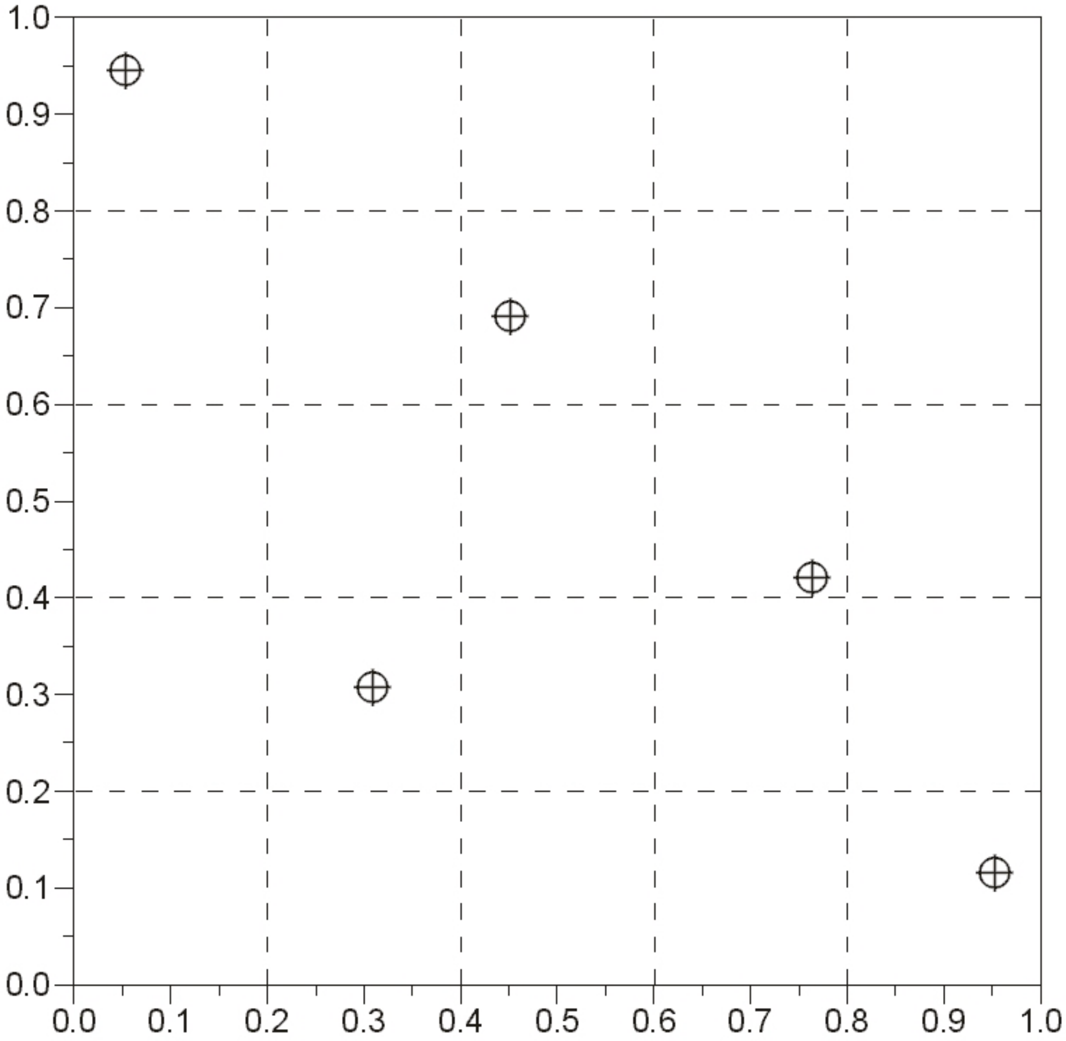


Figure 10 Selection between different parameter distributions by sampling of a control variable.

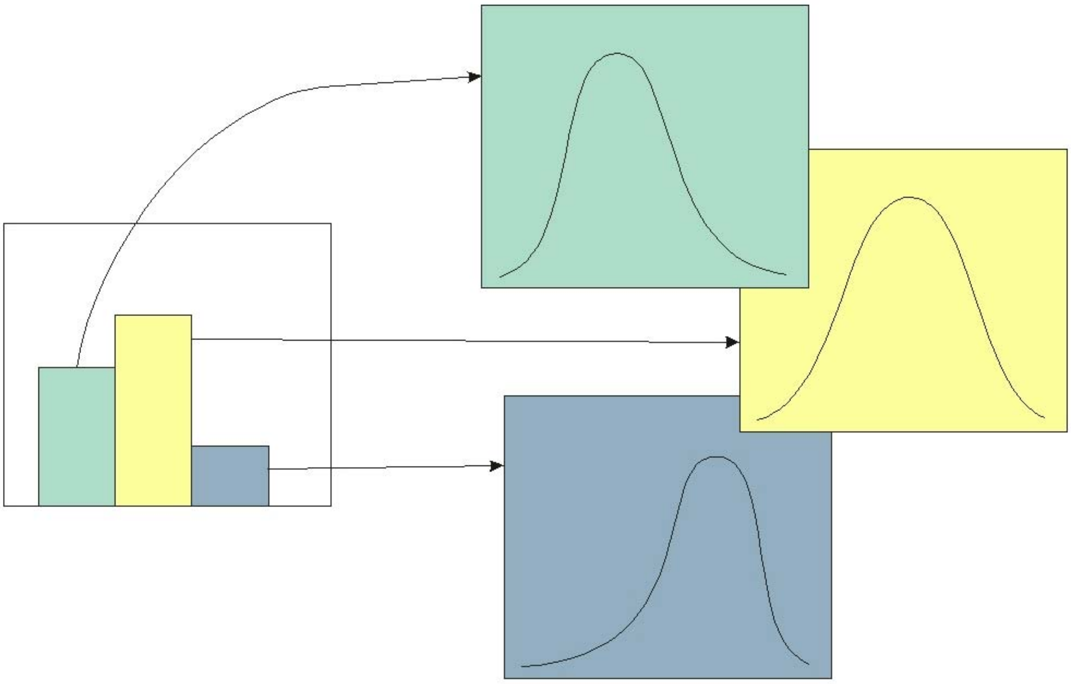


Figure 11 Example of an induced correlation between two normally-distributed parameters.

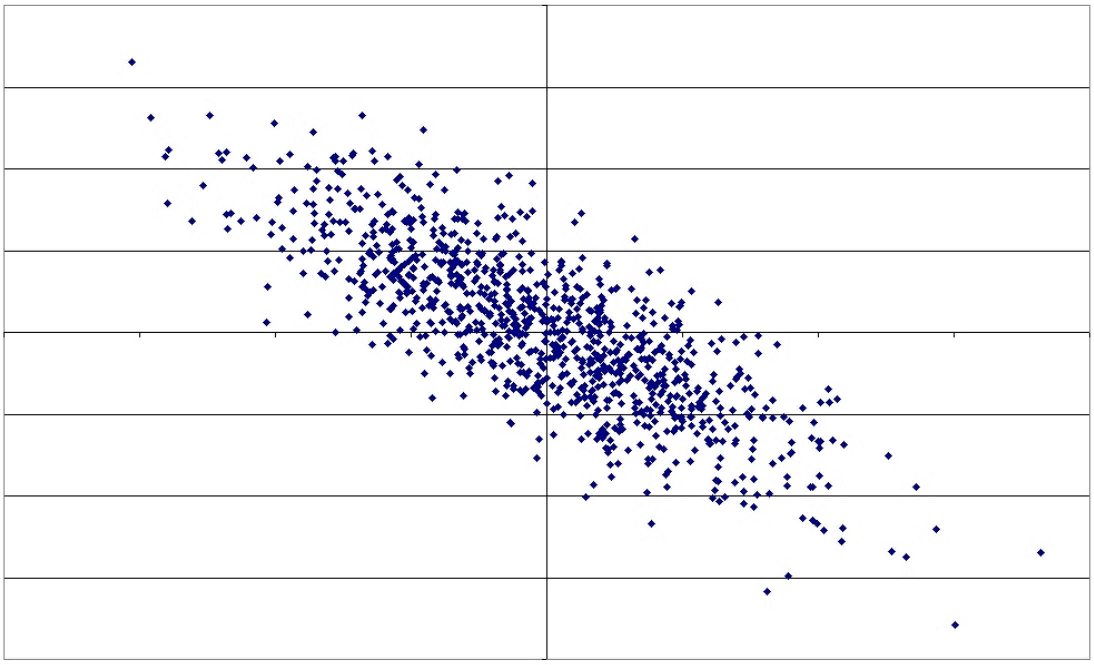


Figure 12 Measures of central tendency and their relationships for different distribution types.

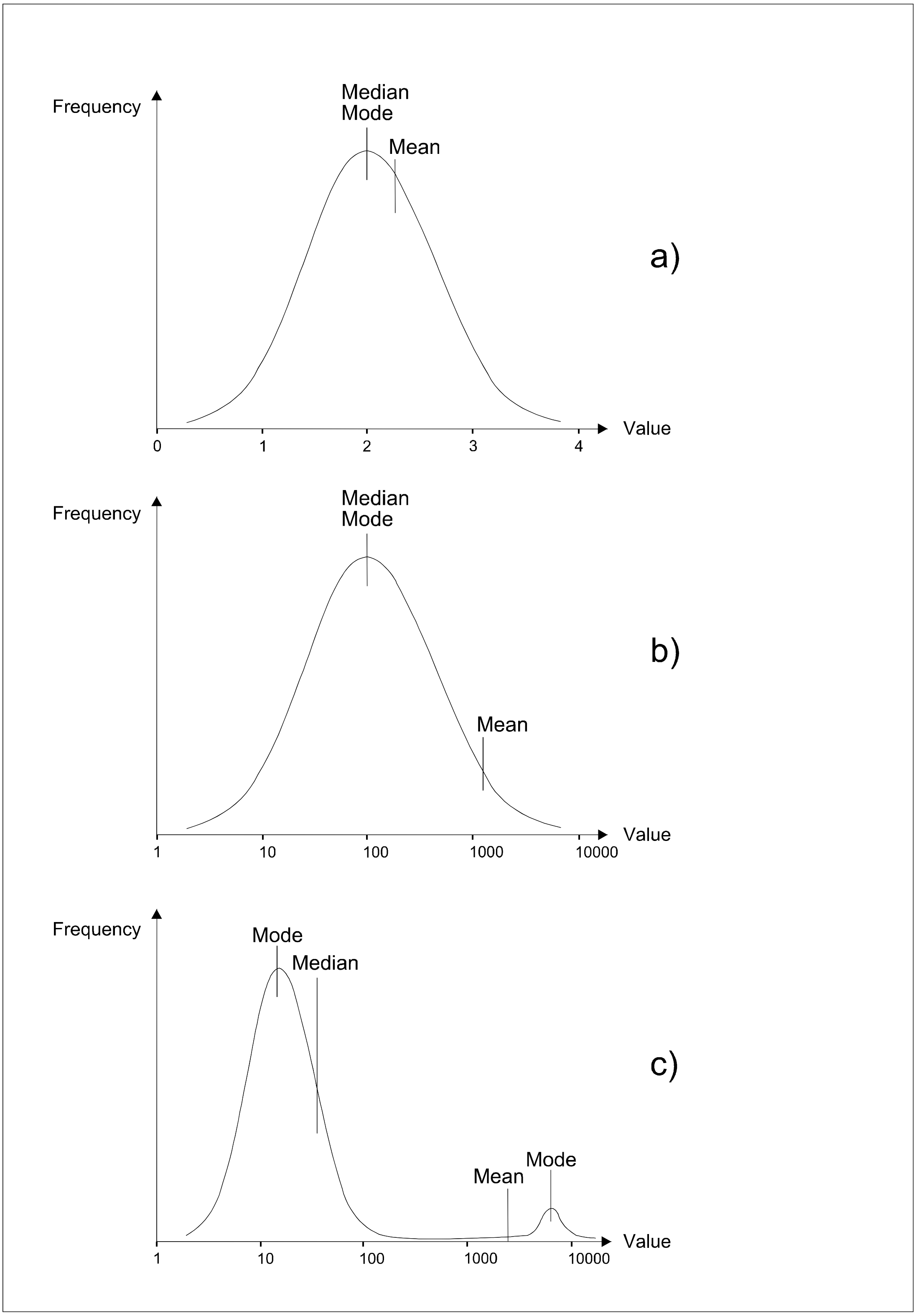


Figure 13 **Effect of varying the bin size on the resolution of a probability distribution function and the corresponding cumulative distribution function. 100 samples from a normal distribution were assigned to 30 (top), 15 (middle) and 7 (bottom) bins.**

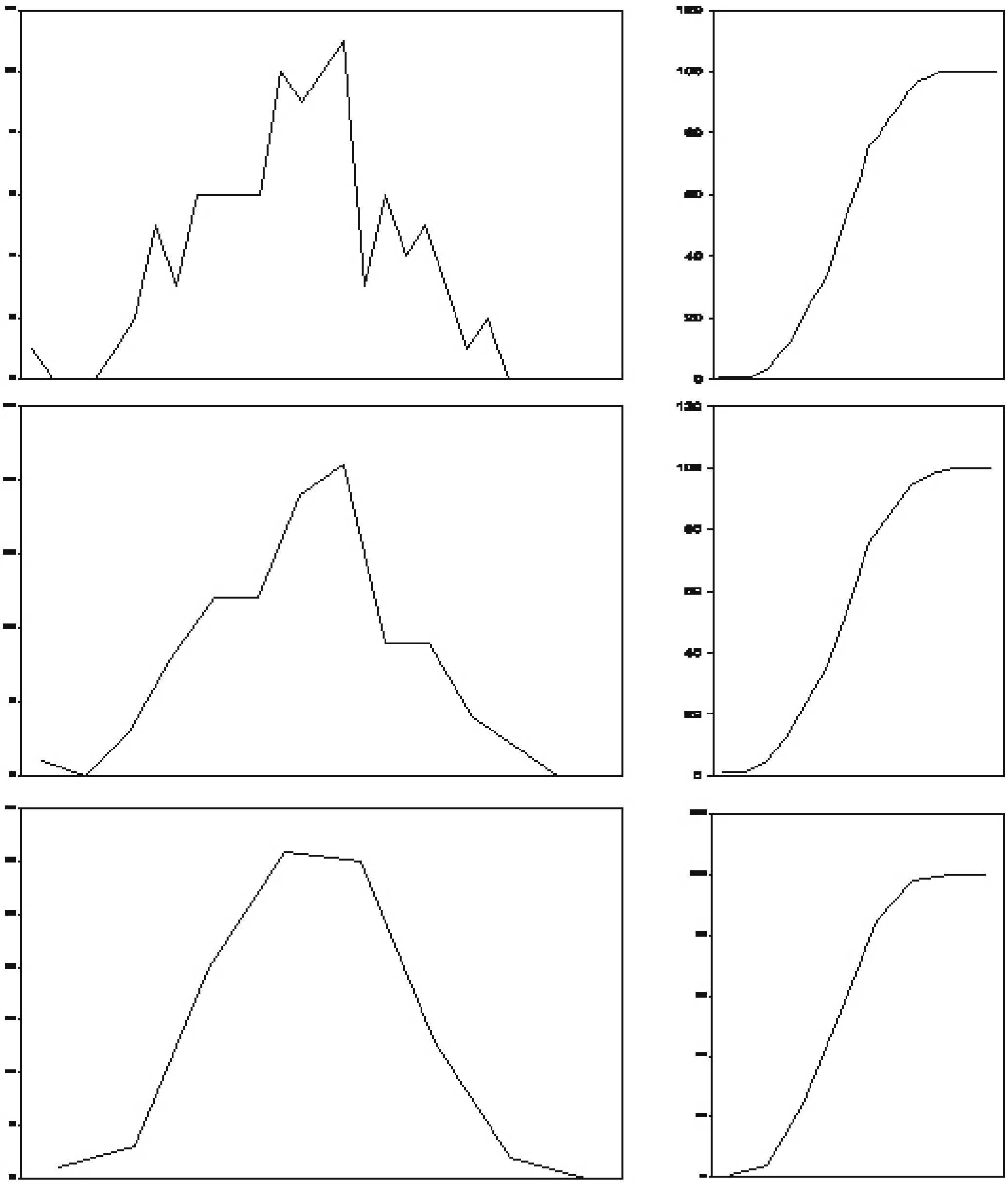


Figure 14 The relationship between the cumulative distribution function and the complementary cumulative distribution function.

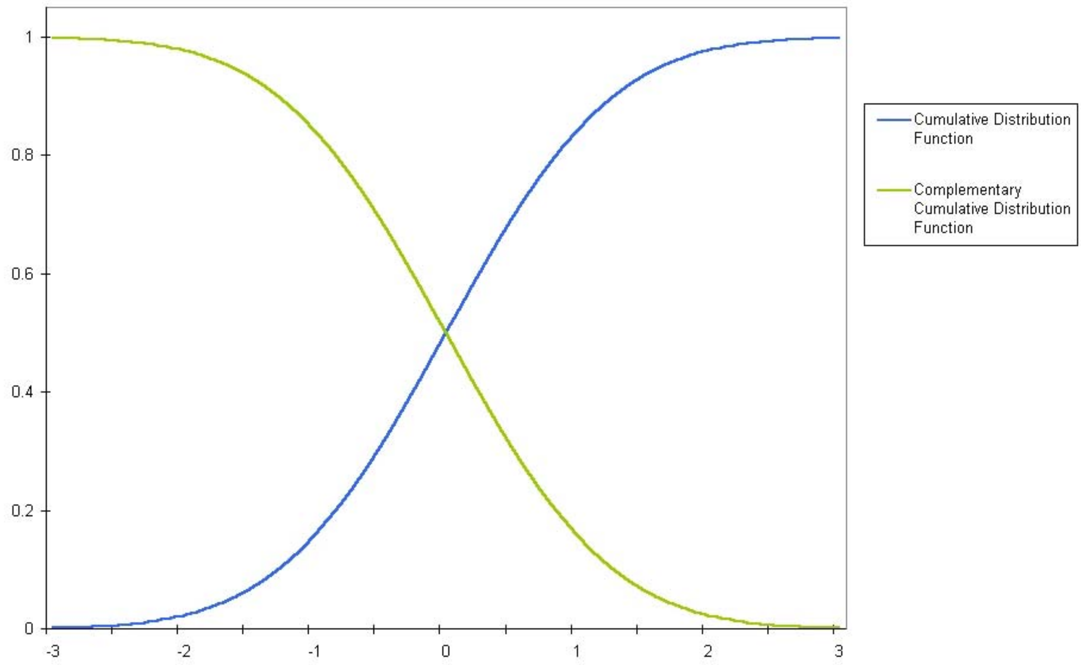


Figure 15 Example of a box-whisker plot summarising the distribution of an output parameter for 5 different cases.

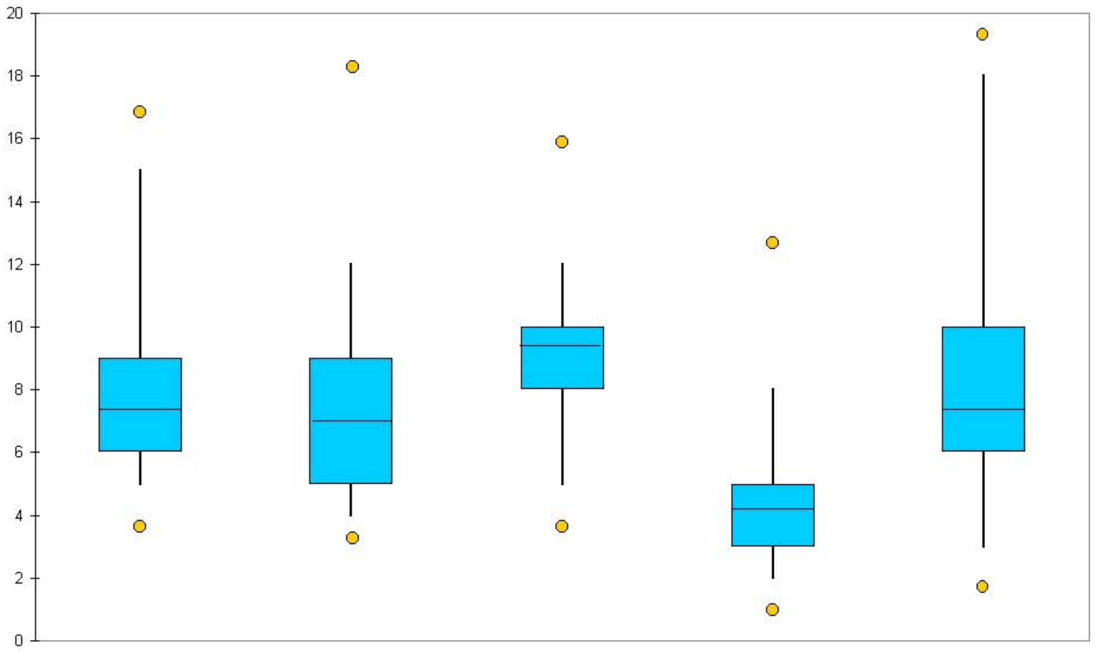


Figure 16 Illustration of how box-whisker plots can be used to display time-dependent distributions.

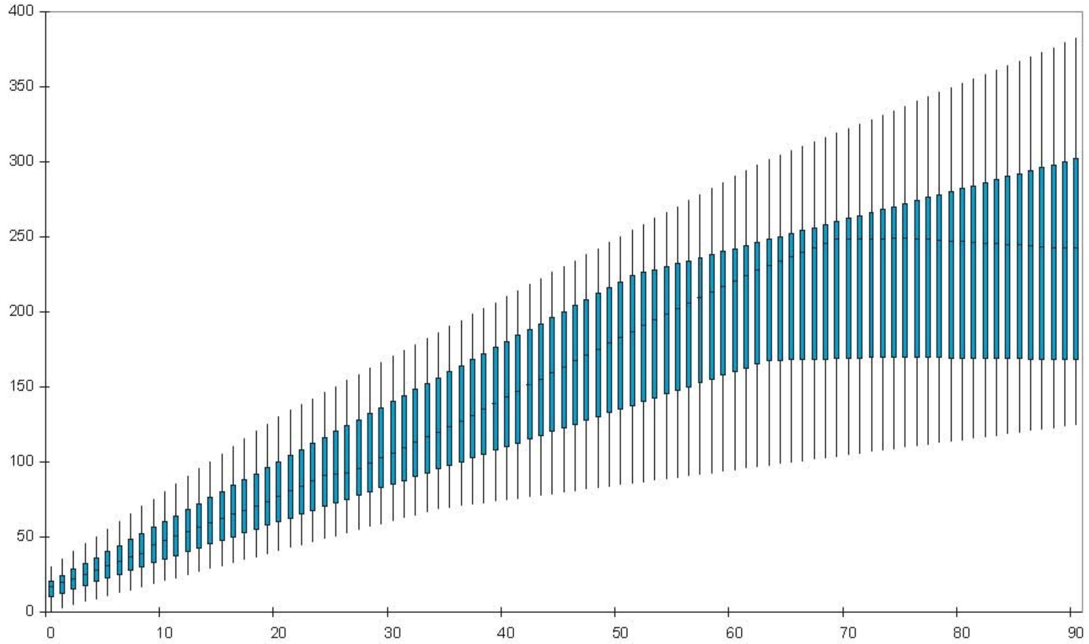


Figure 17 Example of plotting dose vs time curves to indicate overall range of behaviour while obscuring details of individual simulations.

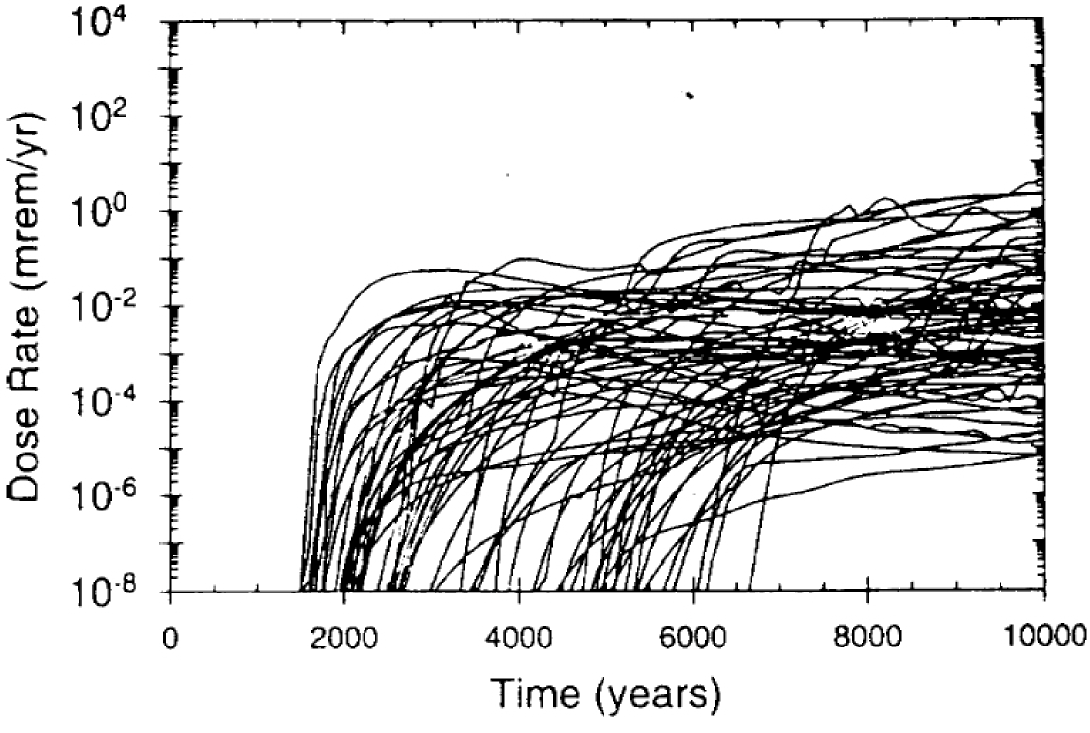
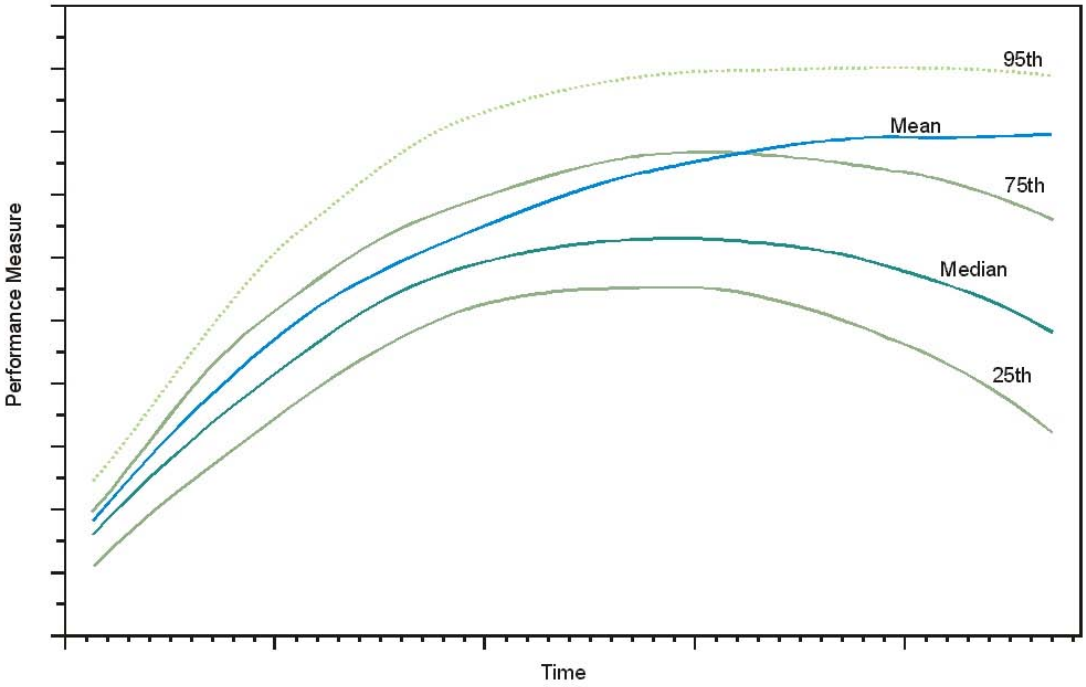


Figure 18 Schematic mean risk or dose vs time curve, illustrating how uncertainty can be shown by plotting quartiles and percentiles in addition to mean.



Appendix 1 Probabilistic Approaches

This report is based on a review of the probabilistic approaches used in a number of key performance assessments, together with a review of supporting documents and descriptions of probabilistic approaches used in other applications.

In the context of performance assessments for radioactive waste disposal, the key reason for adopting a probabilistic approach is generally seen as allowing the calculation of risk or other probabilistic end-point. There is, however, no necessary reason for probabilistic approaches to only be used for the calculation of risk. Any other end-point, such as dose, environmental concentrations or radionuclide fluxes, can be calculated using probabilistic methods. The calculated end-point can be expressed as a single (expectation) value, or as a distribution that shows the uncertainties accounted for in the analysis. Currently, however, there are no published assessments that use a probabilistic approach except where the regulatory end-point is risk or a similar concept. Intermediate results in these assessments may be examined and reported using probabilistic techniques, but these results are not the focus of the assessments.

Similarly, probabilistic calculations can be used for sensitivity studies even if the performance measure is calculated deterministically. Currently, however, there are no published assessments that use a probabilistic approach to sensitivity analysis in support of deterministic calculations of the regulatory end-point.

The assessments using probabilistic approaches that were reviewed were:

- United Kingdom: Her Majesty's Inspectorate of Pollution Dry Run 3 (Summerling 1992).
- United Kingdom: Nirex 97 (Norris *et al.* 1997).
- United States: Compliance Certification Application for the Waste Isolation Pilot Plant (DOE 1996).
- United States: Total System Performance Assessment for the Yucca Mountain Viability Assessment (DOE 1998).
- Canada: Postclosure Assessment of a Reference System (Goodwin *et al.* 1994).
- Sweden: SKB SR97 (SKB 1999)

Probabilistic methods were used in the SR97 assessment, although the approach was criticised for the way in which uncertainties were expressed as probabilities. However the assessment was published shortly after a risk criterion had first been introduced into the regulations, and was conducted as part of the process to develop an understanding of the key issues concerning repository safety, rather than as part of a license application. Further development of the approach is anticipated, and the present report forms part of the work being undertaken by the regulators in Sweden to further understand the issues involved so that these developments can be assessed in due course.