CNWRA *A center of excellence in earth sciences and engineering*

A Division of Southwest Research Institute™
6220 Culebra Road • San Antonio, Texas, U.S.A. 78228-5166
(210) 522-5160 • Fax (210) 522-5155

January 17, 2002
Contract No. NRC-02-97-009
Account No. 20.01402.765

U.S. Nuclear Regulatory Commission
ATTN: Mrs. Deborah A. DeMarco
Two White Flint North
11545 Rockville Pike
Mail Stop T8 A23
Washington, DC 20555

Subject: Programmatic review of paper for PSAM 6 conference titled "A Partitioning Method For Identifying Important Model Parameters"

Dear Mrs. DeMarco:

Attached is the subject paper which will be submitted for publication in the proceedings volume of the Probabilistic Safety Assessment and Management conference (San Juan, Puerto Rico, USA, June 23–28, 2002). The paper describes a novel and simple technique to perform sensitivity analysis of stochastic models. The method is applied to the analysis of the NRC performance assessment model. It is verified that a small set of model parameters is sufficient to explain the nature of the model output.

Please note that the deadline to submit this paper to the conference organizers is February 1, 2002. I will appreciate receiving your programmatic review before that. In the absence of your review we will submit the paper by February 1, with the condition that if found programmatically not acceptable by the NRC, we will withdraw the paper from publication.

Please contact Osvaldo Pensado at (210) 522-6084 if you have any questions regarding this paper.

Sincerely yours,

Budhi Sagar
Technical Director

Enclosure

cc: J.Linehan, E. Whitt, B. Meehan, J. Greeves, J. Piccone, T. Essig, W. Reamer, S. Wastler, R. Codell, D. Esh, C. Grossman, R.K. Johnson, T. McCartin, C. McKenney, J. Peckenpaugh, M. Rahimi, M. Thaggard, W. Patrick, CNWRA Directors, CNWRA Element Managers, S. Mohanty, P. LaPlante, M. Smith, R. Benke, O. Pensado, S. Mayer, O. Povetko, L. Howard

Washington Office • Twinbrook Metro Plaza #210
12300 Twinbrook Parkway • Rockville, Maryland 20852-1606

A Partitioning Method For Identifying Important Model Parameters

Osvaldo Pensado[1], Velin Troshanov[2], Gordon Wittmeyer[1], and Budhi Sagar[1]

[1]Center for Nuclear Waste Regulatory Analyses
Southwest Research Institute, 6220 Culebra Road
San Antonio, Texas 78238-5166, USA
[2]College of Liberal Arts and Sciences,
University of Illinois at Urbana-Champaign, USA

ABSTRACT

A general method for identifying important parameters of complex stochastic models is presented. The method is applied to the analysis of a performance assessment model of a geologic repository. A small set of important parameters is derived and it is verified that this small set is sufficient to explain the nature of the model output.

KEYWORDS

Sensitivity analysis, Monte Carlo simulation, stochastic model, geologic repository, performance assessment

INTRODUCTION

For complex models incorporating multiple stochastic parameters, it is frequently helpful to determine the parameters that most influence their output, and values at which the parameters become influential. In this paper, we describe a novel method for accomplishing this task. This technique, referred to as the partitioning method, has greater power in identifying possible correlations among input and output variables than traditional methods such as linear regression, is computationally simple, and can be efficiently programmed.

The motivation for the partitioning method was to develop a technique to analyze results of a model to assess the performance of a proposed geologic repository at Yucca Mountain, Nevada. The model, implemented in the Total-system Performance Assessment (TPA) code (Mohanty and McCartin, 2000), has over 300 parameters (some of them correlated) that have assigned probability distributions. This code is executed in a Monte Carlo mode using the Latin Hypercube method to sample values of stochastic parameters. The main output of the code is a large number of realizations, each realization consisting of total effective dose equivalent (TEDE) to a reasonably maximally exposed individual as a function of time. The mean and confidence bounds for the TEDE as functions of time can be derived

from these multiple realizations. Each realization is associated with a particular set of values of input parameters. In the United States, disposal regulations applicable to Yucca Mountain require that the peak of the mean TEDE within 10,000 years be below a specified value. In this paper, the partitioning method is used to identify the set of most important stochastic parameters affecting different attributes of the TEDE (simply referred to as the annual dose from here on).

DESCRIPTION OF THE PARTITIONING METHOD

An outline of the partitioning method is provided as follows. Partition the output realizations into two bins, one bin containing those realizations contributing the most to the mean annual dose (contributing realizations) and a second bin containing all the remaining realizations (non-contributing realizations). We explored four different approaches for defining "contributing" and "non-contributing" realizations, discussed later. Let A be the parameter whose importance is to be evaluated. Plot a cumulative distribution function and a complementary cumulative distribution function for the set of values of A that are associated with the contributing and non-contributing realizations, respectively. Let (x_A, P_A) be the coordinates of the intersection of these two curves. The probability value P_A can be used to measure the importance of the parameter A. For example, the importance index for parameter A, z_A, can be defined as

$$z_A = 0.5 - P_A \tag{1}$$

High values of $|z_A|$ (i.e., $|z_A|>0.1$) indicate an evident partitioning of parameter A into two subsets, related to the contributing and non-contributing realizations. The greater the value of $|z_A|$ the more important is variable A. Values of $|z_A| < 0.1$ suggest a lack of partitioning and a lack of importance of the parameter A. If $|z_A|$ is large (i.e., $|z_A|>0.1$) and z_A is positive (negative), then there is a positive (negative) correlation between the parameter and the mean annual dose. Direct comparison of $|z_A|$ yields the ranking of the most important parameters in the stochastic model. The intersection value x_A also has an interesting interpretation. If $|z_A|>0.1$ and z_A is positive, then there is a greater likelihood of $A>x_A$ for contributing realizations and $A<x_A$ for non-contributing realizations. In this sense, the value of the intersection, x_A, defines a partitioning value for the parameter A.

Four methods were explored to define contributing and non-contributing realizations. Method 1 was selected to detect the influence of any given realization on the peak of the mean annual dose, ignoring the time at which the peak may occur. Let p_{-i} be the maximum of the mean annual dose, computed without accounting for ith realization in the determination of the mean; i.e.,

$$p_{-i} = \max\left(\frac{1}{n-1} \sum_{\substack{j=1 \\ j\neq i}}^{n} d_j(t) \right) \tag{2}$$

n is the total number of realizations and $d_j(t)$ is the annual dose as a function of time for the jth realization. The function max has the usual mathematical meaning. The discriminating index associated with the ith realization, a_i, is defined as

$$a_i = \frac{p_{-i} - p_T}{p_T} \tag{3}$$

where p_T represents the peak of the mean annual dose during the simulation time. For Method 1, the contributing realizations are defined as those satisfying $|a_i| > a_\mu$, where a_μ is the mean of the set of $|a_j|$ values (j=1,2, ..., n). The non-contributing realizations are those for which $|a_i| \leq a_\mu$. It can be shown that $a_i \leq 1/n$. For most of the realizations considered in this paper, a_i is close $1/n$; thus, a_μ is also a number close to $1/n$.

Method 2 was selected to detect the influence of a realization on the peak of the mean dose, at the time at which the peak occurs. Let t_T be the time at which the peak dose, p_T, occurs. The discriminating index for the ith realization, b_i, is defined as

$$b_i = \frac{d_i(t_T) - p_T}{p_T} \tag{4}$$

$d_i(t_T)$ is the annual dose for the ith realization evaluated at time t_T. The contributing realizations are defined as those satisfying $b_i > 0$ and the non-contributing are all of the others. Since for the majority of the realizations considered in this paper the annual dose is negligible compared to the peak of the mean annual dose, b_i is in general close to -1.

Method 3 was designed to highlight influences on the mean dose over the complete simulation period. The norm in the space of continuous functions in the interval $[0, t_{max}]$ is defined as

$$\|f\| = \sqrt{\int_0^{t_{max}} \{f(t)\}^2 dt} \tag{5}$$

t_{max} is the maximum time of the simulation period and f is a continuous function. Let $d_\mu(t)$ represent the mean annual dose as function of time. The discriminating index for the ith realization, c_i, is defined as

$$c_i = \frac{\|d_i - d_\mu\|^2}{\|d_\mu\|^2} \tag{6}$$

Chun et al. (2000) used an expression similar to Eqn. 6 to measure changes in output cumulative distribution functions. Let c_μ be the mean value of the set of c_j values (j=1,2,...,n). The contributing realizations are selected as those for which $c_i > c_\mu$ and the non-contributing realizations are all of the others. Since for the majority of the realizations considered in this paper the annual dose is negligible compared to the mean annual dose, c_i is in general close to one.

Method 4 was also designed to highlight the influences on the mean annual dose over the complete simulation period. The discriminating index for the ith realization, $\bar{c}_i$, is defined as

$$\bar{c}_i = \frac{\|d_{-i} - d_\mu\|^2}{\|d_\mu\|^2} \tag{7}$$

d_{-i} is computed as

$$d_{-i} = \frac{1}{n-1} \int_0^{t_{max}} \sum_{\substack{j=1 \\ j \neq i}}^{n} d_j(t) \tag{8}$$

The contributing realizations are defined as those having values of $\bar{c}_i$ greater than the mean of the set of $\bar{c}_j$ values (j=1,2,...,n). It can be shown that $\bar{c}_i \approx c_i/n^2$, provided that the number of realizations, n, is large enough. Thus, Method 4 is equivalent to Method 3; the yield identical results if the number of realizations is large, as is the case in this paper.

RESULTS

Data generated with 4000 realizations of the TPA Code Version 4.1j were used for the analyses. In this version of the TPA code, 330 stochastic input parameters are considered in the non-disruptive base case. In general, the discriminating indices (i.e., a_i, b_i, and c_i) defined above tend to be close to a constant value (i.e., 1/4000, -1, 1, respectively). Thus, linear regression between the discriminating index and parameter values yields a slope that is not clearly different from zero. In other words, linear regression cannot be used to identify a correlation between parameter values and the discriminating index. On the other hand, the partitioning is capable of detecting correlations, if they exist.

The importance indices, z_A, were computed for all of the stochastic parameters using the three methods defined above (methods 3 and 4 are equivalent). The parameters were sorted according to decreasing values of $|z_A|$. The most important parameters are those with highest values of $|z_A|$. The list of the most important parameters for 10,000 year and 100,000 year realizations are included in Table 1.

TABLE 1
LIST OF MOST IMPORTANT PARAMETERS

Parameter	100,000 yr	10,000 yr	Correlated	Meaning
Preexponential_SFDissolutionModel2	×	×		Factor modulating the spent fuel dissolution rate
AlluviumMatrixRD_SAV_Np	×	×	C_1	Retardation coefficient for Np in the alluvium
SubAreaWetFraction	×	×	B_1	Related to the amount of water at the drift
AA_1_1[C/m2/yr]	×			Corrosion rate of Alloy 22
ArealAverageMeanAnnualInfiltrationAtStart [mm/yr]	×	×	B_2	Mean annual infiltration for current climate
AlluviumMatrixRD_SAV_Pu	×	×	C_2	Retardation coefficient for Pu in the alluvium
AlluviumMatrixRD_SAV_Am	×	×	C_2	Retardation coefficient for Am in the alluvium
AlluviumMatrixRD_SAV_U	×	×	C_2	Retardation coefficient for U in the alluvium
DistanceToTuffAlluviumInterface[km]	×	×		Related to location of tuff/alluvium interface
WastePackageFlowMultiplicationFactor	×	×		Related to the amount of water for release
MatrixPermeability_TSw_[m2]	×	×	B_2	Matrix permeability for Topopah Spring tuff-welded
WellPumpingRateAtReceptorGroup20km [gal/day]*	×	×		Well pumping rate for farming receptor group located at a distance greater than 20 km
AlluviumMatrixRD_SAV_Th	×		C_2	Retardation coefficient for Th in the alluvium
FractionOfCondensateTowardRepository[1/yr]	×	×		Fraction of condensed water moving towards the repository
ImmobilePorosityPenetrationFraction_STFF	×			Effective fraction of saturated rock matrix accessible to matrix diffusion
MatrixKD_UCF_Am[m3/kg]	×			Matrix sorption coefficient (Upper Crater Flat) for Am

SolubilityNp[kg/m3]	×			Solubility of Np
InterceptionFraction/Irrigate	×	×		Fraction of irrigation interception
DefectiveFractionOfWPs/cell		×		Related to the number of waste packages assumed initially failed
DripShieldFailureTime[yr]		×		Time of failure of the drip shield
FractionOfCondensateRemoved[1/yr]		×		Fraction of condensed water not intersecting the drifts
RntoDetermineFaultOrientation		×		Random number to determine fault orientation
* In the TPA Code Version 4.1j, non-disruptive base case, the pumping rate is sampled from a probability distribution function. United States regulations for the proposed Yucca Mountain Site require that the pumping rate of well water considered for performance assessment analysis be a particular fixed value. Future versions of the TPA code will be made consistent with this recent regulatory requirement.				

The parameter names in Table 1 are the same as those used in the TPA Code. The three methods coincide in that pre-exponential factor modulating the rate of spent fuel dissolution is the most important parameter (listed as the first entry in Table 1). If a parameter ranked within the first 20, for at least two of the three methods, such a parameter was included in Table 1. The three methods coincide in the top nine (five) parameters —indicated in bold (italic) font in Table 1— for 100,000 (10,000) year simulations, although the ranking is slightly different from method to method. The 18 (17) most important parameters for 100,000 (10,000) year simulations are indicated by the label × under the 100,000 (10,000) year column in Table 1. In the non-disruptive base case, the parameters labeled with B_2 and C_2, under the Correlated column, are correlated to the parameters labeled with B_1 and C_1, respectively. Parameters labeled with B_2 and C_2 appear important because they are correlated to the parameters labeled with B_1 and C_1, as is shown later. Several runs of the TPA code were completed to verify that the parameters in Table 1 are sufficient to reproduce the variance of the annual dose and the magnitude of the mean annual dose. The results are reported in Figure 1.

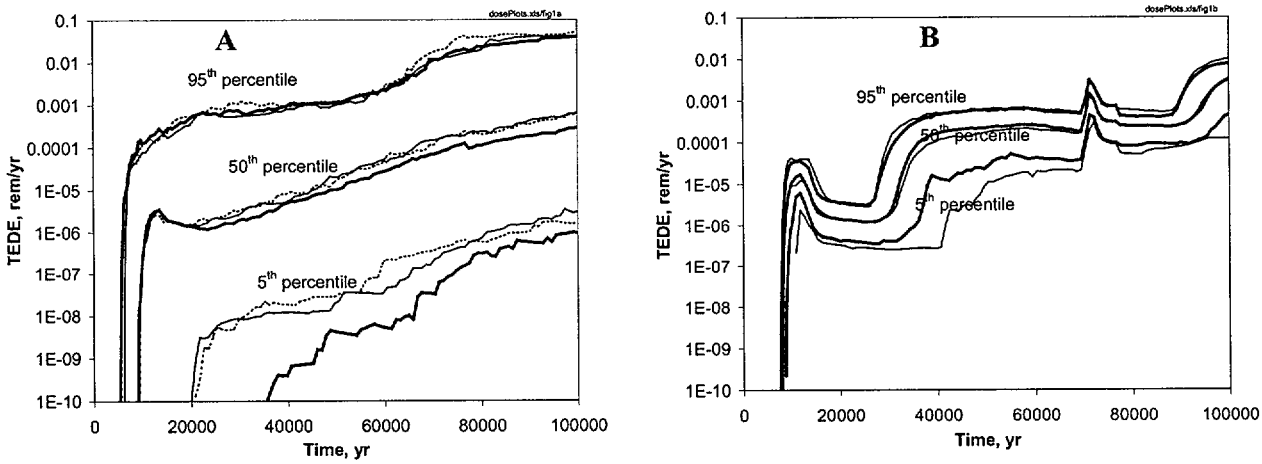

Figure 1: Plots of the 5th, 50th, and 95th percentile of the annual dose versus time.
(A) Base case, and cases 1 and 2.
(B) Cases 3 and 4. See main text for the definition of cases 1 to 4.

Figure 1 includes 5th, 50th, and 95th percentile curves for the annual dose versus time. Figure 1-A presents results for the base case and cases 1 and 2. The thick lines are associated to results of the non-disruptive base case (500 realizations). For the case 1 (thin lines), the important parameters in Table 1 were sampled stochastically (300 realizations) in the range defined in the base case, with the exception of those parameters labeled with B_2 and C_2. All of the other parameters, including those labeled with B_2 and C_2, were fixed at their mean values. A total of 16 parameters were sampled. For the case 2 (dotted lines in Figure 1-A), the parameters in Table 1 in bold or italic font, except those labeled with B_2 and C_2, and the parameter DripShieldFailureTime[yr] were sampled stochastically

(300 realizations). All of the other parameters, including the B_2 and C_2 parameters, were fixed at their mean values. Thus, a total of 8 parameters were sampled for the case 2. The confidence intervals for the base case and cases 1 and 2 are very similar, with relevant variations only in the 5^{th} percentile curves. It is concluded that 8 parameters suffice to account for the gross variance of the annual dose. The 50^{th} and 95^{th} percentile curves compare within less than an order of magnitude for the three cases. Furthermore, the mean annual dose curves (not included in Figure 1) are almost the same for the three cases (they differ by much less that an order of magnitude at all times for the three cases).

Figure 1-B, includes results for cases 3 and 4. For case 3 (thick lines), all of the important parameters in Table 1 were fixed at their mean values (a total of 22 parameters) and all of the others were sampled. Case 4 (thin lines) is the reverse to case 2; the 8 parameters of the case 2 were fixed at their mean values, and the remaining 322 input parameters were sampled. Figure 1-B summarizes 300-realization runs.

In Figure 1 it is noted that the variance in the annual dose deriving from the variance of 322 parameters is small compared to that resulting from the variance of the 8 most important parameters identified in cases 2 and 4. Some of the parameters (those with labels B_2 and C_2 in Table 1) are ranked high by the partitioning method because they are correlated to important parameters. Method 3 was capable of ranking the parameter associated with the failure time of the drip shield within the highest six parameters, because it was designed to identify parameters affecting the annual dose in the complete simulation period, as opposed to methods 1 and 2, which focus on the peak of the mean annual dose. The partitioning method succeeded in identifying a small set of parameters controlling the variance of the annual dose.

CONCLUSIONS

The partitioning method for identifying important parameters was discussed. Although there was a direct motivation to analyze the performance assessment model of a geologic repository, the method is quite general and can be applied to the analysis of any data in which the output depends upon stochastic input parameters. In the particular example of the geologic repository, the partitioning method indicates that the mean annual dose rate is influenced most by parameters controlling the rate of release of radionuclides, corrosion rates of container materials, the amount of water available for radionuclide transport, and retardation coefficients for neptunium.

ACKNOWLEDGEMENTS

This paper was prepared to document work performed on behalf of the U.S. Nuclear Regulatory Commission (NRC), Office of Nuclear Material Safety and Safeguards, under Contract No. 02–97–009. The views expressed in this paper are those of the authors and do not necessarily reflect the views or regulatory position of the NRC.

REFERENCES

Chun M.-H., Han S.-J., and Tak N.-I. (2000). An uncertainty importance measure using a distance metric for the change in cumulative distribution function. *Reliability Engineering and System Safety.* **70**, 313-321

Mohanty S. and McCartin T. J. Coordinators. (2000). Total-system Performance Assessment (TPA) Version 4.0 Code: Module Description and User's Guide (Draft), Center for Nuclear Waste Regulatory Analyses, San Antonio, Texas