EPA/600/R-98/106

QUALITY ASSURANCE OF MULTI-MEDIA MODEL FOR PREDICTIVE SCREENING TASKS

by

J Chen and M B Beck*

Department of Civil Engineering
Imperial College, London SW7 2BU, UK

* Warnell School of Forest Resources, University of Georgia
Athens, Georgia 30602-2152, USA

Cooperative Agreement # CR 816572-010
"Analysis of Uncertainty in Environmental Simulation"

Project Officer

Thomas O Barnwell Jr.
Assistant Laboratory Director
National Exposure Research Laboratory
960 College Station Road
Athens, Georgia 30605-2700

Notice

The U.S. Environmental Protection Agency through its Office of Research and Development partially funded and collaborated in the research described here under Cooperative Agreement # CR 816572-010 with Imperial College of Science, Technology and Medicine, London, UK. It has been subjected to the Agency's peer and administrative review and has been approved for publication as an EPA document.

Abstract

Priorities must be determined for the ways in which limited resources can be deployed in the most cost-effective manner. In the case of potential contamination of groundwater by leachates from facilities for storing hazardous materials, there are many more sites where action might be taken to reduce risks of exposure than there are funds to support all such actions. There is a need to rank the sites of potential action in terms of achieving the greatest reduction in the risk of exposure for a given sum of money. In situations such as this, which are characterised by gross uncertainty, assessing the reliability of a model in performing the task of a screening analysis is especially important. The risks of ranking the sites for remedial action in an erroneous order are significant. The paper explores three groups of tests that might be formulated to determine model reliability. The first of these is concerned with establishing whether the uncertainties surrounding the parameterisation of the model render it impotent in discriminating between which of two sites, say, gives the significantly higher predicted receptor concentration of contaminant, in conditions where this result would generally be expected. The second test is a straightforward form of regionalised sensitivity analysis designed to identify which of the model's parameters are critical to the task of predicting exceedance, or otherwise, of prescribed (regulatory) receptor-site concentrations. The third test is designed to achieve a more global form of sensitivity analysis in which the dependence of selected statistical properties of the distributions of predicted concentrations (mean, variance, and 95th-percentile) on specific model parameters can be investigated. The results of these tests suggest that it may be possible to develop a novel form of statistic for assisting in judging the trustworthiness of a candidate model for performing predictive exposure assessments.

FOREWORD

As environmental controls become more costly to implement and the penalties of judgment errors become more severe, environmental quality management relies increasingly on the use of predictive models to estimate the impact of contaminant releases to the environment. Further, as the questions of exposure and risk become more comprehensive the tools necessarily become more complex. This, in turn, puts increased pressure on appreciating and quantifying the uncertainties associated with model predictions. As part of this Division's research on the occurrence, movement, transformation, impact, and control of environmental contaminants, the Regulatory Support Branch develops engineering tools to help pollution control officials address environmental problems.

The first step in addressing the potential impact of environmental releases of contaminants is the selection of an appropriate predictive tool (i.e., model). The model must satisfy various criteria including relevance, reliability, and validity. Using EPAMMM, a multimedia model for simulating the fate and transport of contaminants, this report explores three groups of tests, that combined, formulate a measure of reliability, or trustworthiness of a model. The results of these tests suggest that it may be possible to develop a novel form of statistic for assisting in judging the trustworthiness of a candidate model for performing predictive exposure assessments.

Rosemarie C. Russo, Ph.D.
Director
Ecosystems Research Division
Athens, Georgia

Contents

Tables

Figures

1 Introduction

Contamination of the subsurface land environment as a result of leakage from sites used to contain and store hazardous materials has been a predominant feature in the development of mathematical models of soil and groundwater systems over the past fifteen years or so (Onishi *et al*, 1990; National Research Council, 1990; Gee *et al*, 1991; McLaughlin *et al*, 1993). At a strategic level, the key questions to be answered are: what are the direction, rate of movement, and attenuation of contaminants in the plume; what level of contaminant concentration will result at a particular receptor site; and, when contamination is forecast to be unacceptable, at which repositories will the greatest reduction in the risk of adverse exposure be achieved through expenditures on control and remediation measures?

Given the large number of storage sites and a debilitating lack of in situ field observations -- that is, given conditions of gross uncertainty -- this last question is extremely difficult to answer. Yet answered it must be, in spite of such difficulties, if decision-making is to be guided by the support of the best possible information and quantitative analysis. To this end, the United States Environmental Protection Agency (EPA) has developed a Multi-Media model, which is in principle capable of predicting the propagation of a contaminant via several pathways through multiple compartments of the environment (subsurface water, surface water, atmosphere) from source to receptor (Salhotra *et al*, 1990; Sharp-Hansen *et al*, 1990). In this paper, we explore the suitability of this Multi-Media model (abbreviated as EPAMMM herein) for performing the various tasks of a screening-level analysis, where, according to the EPA's guidelines on exposure assessment, such an analysis is defined in the following terms (EPA, 1991):

> A primary concern in selecting a model is whether to perform a screening study or to perform a detailed study.
>
> The value of the screening-level analysis is that it is simple to perform and may indicate that no significant contamination problem exists. Screening-level models are frequently used to get a first approximation of the concentrations that may be present. Often these models use very conservative assumptions; that is, they tend to over-predict concentrations or exposures. If the results of a conservative screening procedure indicate that predicted concentrations or exposures are less than some predetermined "no concern" level, then a more detailed analysis is probably not necessary. If the screening estimates are above that level, refinement of the assumptions or a more sophisticated model are necessary in further iterations for a more realistic estimate.

While EPAMMM is not what would normally be described as the simple form of model best suited to a screening-level analysis (such as that of, say, Schanz and Salhotra (1990)), it is nevertheless nowhere nearly as complex as some of the contemporary alternatives for simulating subsurface contaminant transport (e.g., Ewen, 1995). Even so,

despite its relative simplicity, there is still a need to establish the relevance (or "legitimacy", "validity", or "trustworthiness") of EPAMMM for performing a screening-level analysis[1]. Issues of model validation are of very considerable topical interest, not least in the present subject area of contamination of the subsurface environment (Konikow and Bredehoeft, 1992; Dougherty and Bagtzoglou, 1993; Oreskes *et al*, 1994; Beck *et al*, 1995; Armstrong *et al*, 1995). Our purposes herein are, therefore, several: to present key elements in a protocol for validating models for predictive exposure assessment (Beck *et al*, 1997); to illustrate the application of these in the specific case of EPAMMM; and thence to examine the process of coming to a judgement on the trustworthiness of EPAMMM in fulfilling its designated task of identifying sites that would be prime candidates for risk reduction.

The paper begins by summarising the properties and assumptions of EPAMMM. We then set out the kinds of questions to be answered in order to assure the quality, or establish the trustworthiness, of this model as a tool designed for performing the tasks of a screening-level analysis. In particular, these questions are discussed in relation to what is normally understood as an analysis of sensitivity and model uncertainty (Beck, 1987). Computational results illustrating the performance of the model for a generic form of a Subtitle D storage facility are presented. It may be helpful to view this analysis as rather like the testing of prototype air-frame designs in a laboratory wind tunnel prior to construction of the airplane that is actually to perform the task of achieving flight. Our central concern is to illustrate how the potential "flight-worthiness" of the model might be established prior to its use in a practical decision-making context. In closing the paper, we shall discuss our results and the form of the analysis in the light of the ideas expressed elsewhere with regard to quality assurance and the development of a protocol for model validation in predictive exposure assessments (Beck *et al*, 1995, 1997). At issue here is what we see as the pressing need to broaden the procedure of validation in two important directions: first, in developing quantitative measures of the trustworthiness of a model when there are no historical data to be matched by its simulated responses; and, second, in augmenting and buttressing the process of peer review, which may be critical in such contexts where extrapolation into the utterly unknown is the essential part of the problem definition (as it is in predictive exposure assessments; Beck *et al*, 1997).

It is important to emphasise that this analysis is of a general nature. It is applicable, in principle, to any form of simulation model, not merely EPAMMM in the assessment of the risks associated with sub-surface contamination. Further, we make no specific statements about the suitability of a particular disposal facility with given surrounding soils and hydrological regime for attenuating the off-site mobility of a given contaminant.

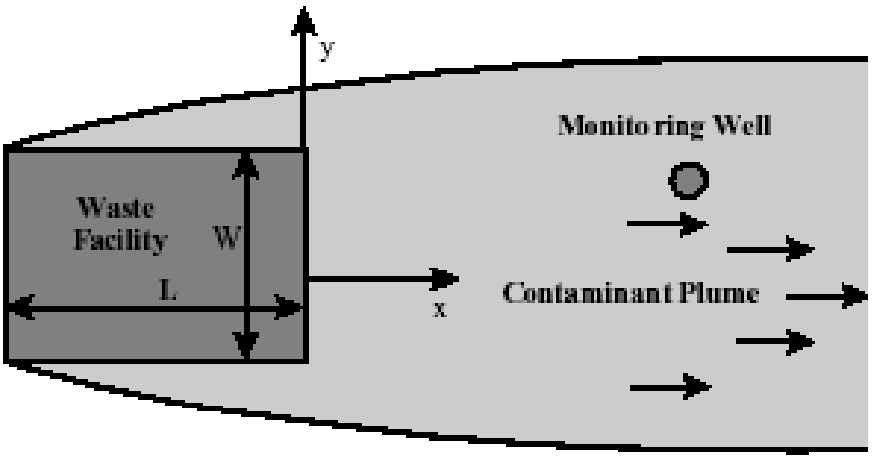
[1] As noted elsewhere (Beck *et al*, 1997), it seems the intractability of the problem of model validation has given rise to many labels for the process yet few entirely satisfactory procedures for its resolution.

This was not our purpose; site-specific data were not available to us for evaluation of the model; and, in any case, a screening-level analysis is by definition a generic, non-specific problem.

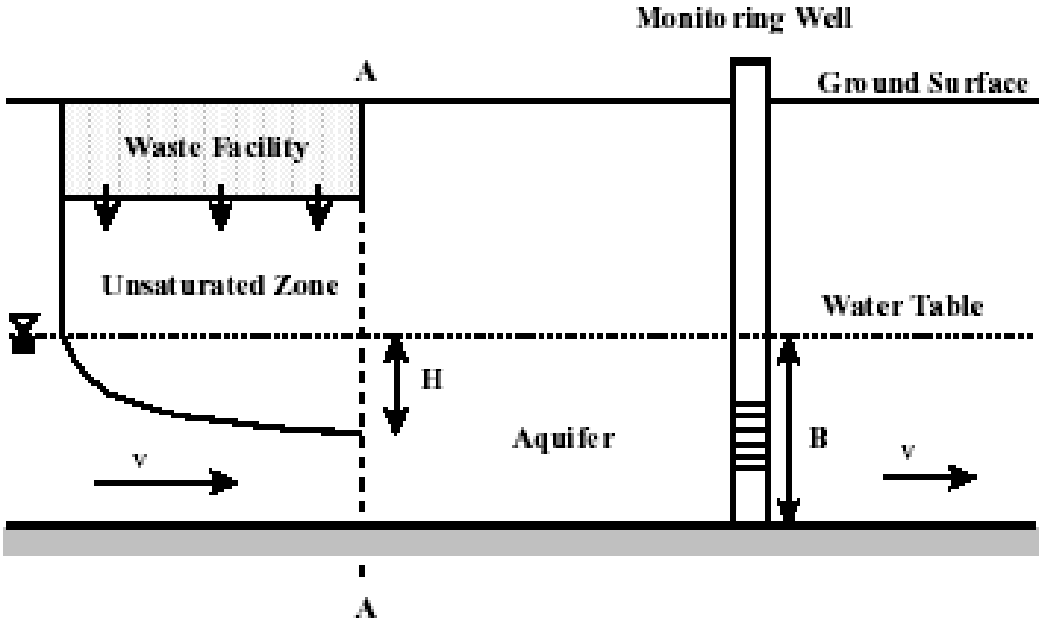
2 The Model

EPAMMM, as we have already noted, is a tool for predicting the transport and fate of contaminants released from a waste disposal facility into an environment composed of several media. Releases may be to the air or subsurface environment, the latter including both unsaturated and saturated zones, with the possibility of interception of the subsurface contaminant plume by a surface water system. The model contains seven modules: the landfill unit; the flow field in the unsaturated zone; the transport of solutes in the unsaturated zone; the transport of solutes in the saturated zone; the surface water system; an air emissions module; and the advective transport and dispersion of the contaminant in the atmospheric environment. Parallel developments on other forms of multi-media models are reviewed in Onishi *et al* (1990), Smith (1992), and Davis *et al* (1993). In general, analytical and semi-analytical techniques are used to solve the basic partial differential equations of fluid flow and solute transport. As a consequence of the associated simplifications, the model cannot account **explicitly** for the following: spatial variability of either soil or hydrological properties, such as, for example, porosity and hydraulic conductivity; specific geometries of the landfill site (other than rectangular); site-specific boundary conditions, i.e., inter alia, the spatial variability of both the infiltration rate and the depth of the unsaturated zone; multiple aquifer bodies; or wells at which pumping operations may take place. Further, flow through fractured media and chemical interactions **among** multiple contaminants cannot formally be simulated by the model.

As applied herein, that is, to the characterisation of a Subtitle D facility, only three of the above seven modules of EPAMMM will be used: flow in the unsaturated zone; transport of solute in the unsaturated zone; and transport of the solute in the saturated zone. Thus, the following generic situation is simulated (Figure 2-1). A storage facility of rectangular shape ($L \times W$, in plan view in Figure 2-1(a)) is sited above the unsaturated zone. The contaminated leachate from the facility infiltrates the ground uniformly from this area (alone), passing into the underlying unsaturated zone in the vertical direction only. When the leachate reaches the saturated zone it enters a horizontal flow of water driven by input, laterally-oriented recharge (Figure 2-1(b)). Mixing of the two flows is such that the contaminant in the leachate is assumed to penetrate the saturated zone to a maximum depth (H, the source penetration depth) at the rightward boundary of the source area in Figure 2-1(b). This rightward (downstream) boundary of the storage area, i.e., vertical section A-A in Figure 2-1(b), is given as the leftward (upstream) boundary of flow in the saturated zone in Figure 2-2, in which the contaminant from the leachate can be seen to occupy the shaded portion ($H \times W$) of the aquifer cross-section.



(a) Plan view



(b) Section view

Figure 2-1 Schematic of the waste facility and leachate migration through the unsaturated and saturated zones: (a) plan view; (b) section view.

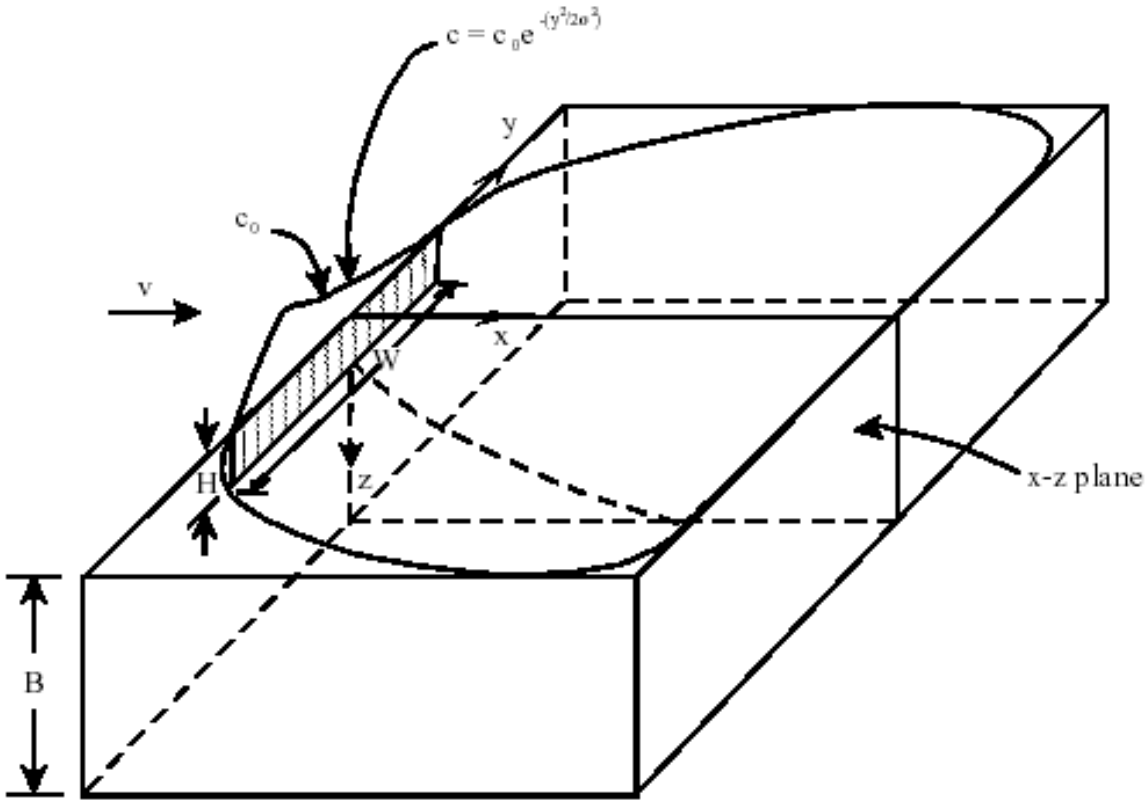


Figure 2-2 A schematic diagram of the Gaussian source boundary condition for the saturated zone transport module.

However, the concentration of the contaminant is not uniform across this area, but is instead assumed to be distributed in a Gaussian sense in the horizontal (in the *y* direction in Figure 2-2). It is from this location, with this boundary condition, that the contaminant is subsequently transported with flow in the saturated zone to the downstream receptor site. The lateral input recharge is uniform spatially and invariant with time, so that a steady, uniform flow field exists throughout the saturated zone.

As the contaminated leachate passes through the unsaturated zone, its solute (considered simply as a single contaminant) undergoes attenuation and redistribution through dispersion, biodegradation (according to linear first-order kinetics), hydrolysis, and adsorption, this last being prescribed as a function of the organic matter content of the unsaturated zone. Dispersion in the unsaturated zone, i.e., the dispersion coefficient, is computed as a function of the dispersivities and (vertical) seepage velocity. The former may be specified as a given, or derived from other attributes, such as the specified thickness of the unsaturated zone. The latter (seepage velocity) is calculated as a function of the infiltration rate, the porosity, and the water content (% saturation) of the unsaturated zone, which itself is a function of the specified saturated hydraulic conductivity. In other words, seepage velocity is derived from Darcy's law, as is customary. Hydrolysis in the unsaturated zone is computed as a function of the bulk density and the distribution coefficient of the contaminant between the dissolved and sorbed phases; it is assumed to occur in both phases with first-order kinetics that are a function of both temperature and pH. In the saturated zone, the solute undergoes the same processes of attenuation.

The source term for the infiltration of contaminated leachate into the unsaturated zone is invariant with time, both in respect to its volumetric rate of flow and its composition. It is also invariant in space, so that from the upper surface of the unsaturated zone down to the upper surface of the saturated zone it is similarly uniform in the horizontal plane, as the leachate moves vertically downward. The receptor site is assumed to be at a "worst-case" location, being directly downstream of the source, on the centre-line of the plume, and at the top of the saturated zone. The relationship between relative hydraulic conductivity and water saturation and the water moisture curve of the unsaturated zone is assumed to have the form first proposed by van Genuchten (1976).

In general, the structure of a model may be defined by the following equation for the dynamics of the state vector $\boldsymbol{x}$,

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}\{\boldsymbol{x},\boldsymbol{u},\alpha;t\} \qquad (2\text{-}1)$$

in which, in principle, $\boldsymbol{x}$ denotes the field of contaminant concentrations in the subsurface environment, $\boldsymbol{u}$ is a vector of inputs to the system (here the infiltration rate from the

storage facility or the rate of recharge to the aquifer), α is a vector of model parameters, such as a rate constant of contaminant degradation, and the dot notation in $\dot{x}$ denotes differentiation with respect to time *t*. Strictly speaking, however, since EPAMMM is based ultimately on a distributed-parameter, partial-differential equation set, such differentiation is of a partial (as opposed to a total) nature and $\boldsymbol{f}\{\}$ may contain partial derivatives of the state $\boldsymbol{x}$ with respect to distance. It is also noted that even this simple model requires the specification of more than 30 parameters (α) for its application to a Subtitle D facility.

More specifically, the inputs $\boldsymbol{u}$ are assumed for the present analysis to be invariant with time, so that computational results are concerned merely with the steady-state solution of equation (2-1) and, more specifically, with the value of the residual contaminant concentration at the receptor site, i.e., *y*, where *y* is given by

$$y = g\{\boldsymbol{x},\boldsymbol{u},\alpha\} \qquad (2\text{-}2)$$

and now $\boldsymbol{x}$, $\boldsymbol{u}$, and *y* are invariant with time. In the computational exercises that follow, equation (2-2) will be solved in the context of a Monte Carlo simulation, thus generating distributions for *y* as a function of the assumed uncertainty associated with α and $\boldsymbol{u}$. The possibility of any structural, i.e., conceptual, error in the forms of $\boldsymbol{f}$ in equation (2-1) and *g* in equation (2-2) will not be considered herein, although we acknowledge that errors of this kind are currently of some interest (Beck, 1987, 1994; Beck *et al*, 1993; Konikow and Bredehoeft, 1992). All models, by definition, will suffer from structural errors, in the sense that all models are approximations of the truth, and as such one can immediately recognise the impossibility of quantifying them. In the specific case of EPAMMM, a portion of these errors will be attributable to the errors in classifying a facility under Subtitle D when its features do not conform all that well to the idealised properties of this category, for example, when the facility does not have a rectangular source area, or the flow in the underlying saturated zone is not strictly in the horizontal plane alone.

A more complete description of EPAMMM can be found in Sharp-Hansen *et al* (1990) and Salhotra *et al* (1990).

3 Tests for the Assurance of Quality in a Model's Predictive Performance

In order to be effective as a tool for determining whether contamination arising from a storage facility will be significant and, in that event, what may be done to remedy such an unacceptable situation, EPAMMM must be able to demonstrate that uncertainty about the value of *y*, as a result of the substantial uncertainties in $\boldsymbol{u}$ and α, does not undermine the basis of decision-making. In the extreme, for example, the outcome that more or less

any value of *y* is equally probable under any given combination of soil, contaminant and hydrological regimes is hardly a secure basis on which to construct a decision. There are several issues to be addressed in assuring the quality of the model's predictive performance. We provide computational results for three such issues and indicate a fourth promising line of analysis.

Output Uncertainty as a Function of Different Site Characteristics

Let us suppose that the same contaminant is stored at several sites, with each site having a different underlying soil, aquifer and hydrological regime. From the perspective of making a decision relating to the performance of each such facility, interest would focus on the capacity to predict the residual contaminant concentrations *y* at the respective receptor sites in order to establish which facility is the most or least effective in containing the particular contaminant. Formally, it is necessary to determine whether the model is able to separate the respective distributions of *y*, let us say y_A and y_B, for two sites A and B respectively parameterised in soil and hydrological terms through θ_A and θ_B in equation (2-2) (much the same problem, albeit in a slightly different setting, is addressed in Beck and Halfon (1991)). By "separation", we mean that the probability of identical values of *y* being generated under the two (storage site) scenarios is less than some threshold, such as 0.01, 0.05, or 0.10 (as illustrated in Figure 3-1). Alternatively, it may be desirable to explore the scope of the model in discriminating (for a single site) between the residual concentrations of two (or more) contaminants with differing degradabilities, likewise parameterised through different ranges of values for θ.

From the practical perspective of making a decision -- for example, to rectify inadequate performance at site A or B -- such an analysis could be used to quantify the risk of taking the (wrong) action, say, at site A, when in reality site B is the more poorly performing storage system (see also Skiles *et al*, 1991; Goodrich and McCord, 1995).

Here our concern is primarily with what this kind of analysis may illuminate with respect to the power of the given model, in this case EPAMMM, to discriminate the predicted behaviour of one site from that of another. Given that there are strong prior beliefs that site or contaminant characteristics ought to generate distinctly different receptor site concentrations under reasonable model parameter uncertainty, the result that this is so (or not so) is revealing of the discriminating power, or relevance, of the model in performing the stated task. Indeed, some formal manipulation of the probability of coincident, i.e., indistinct, values of *y* might be used as a quantitative measure of this power.

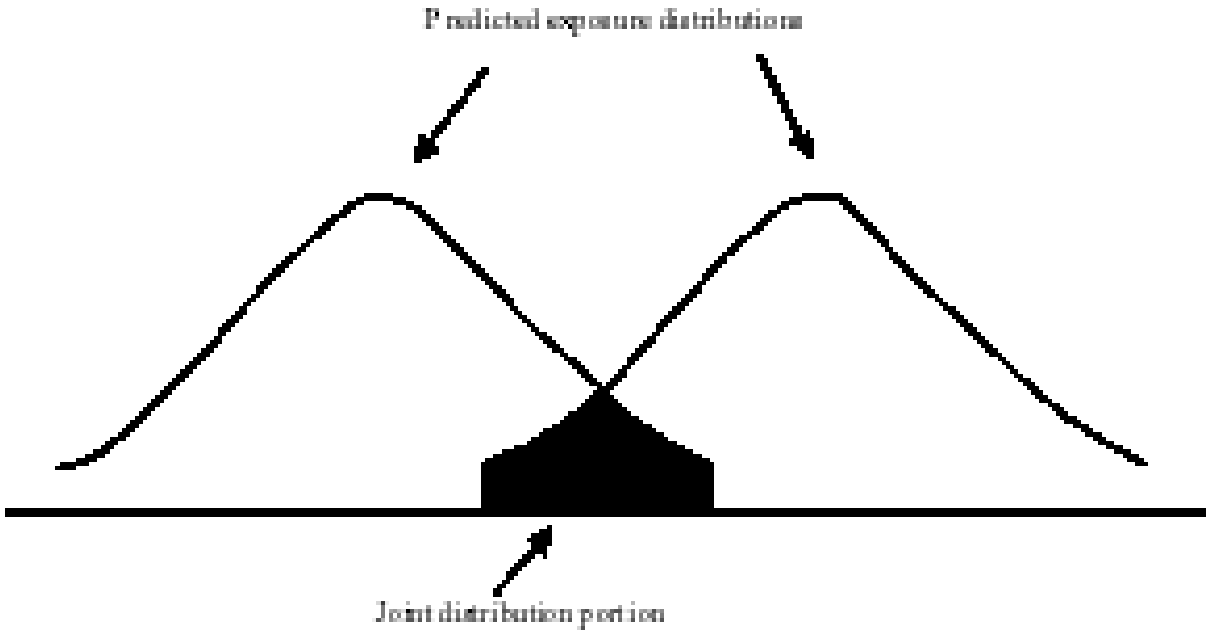


Figure 3-1 Diagram indicating the separation of two predicted exposure distributions.

Key and Redundant Parameters in Predicting a Percentile Concentration

The latter analysis can be viewed as follows. A screening-level assessment of the risk of adverse exposure at the receptor site is concerned with knowledge of the probability that a particular contaminant concentration, say ψ, will be exceeded. The choice of specific values for some of the parameters in the model, within the range of values they might assume, may be key to governing whether the resulting prediction of y falls above or below ψ. For other parameters, the choice of a specific value may be immaterial to such discrimination in terms of y being above or below ψ. The quality of the model in performing this screening task might, therefore, be related to the relative numbers of key and redundant model parameters, $\{\theta^K\}$ and $\{\theta^R\}$ respectively, that are so found (as outlined in the protocol of Beck *et al* (1995, 1997)).

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

In order to answer these questions, we shall use the algorithm of Hornberger, Spear, and Young (HSY), often referred to as regionalised sensitivity analysis (Young *et al*, 1978; Hornberger and Spear, 1980; and Spear and Hornberger, 1980), further brief details of which are given in Section 5. Indeed, the same form of analysis has been extended in a recent application to the model MMSOILS, a close relative of EPAMMM, in which the goal was to identify and explore how particular **clusters** (or aggregate assemblies) of parameters, as opposed to **individual** parameters, might be key or redundant in the above discriminating function (Spear *et al*, 1994). In the present study, we shall use merely the "basic" form of the regionalised sensitivity analysis, but note that any interpretations of its results will be subject to the limiting qualifications illuminated by Spear *et al* (1994).

Towards a Global Form of Sensitivity Analysis

A **local** analysis of a model's sensitivity seeks to compute the extent to which the output of the model may change, say ∂y_i, as a function of a change, $\partial \theta_i$, in the value of this

model parameter about its nominal (or "best") point estimate, θ_i^e. In such a test, all other values of the parameters and properties of the model are specified as those of their respective single, nominal, best estimates; the output deviation, δy_i, is, therefore, a change from the accompanying single, nominal, best estimate of the output y^e. For each parameter θ_i in the model, a measure of the sensitivity of the model's output -- as gauged by, say, $(\delta y_i / \delta \theta_i)$ -- can be computed, for **small** deviations from the **nominal** parameterisation of the model. This analysis, as is well known, is valid only within the local neighbourhood of the nominal parameterisation, and the domain of this validity will be narrower the more nonlinear the relationship between θ_i and y in the vicinity of θ_i^e. Nevertheless, our judgement about the validity of the model would differ significantly according to whether poorly known or well known parameters dominate behaviour in this neighbourhood. That is, we assert that most users would view a model with unease if its performance is critically dependent on the precise value of a parameter subject to great uncertainty and not easily identifiable from past observations of performance (Beck *et al*, 1995, 1997).

A **regional** analysis of sensitivity, as proposed by Hornberger and colleagues, can be thought of as an attempt to answer a similar question, but more generally over a larger domain of feasible values for the model's parameters and, significantly, with respect to a particular task expressed as some form of constraint on, or categorisation of, the output performance (y) of the model. Nonlinearities in the model are not, in principle, a problem, but the conclusions from the analysis are inevitably task-specific.

Herein, we introduce and illustrate yet a third form of sensitivity analysis. For each individual parameter θ_i in EPAMMM, we explore to what extent the **statistical properties** of the predicted distribution of the residual contaminant concentration, y, i.e., its mean μ_y and standard deviation σ_y, vary as a function of the point estimates assumed for θ_i across the range of its feasible values. All the remaining parameters of the model, other than the parameter i under examination, are treated as random variables within the framework of a Monte Carlo simulation. Unlike the regional analysis of Hornberger, Spear and Young, yet in line with a classical, local sensitivity analysis, no reference is made to a particular predictive task that the model must perform (as expressed in terms of y). Again, nonlinearity in the model structure is, in principle, not a problem. In particular, our extension permits quantification of that proportion of the uncertainty attached to the output (y) that derives from uncertainty in the knowledge of a given parameter θ_i. Previous approaches to this familiar problem of ranking the sources of uncertainty have been restricted to a ranking either of the local sensitivity coefficients, with all the restrictive assumptions thereby entailed, or of the coefficients estimated from an approximate relationship between the output (y) regressed upon the set of parameters θ_i (Beck, 1987; Janssen, 1994). To the extent that the regression relationship of the latter is a good or a bad approximation of the underlying (more complex) model from which it is derived, the resulting ranking will be more or less reliable. Our new approach to this problem is relatively free of any such restriction on the interpretation of its results.

Key and Redundant Controls in Achieving a Given Level of Site Performance

The problem of predicting a percentile concentration of the residual contaminant at the receptor site is akin to the problem of designing a system to achieve some "target" (or desired, or "optimal") category of performance, denoted *T*, for example. As discussed previously, we have considered the question of what choices of model parameter values will make such pre-specified performance achievable, not what values of the **controls** (***u***) associated with the system might be successful in this regard. This latter question is much more readily recognised, not as a matter of sensitivity analysis, but as the classical problem of control system design: of choosing a particular set, or sequence, of values for the input variables in order to bring about some desired output response. The distinctive feature of the present problem setting, however, is that our interest lies in achieving a broad band of target performance (*T*), namely a residual contaminant concentration $y \# \mathbf{y}$, as distinct from inducing the complement "not-target performance" ($\mathbf{T}$), i.e., $y > \mathbf{y}$, and to do so in the face of gross uncertainty about the site's conditions as reflected in the model's parameters $\mathbf{''}$.

In just the same way as the HSY algorithm enables discrimination between key and redundant parameters, so discrimination can be made between key $\{\mathbf{u}^K\}$ and redundant $\{\mathbf{u}^R\}$ controls in achieving/not achieving the target performance. One might further explore -- through the Monte Carlo framework of the HSY algorithm, as now seems more feasible than previously (Spear *et al*, 1994) -- what sets of values $\{\mathbf{u}(T)\}$ yield the target behaviour (*T*) and whether, through induced correlation with some of the model's parameters, i.e., $\{\mathbf{''}(T)\}$, successful control is critically dependent upon site characteristics such as hydrological regime, soil type, contaminant degradability, etc.

We do not, however, investigate this particular issue herein.

Closing Remarks

Subjecting the model to the described battery of tests is designed to establish its validity, trustworthiness, and relevance in performing a **prospective** task of prediction. These are tests designed to ensure -- to the maximum extent possible -- that when the model is used to fulfil this predictive task it will, so to speak, "achieve flight". The terms of these tests contain **nowhere** the classical notion of "matching history" (of evaluating the model's quality and performance with respect to observations of past behaviour; Konikow and Bredehoeft, 1992). We ask **not** whether flight has been achieved in the past under an observed set of flight conditions (although we are by no means disinterested in the answers to such questions, for they too are part of the basis of our judgement on the quality of the model). The purpose of our tests is to expose weaknesses and limitations in the model relative to whether flight is likely to be achievable under some other

(imagined, or required) set of flight conditions in the future. This does not ignore past experience, all of which may be thought of as being incorporated into the prior beliefs (expectations) surrounding the model, i.e., its prior validation status. Success (failure) of the model when subjected to our tests, while clearly not a matter of the matching of a particular, observed history, will accordingly increase (decrease) the validation status of the model. If we were working with a new model, many of our conditions may well reflect those observed in the past. But when working with projection of the model into novel situations -- as is of critical interest in predictive exposure assessment for the movement of novel substances into the environment (Beck *et al*, 1997) -- the power of our tests will derive from the richness we can bring to bear in imagining what conditions might occur in the future (and these may be derived in large part from manipulating (subjective) belief networks; Varis, 1995).

The purpose of our battery of tests is to expose, and possibly to rectify, these weaknesses before a specific prediction is made on which a decision, with costs of actions and risks of failure and subsequent damages, is made. Yet when so applied the model may still be shown in the event to have failed to predict specifically what may come to pass in the future. Indeed, this is almost inevitable: seen from the present, some, if not much, of the future is unknowable. As in the design of an aircraft, the previously described battery of tests is intended to minimise the damage of such failure and, perhaps, to maximise the ease of subsequently adapting the model in the light of what will be learned from the experience.

If the task of prediction is not richly specified, if only weak prior beliefs about expected site performance are held, and if there is great uncertainty attached to the model's constituent parameters, judgements about the legitimacy of using this model to fulfil the task will be equally bland. Similarly, if the cost of action, the risk of failure, and the magnitude of the ensuing damage, are minimal, then only minimal effort might need to be expended in carrying out such tests of the model. If the converses of all these conditions are true, very great effort might be justified: in assessing the model, in collecting more appropriate field data, and in changing the model to meet better the terms of the task specification. This is, of course, merely a juxtaposition of **qualitatively** quite different problem contexts, with no intention -- at this stage -- of seeking to quantify how much effort should be invested in re-designing the model as a function of (a) the richness of the task specification, (b) the uncertainty of our prior knowledge, or (c) the costs of action and failure. However, it **is** our purpose to project the discussion of model validity into a broader domain in which there is a concern for assuring the quality of a tool being designed against some (predictive) task specification.

4 Output Uncertainty and Discriminating Power

The question of central interest here is: does a reasonable range of model parameter

uncertainty render ineffective the power of the model to discriminate between the performance of containment facilities under quite different subsurface soil, hydrological, and contaminant-degradability regimes? For it is, we should note, our collective, strong, prior belief that such different regimes, as characterised by the model's parameters, should lead to significant differences in a facility's behaviour. The siting of a storage facility should, itself, be strongly conditioned upon finding a hydrogeological regime that is maximally resistant to the possibility of leakage and rupture.

In conducting a computational analysis in order to respond to this question, let us recall first that EPAMMM contains over 30 parameters (″) (for a Subtitle D facility). Our immediate need is to restrict the ranges of values these parameters may assume, in accordance with certain "distinct" categories of soil types and contaminant degradability properties.

Discrimination Among Different Soil Types

There is a substantial body of literature dealing with the textural and hydraulic characteristics of the various soil types (such as, for example, Jury (1982) or Carsel and Parrish (1988)). For present purposes, the textural classifications of the US Soil Conservation Service (SCS) will be used (Soil Conservation Service, 1972), i.e., clay, clay-loam, loam, loamy sand, silt, silt-loam, silty clay, silty clay-loam, sand, sandy clay, sandy clay-loam, and sandy loam. Only five of these categories, representing a supposedly wide array of characteristics, have been selected for this illustrative analysis: clay, loam, silt, sand, and sandy clay-loam.

Within each soil type the values that might be assigned to EPAMMM's parameters exhibit strong correlations and it is, therefore, important to account for this by restricting the ranges of values permitted for the sampling of the Monte Carlo simulation. In other words, the "signature" of a particular soil type is reflected in the choice of upper and lower bounds for the model's parameters. Thereafter, no further correlation structure will be imposed through the sampling procedure of the simulation; parameter values may be chosen independently from their respective ranges. Such restrictions are relevant for the following subset of 11 (out of the total of more than 30) parameters: saturated conductivity, porosity, residual water content, bulk density, and percent organic matter of the unsaturated zone; the two parameters of van Genuchten's expression; and particle diameter, bulk density, hydraulic conductivity, and organic carbon content of the saturated zone. More specifically, the numerical bounds placed on these parameters are drawn from the data bases of Sharp-Hansen *et al* (1990) and Carsel and Parrish (1988). For example, those used herein for the sandy clay-loam soil are given in Table 4-1.

The other parameters will be sampled from ranges that are independent of the soil type (but the same for each test), representing what might, therefore, be termed average

Table 4-1. Statistical Distributions of Inputs/Parameters
for Sensitivity Analysis

Parameter	Unit	Distribution	Mean	Standard	Minimum	Maximum
Aquifer Parameters						
Aquifer thickness	m	normal	12.0	3.5	5.0	19.0
Particle diameter	m	normal	0.0125	0.01	0.001	0.05
Bulk density		normal	1.4	0.25	0.94	1.76
Hydraulic conductivity	m/yr	lognormal	115	240	10	390
Hydraulic gradient		uniform			0.001	0.002
Temperature	°C	uniform			15	20
pH		uniform			6.0	8.5
Organic carbon content	%	uniform			0.001	0.01
Receptor from site	m	uniform			50.0	100.0
Source-Specific Parameters						
Recharge rate	m/yr	normal	0.015	0.01	0	0.03
Infiltration rate	m/yr	normal	0.25	0.07	0.13	0.38
Waste disposal area	m^2	fixed	100			
Initial concentration	mg/l	fixed	1.0			
Chemical Parameters						
Acid-catalysis rate	1/mole-yr	fixed	0.0			
Neutral-catalysis rate	1/mole-yr	fixed	0.0			
Base-catalysis rate	1/mole-yr	fixed	0.0			
Reference temperature	°C	fixed	25			
Bio-degradation rate	1/yr	fixed	0.0			
Normal distribution coefficient	cc/g	uniform			71	178
UZ Parameters						
Saturated conductivity	cm/h	lognormal	1.31	2.74	0.02	4.42
Porosity		normal	0.40	0.05	0.30	0.50
Residual water content		normal	0.1	0.006	0.088	0.112
Depth of UZ	m	uniform			5	10.0
Alpha coefficient "		normal	0.059	0.038	0.005	0.124
Beta coefficient ß		normal	1.48	0.13	1.09	1.87
Air entry pressure	m	fixed	0.0			
Percent organic matter	%	normal	0.26	0.25	0.01	1.5
Biodegradation in UZ	1/yr	fixed	0.0			
Bulk density	g/cc	normal	1.40	0.2	1.0	1.6

ranges for all but two of these parameters. Recharge of the saturated zone at its upstream boundary is given a low range of permissible values, reflecting a "worst-case" scenario in which there is a below-average flow of water for dilution of any contaminant reaching the saturated zone. The presumption here is that the site may well have been chosen for the storage of materials precisely because of the limited scope for widespread propagation of the contaminants in the event of leakage. Similarly, under the assumption that the majority of facilities will already have some form of lining installed, the rate of infiltration of the leachate into the unsaturated zone is assigned a relatively low range of permissible values. We acknowledge that assuming the recharge and infiltration rates to be independent of soil type is somewhat contentious, the role of the facility lining notwithstanding. These water fluxes, however, will also be a function of variations in site-specific precipitation, evapotranspiration, and soil moisture content, such that the local soil types may appear to have little bearing on their magnitudes.

One further salient feature of the test formulation is the assumption of a conservative, non-degradable contaminant in the leachate, such as benzene, whose rates of hydrolysis and bio-degradation have been found to be very small (Schnoor ***et al***, 1987). Again, we recognise that this may appear to some to be a strong assumption. The issue is rather one of the period of time over which the absence of significant degradation can reasonably be assumed to occur (for given sufficient time virtually all chemical species might be deemed to be "degradable"). For our present purposes, benzene is used as the archetype of a very slowly degradable, volatile contaminant.

In this test, then, our prior expectation -- conditioned on all that has gone before -- is that (radically) different soil types should be an important feature in discriminating between the extent to which the contaminant is leached from the disposal facility. Furthermore, the implication is that, if EPAMMM is to be credible in a screening-level analysis, such discrimination should be apparent against a background of a reasonable level of uncertainty in site characterisation.

Results

The computed distributions for *y* at the receptor site, as a function of the five soil types, are shown in Figure 4-1. It appears that only the distribution for the sand can be distinguished from the other four soil types, which rather confounds our prior beliefs. Moreover, the

predicted exposure concentration for the sand is significantly lower than the other responses, which again is at variance with prior expectations. However, according to the formulation of the model, the low adsorption capacity and high hydraulic conductivity of the sand are such that a large dispersion coefficient is computed and used in the equations for contaminant transport. The expression for the dispersion coefficient is nonlinearly related to the soil properties, and notably so towards the sand end of the spectrum. The consequence is a relatively low exposure concentration, even

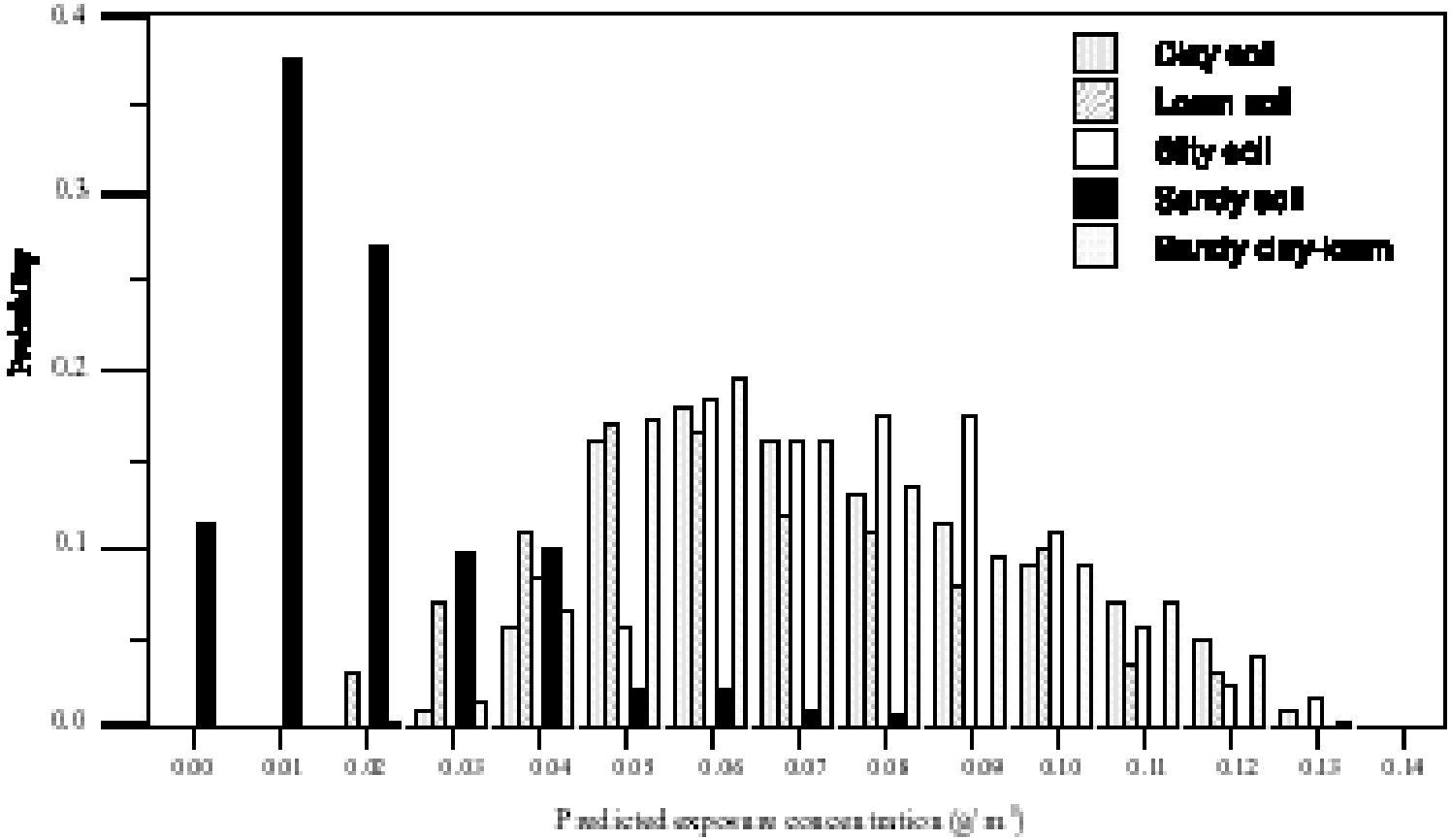


Figure4-1 Distributions of exposure concentrations for different soil types.

at the relatively very closely located receptor site, just some 50-100m distant from the source. In the case of the sand regime, the contaminant plume will be dilute, but extensive, implying that at many other locations the associated exposure concentration will exceed those in the less dilute, but less extensive plumes in the other soil types. Figure 4-2 shows that the lack of power in discriminating among the responses of the different soils is not altered when the receptor is further away from the source, at a (fixed, i.e., certain) distance of 1000m.

This low discriminating power of the model, which confounds our prior beliefs, may have several origins. First, in view of the relative simplicity of the features incorporated in EPAMMM, it may be that the true richness, diversity, and distinctive features of the soil types cannot be properly reflected in the model's limited set of parameters. Second, the soils are classified essentially in terms of characteristics strictly relevant to soil science (such as particle size) and not to the description of a hydrological regime. Consequently, the hydrological features associated with the soil types may not be all that strongly distinguished, and it could be argued that our results are unsurprising. However, we should not overlook the fact that a systematic means of mapping soil properties into hydrological properties has long been the subject of intensive study (for example, Carsel and Parrish, 1988). For some, therefore, the prior belief that a storage facility located in one type of soil will perform significantly better than another located in a significantly different soil type could be quite strong indeed. Third, what has been considered a reasonable level of uncertainty attached to the model's parameters may simply not be sufficiently small to permit the expected separation of the distributions of exposure concentrations. Fourth, the model may, in fact, be a reasonable representation, but the test under steady-state conditions of saturated flow and input leachate rate may not be a sufficiently exciting (or discriminating) test. In other words, certain transient modes of behaviour that are especially sensitive to some of the model's parameters, and which would lead to distinctly different responses among the soil types, are not being sufficiently exercised.

If, therefore, we have no reason to overturn our prior beliefs, if the steady-state condition is the most relevant style of test, and if no better (less uncertain) information can be made available for implementing the model, which may well be the case at the level of a screening analysis, EPAMMM may **not** be appropriate for this predictive task. We argue, however, that this would be a premature conclusion and that the model should undergo more extensive testing, as follows in Sections 5 and 6.

Discrimination Among Different Contaminant Types

In the preceding assessment, the properties of benzene were used to illustrate a category of conservative contaminant types. In a parallel effort, the predicted off-site movement of four other chemicals, DDT (as a representative pesticide), Aroclor 1248 (as

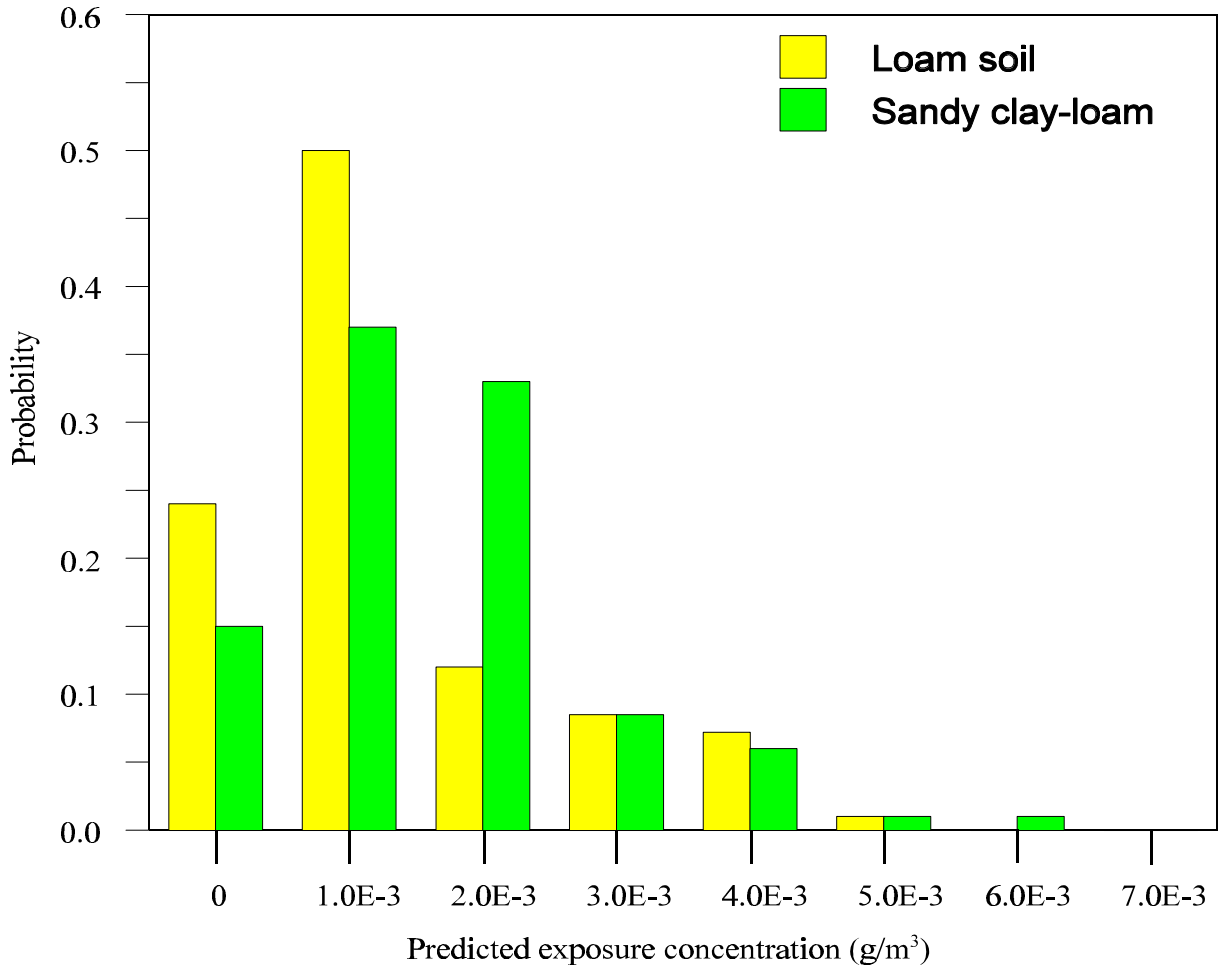


Figure 4-2 Exposure concentration distributions when receptor from site is fixed at 1000m away.

a representative PCB), chloroform (as a halogenated aliphatic hydrocarbon) and 2,4-dimethylphenol (as a monocyclic aromatic compound), was compared to that of benzene. The chemical properties of these four substances are given in Table 4-2. In terms of biodegradability they span a diversity of values, with benzene being effectively zero, DDT,

Aroclor 1248, and chloroform being low, while that of 2,4-dimethylphenol can be regarded as relatively high. The uncertainty attached to the rate of decay also varies significantly among the five chemicals. The range of values permitted for 2,4-dimethylphenol is high; that for DDT is medium; and those for Aroclor 1248 and chloroform are low. In fact, the range of degradation rates for DDT covers the total range occupied by Arochlor 1248 and chloroform. Similar ranges of values are assigned to the rates of hydrolysis and adsorption capacities of the five contaminants. In short, the substances chosen for analysis should reflect a wide range of chemical behaviours.

Results

The predicted distributions of residual contaminant concentrations at the receptor site (for an identical input leachate concentration) are shown in Figure 4-3. In this test, there is a more distinctive separation among the responses, except for the cases of DDT and Arochlor 1248, which are so low as to be aggregated into a single category of the plot in Figure 4-3(a).

There are several salient points to note in respect of these results. First, we might conclude that the model **is**, in this instance, appropriate for performing the given task, because of its greater power -- relative to the test with the different soil types -- in generating responses that are (and are expected to be) distinct. However, *in situ* field estimates of the parameters associated with biodegradation, hydrolysis, and adsorption are not well known and may deviate substantially from those determined under laboratory conditions (Blackburn, 1989). If, therefore, the performance of the model is more sensitive to a group of parameters believed a priori **not** to be known relatively well, this is disquieting, as we have observed previously. One is likely to feel less comfortable about the validity of a model in performing a given task, when that performance is dominated by features of the model in which there is less confidence and for which appropriate parameter estimates are hard to obtain (Beck *et al*, 1995, 1997). The model may still be a useful model in principle; our point, however, is that it may not be well suited to the given, specific predictive task. A better conclusion is that further experimental effort would be most profitably allocated to narrowing the ranges of values for the parameters of contaminant removal mechanisms. The trustworthiness of using the model for predictive exposure assessments at the screening level would, thereby, be relatively more significantly enhanced.

In addition, we note that there is substantial separation, of the order of 10^5 to 10^{12} times magnification, between the predicted exposure concentrations of conservative and non-

Table 4-2. The Uncertainty of Chemical Parameters for Several Typical Chemicals

Chemicals	**Biodegrad-ation rate (1/yr)**	**Acid-catalysis rate (1/mole-yr)**	**Neutral-catalysis rate (1/mole-yr)**	**Base-catalysis rate (1/mole-yr)**	**Normal distribution coefficient (cc/g)**
Benzene	0.0	0.0	0.0	0.0	71-178
Pesticide (DDT)	0.0-0.10	0.06	0.0	31186-311856	47863-5011872
PCB (Aroclor 1248)	0.0-0.007	0.0	0.0	0.0	346737-794328
Halogenated Aliphatic Hydrocarbon (Chloroform)	0.09-0.10	4.3	0.5	1892	49-58
Monocyclic Aromatic (2,4-Dimethylphenol)	0.24-0.66	0.0	0.0	0.0	123-195

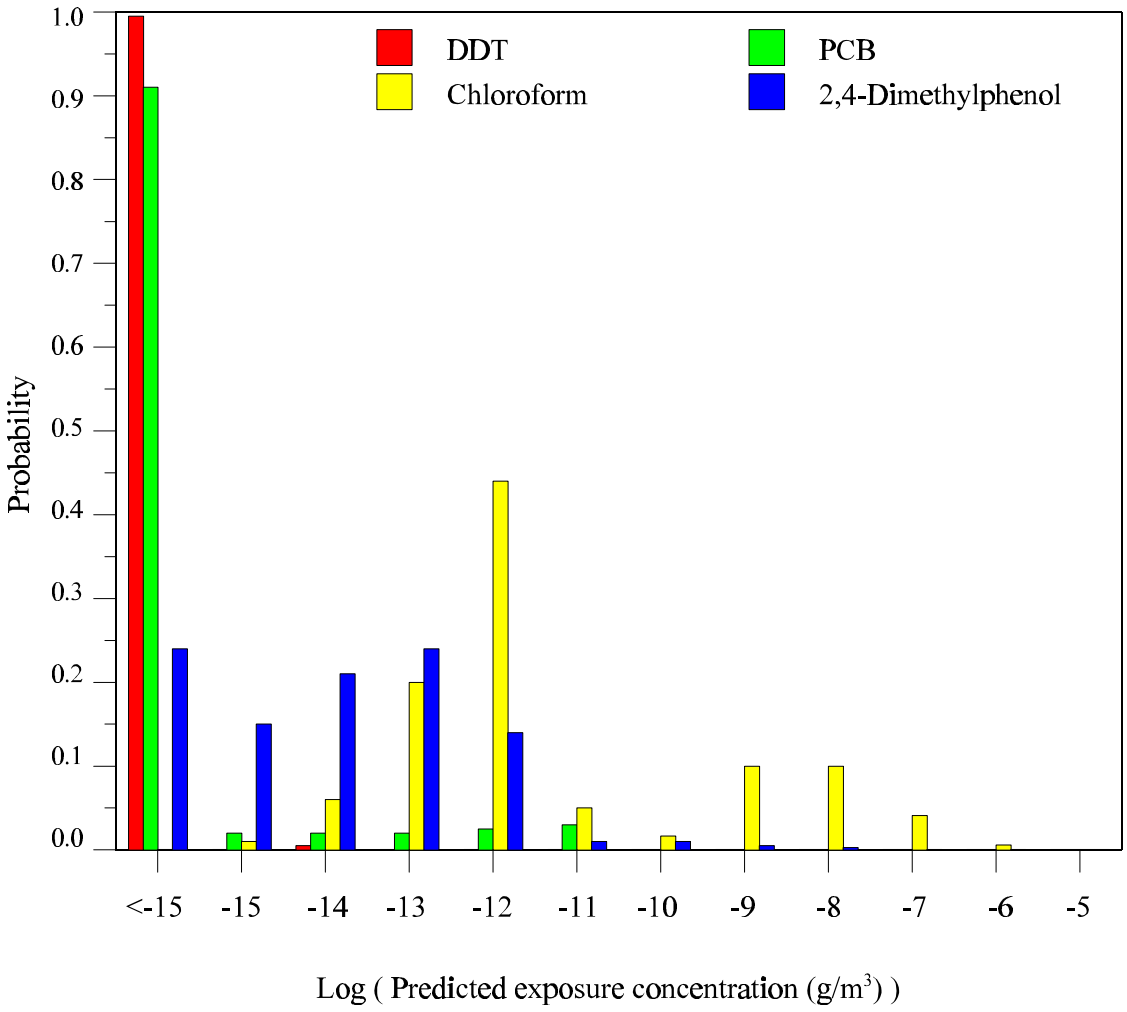


Figure 4-3(a) Exposure concentration distributions for different chemicals.

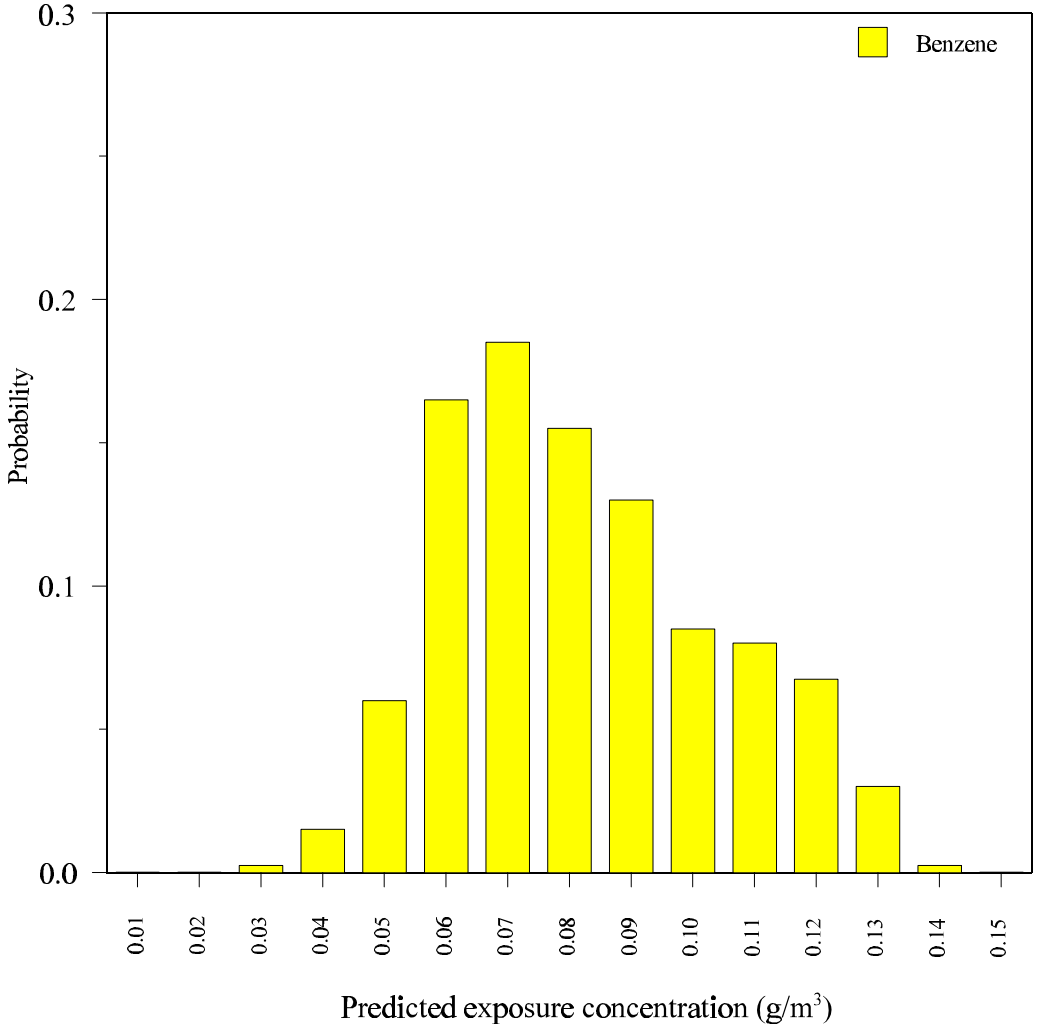


Figure 4-3 (b) Exposure concentration distribution for a conservative chemical species.

conservative contaminants. If, therefore, an unacceptable prediction of exposure concentration is obtained with respect to the (safe) assumption of a conservative substance, the implication is that the acquisition of reliable information about the degradability of the substance is likely to be highly cost-effective, especially when set against the otherwise potentially very large costs of remedial action.

Another observation is that the predicted exposure concentrations are rather uncertain, to say the least. Except for the special case of DDT (whose concentration is categorised in an undifferentiated manner, as below $10^{-15}g/m^3$), predicted concentrations range over 6-9 orders of magnitude. Once again, the uncertainty attached to the parameters of the removal mechanisms may be unreasonably high. On the other hand, the range of values permitted for the biodegradation of chloroform is small, so that the uncertainty in the predicted residual concentration at the receptor site may not stem directly from this particular parameter. Rather, it may be dominated by other factors, such as a highly uncertain "residence time" for the contaminant (between source and receptor), over which this relatively well known rate of degradation is acting.

We note that each of the mechanisms of contaminant removal from the aqueous phase, i.e., biodegradation, hydrolysis, and adsorption, is represented in the model in a different functional form. The aggregate rate of removal of each contaminant is dominated by a different mechanism: hydroloysis in the case of chloroform; biodegradation for 2,4-dimethylphenol; and adsorption for Aroclor 1248. Not only, then, is such removal in the **aggregate** important, *vis à vis* the background uncertainty in the facility's soil and hydrological characterisation, but the discriminating power of EPAMMM is dependent upon the inclusion of such "richness" in the **constituent** mechanisms of overall removal. In other words, this richness should not be subsumed under the umbrella of a single attenuating factor. For while it may not be vital in assessing which sites are most in need of remedial action, it may be essential in determining what particular form this action should then assume, given that quite different costs may attach to different engineering controls.

To summarise, the model possesses some discriminating power, particularly in regard to the attenuation of contaminant migration through mechanisms other than advection and dispersion in the flow of groundwater. However, this conclusion must be qualified by noting that the variety of hydrological regimes simulated by the model may be restricted as a result of the limited scope for properly parameterising the diversity of soil properties (although many parameters in the model are devoted to an adequate characterisation of soil properties).

Yet this conclusion is also strongly conditioned upon a set of prior beliefs about how such facilities for waste disposal ought to behave. It is also strongly conditioned upon what is believed to be a reasonable measure of uncertainty attached to EPAMMM's parameters. Furthermore, this is a conclusion drawn largely independently of any

specific, predictive task that the model must perform. The purpose of the analysis that follows in Section 5 is to examine the discriminating power of the model when its performance is made task-specific.

5 Key and Redundant Model Parameters

In the preceding analysis, Monte Carlo simulation was implemented through the repeated sampling of values for the model's parameters between upper (θ_i^u) and lower (θ_i^l) bounds for each parameter *i*. What was of interest was the resulting distribution of *y*, the residual contaminant concentration at the receptor site, irrespective of any constraint attached to the output. Decisions, however, are based on some desired level of performance from the system. In the present case, the desired performance is simply that the residual contaminant concentration should not exceed a pre-specified value, $\bar{y}$, with probability *p*. What we want to know is: for which (key) constituent parameters of the model would we most want to have good knowledge readily available? What is it, in other words, that is most critical in the design of the model with respect to successful achievement of this particular task?

We can, in part, answer such questions through the algorithm of Hornberger, Spear and Young (Young *et al*, 1978; Hornberger and Spear, 1980; Spear and Hornberger, 1980), in which it is necessary to discriminate between two classes of behaviour from the model, that which would be "acceptable", i.e.,

$$0 \le y \le \bar{y} \tag{5-1}$$

and its complement, "unacceptable" behaviour, being classified as $y > \bar{y}$. We might also speak of the "acceptable" as the "target" behaviour, as earlier. For each random realisation (*j*) of a candidate parameterisation of the model, i.e., θ^j, y^j is obtained and, according to equation (5-1), is associated with either giving, or not giving, the target behaviour. For a sufficiently large sample of realisations, two sets of candidate parameterisations of the model can be distinguished: those *m* samples $\{\theta(T)\}$ that give the target behaviour and those *n* samples that do not, i.e., $\{\theta(\bar{T})\}$. For each constituent parameter, θ_i, the maximum separation distance, d_{max}, between the respective cumulative distributions of $\{\theta_i(T)\}$ and $\{\theta_i(\bar{T})\}$ may be determined and the Kolmogorov-Smirnov statistic, $d_{m,n}$, used to discriminate between significant and insignificant separations for a chosen level of confidence (Kendall and Stuart, 1961; Spear and Hornberger, 1980). Relatively large separation implies that assigning a particular value to the given parameter is key to discriminating whether the model does, or does not, generate the target behaviour. Relatively small separation of the two distributions implies that evaluation of the associated parameter is redundant in discriminating the performances

of the model. For the latter, it matters not, in effect, what value is given to the parameter; the giving or not giving of the target performance is more or less equally probable whatever value of the parameter is assigned. We have already referred to the sets of key and redundant model parameters as $\{\theta^K\}$ and $\{\theta^R\}$, respectively.

Put simply, our approach combines a Monte Carlo simulation with an analysis of the (posterior) parameter distributions resulting from the classification of equation (5-1). We employed a computationally more efficient version of the approach due to Chen (1993) in this study. Moreover, since our interest lies in how the identification of key and redundant parameters varies as a function of the exposure concentration $y(p)$ not to be exceeded with probability p, the total sample of realisations $\{\theta^j\}$ associated with $\{y^j\}$ must be stored for subsequent classification according to equation (5-1) as y varies as a function of the chosen percentile values of the overall distribution of y. In this way it should be possible to assess whether the same constituent parameters of the model are key to the task of predicting high-end, mean, or low-end exposure concentrations, for example. It is, of course, the high-end exposure concentrations that are of particular interest to decision-making.

For illustration, a conservative (i.e., non-biodegradable) contaminant (again, benzene) associated with a facility located in a sandy clay-loam soil, where the 95th-percentile exposure concentration is deemed to be the upper bound on a tolerable high-end exposure, is assumed for the task description. The ranges of EPAMMM parameter values to be sampled for the Monte Carlo simulation were those of Table 4-1. In order to obtain a better understanding of the mechanisms that govern the model's performance across the spectrum of exposures, classification into key and redundant parameters was also undertaken for the 10th-, 30th-, 50th-, 70th-, and 90th-percentile concentrations of the contaminant at the receptor site.

Tables 5-1(a) through 5-1(f) summarise the resulting rankings of the parameters, according to their significance in discriminating target from non-target performance. Three categories of parameters are identified: "key", "important", and "redundant".

Results: High-end Exposure Concentration

Five parameters were found to be key for the 95th-percentile (Table 5-1(a)): the coefficients of dispersion in all three directions in the saturated zone; the distance of the receptor from the site; and the rate of leachate infiltration. For a conservative contaminant this result is unsurprising. However, it was somewhat unexpected that the hydraulic gradient and rate of recharge of the saturated zone were not found to be key parameters (a point to which our discussion will return later).

Exploring the details of these results, it is pertinent to observe first that the coefficients

Table 5-1 (a). Classification of Model Sensitivity at 95 Percentile

Key Parameters (significance level=0.001, $d_{m,n}$=0.1864):	d_{max}=
Transverse dispersivity in aquifer	0.8905
Receptor from site (distance)	0.8874
Longitudinal dispersivity in aquifer	0.8874
Vertical dispersivity in aquifer	0.8695
Infiltration rate	0.2253
Important Parameters (significance level=0.1, $d_{m,n}$=0.1170):	
Recharge rate	0.1747
Seepage velocity	0.1632
Bulk density in aquifer	0.1590
Source penetration depth in aquifer	0.1484
Hydraulic conductivity in aquifer	0.1347
Redundant Parameters (significance level>0.1):	
Aquifer porosity	0.1168
Organic carbon content in aquifer	0.1032
Bulk density in UZ	0.0968
Temperature in aquifer	0.0947
Longitudinal dispersivity in UZ	0.0937
Retardation coefficient in aquifer	0.0905
Percent organic matter in UZ	0.0895
Normal distribution coefficient	0.0768
Distributed coefficient	0.0758
Particle diameter in aquifer	0.0758
pH in aquifer	0.0737
Depth of UZ	0.0737
Aquifer thickness	0.0621
Beta	0.0558
Porosity in UZ	0.0547
Hydraulic gradient	0.0505
Saturated conductivity in UZ	0.0379
Residual water content	0.0368
Alpha	0.0190

Table 5-1 (b). Classification of Model Sensitivity at 90 Percentile

Key Parameters (significance level=0.001, $d_{m,n}$=0.1354):	d_{max}=
Transverse dispersivity in aquifer	0.76222
Receptor from site (distance)	0.75889
Longitudinal dispersivity in aquifer	0.75889
Vertical dispersivity in aquifer	0.75111
Seepage velocity	0.16111
Important Parameters (significance level=0.1, $d_{m,n}$=0.0850):	
Infiltration rate	0.12667
Recharge rate	0.11778
Hydraulic conductivity in aquifer	0.10889
Source penetration depth in aquifer	0.10444
Beta	0.08778
Redundant Parameters (significance level>0.1):	
Porosity in UZ	0.08000
Saturated conductivity in UZ	0.07111
Bulk density in UZ	0.06889
Organic carbon content in aquifer	0.06000
Retardation coefficient in aquifer	0.05667
Aquifer porosity	0.05667
Temperature in aquifer	0.05556
Normal distribution coefficient	0.05222
Hydraulic gradient	0.05222
Percent organic matter in UZ	0.05000
Longitudinal dispersivity in UZ	0.04667
Residual water content	0.04667
Bulk density in aquifer	0.04556
pH in aquifer	0.04445
Distributed coefficient	0.04222
Particle diameter in aquifer	0.03556
Aquifer thickness	0.03222
Depth of UZ	0.02667
Alpha	0.00222

Table 5-1 (c). Classification of Model Sensitivity at 70 Percentile

Key Parameters (significance level=0.001, $d_{m,n}$=0.08865):	d_{max}=
Transverse dispersivity in aquifer	0.35857
Receptor from site (distance)	0.34857
Longitudinal dispersivity in aquifer	0.34857
Vertical dispersivity in aquifer	0.34857
Infiltration rate	0.10095
Important Parameters (significance level=0.1, $d_{m,n}$=0.05566):	
Recharge rate	0.08095
Particle diameter in aquifer	0.07381
Hydraulic conductivity in aquifer	0.06714
Aquifer porosity	0.06429
Seepage velocity	0.05952
Redundant Parameters (significance level>0.1):	
Porosity in UZ	0.05524
Source penetration depth in aquifer	0.04857
Hydraulic gradient	0.04476
Saturated conductivity in UZ	0.03429
Bulk density in UZ	0.03429
Organic carbon content in aquifer	0.03381
Beta	0.03191
Distributed coefficient	0.03048
pH in aquifer	0.02905
Aquifer thickness	0.02857
Normal distribution coefficient	0.02810
Temperature in aquifer	0.02762
Longitudinal dispersivity in UZ	0.02714
Residual water content	0.02667
Retardation coefficient in aquifer	0.02429
Percent organic matter in UZ	0.02143
Bulk density in aquifer	0.01476
Depth of UZ	0.01238
Alpha	0.00667

Table 5-1 (d). Classification of Model Sensitivity at 50 Percentile

Key Parameters (significance level=0.001, $d_{m,n}$=0.08125):	d_{max}=
Transverse dispersivity in aquifer	0.22600
Receptor from site (distance)	0.22000
Longitudinal dispersivity in aquifer	0.22000
Vertical dispersivity in aquifer	0.22000
Recharge rate	0.08800
Important Parameters (significance level=0.1, $d_{m,n}$=0.05101):	
Infiltration rate	0.08000
Source penetration depth in aquifer	0.06400
Hydraulic conductivity in aquifer	0.05800
Redundant Parameters (significance level>0.1):	
Seepage velocity	0.05000
Depth of UZ	0.04800
Normal distribution coefficient	0.04600
Aquifer porosity	0.04400
Distributed coefficient	0.04000
Particle diameter in aquifer	0.03800
Hydraulic gradient	0.03600
Bulk density in UZ	0.03600
Longitudinal dispersivity in UZ	0.03600
Saturated conductivity in UZ	0.03400
Porosity in UZ	0.03200
Temperature in aquifer	0.03000
Residual water content	0.03000
Percent organic matter in UZ	0.03000
Organic carbon content in aquifer	0.02600
Bulk density in aquifer	0.02200
Beta	0.01800
Aquifer thickness	0.01800
Retardation coefficient in aquifer	0.01800
pH in aquifer	0.01400
Alpha	0.00200

Table 5-1 (e). Classification of Model Sensitivity at 30 Percentile

Key Parameters (significance level=0.001, $d_{m,n}$=0.08865):	d_{max}=
Transverse dispersivity in aquifer	0.30238
Receptor from site (distance)	0.30238
Longitudinal dispersivity in aquifer	0.30238
Vertical dispersivity in aquifer	0.30238
Important Parameters (significance level=0.1, $d_{m,n}$=0.05566):	
Recharge rate	0.08095
Source penetration depth in aquifer	0.07667
Seepage velocity	0.07000
Infiltration rate	0.06857
Depth of UZ	0.06238
Redundant Parameters (significance level>0.1):	
Hydraulic conductivity in aquifer	0.05191
Aquifer porosity	0.05143
Organic carbon content in aquifer	0.04714
Bulk density in aquifer	0.04619
Distributed coefficient	0.04191
Bulk density in UZ	0.04048
Normal distribution coefficient	0.03857
Porosity in UZ	0.03571
pH in aquifer	0.03571
Temperature in aquifer	0.03476
Aquifer thickness	0.03286
Hydraulic gradient	0.03143
Retardation coefficient in aquifer	0.03143
Particle diameter in aquifer	0.03095
Beta	0.03048
Residual water content	0.02905
Longitudinal dispersivity in UZ	0.02762
Percent organic matter in UZ	0.02619
Saturated conductivity in UZ	0.02429
Alpha	0.00191

Table 5-1 (f). Classification of Model Sensitivity at 10 Percentile

Key Parameters (significance level=0.001, $d_{m,n}$=0.1354):	d_{max}=
Transverse dispersivity in aquifer	0.37222
Receptor from site (distance)	0.37222
Longitudinal dispersivity in aquifer	0.37222
Vertical dispersivity in aquifer	0.37222
Source penetration depth in aquifer	0.19667
Infiltration rate	0.18667
Important Parameters (significance level=0.1, $d_{m,n}$=0.08502):	
Percent organic matter in UZ	0.11111
Hydraulic conductivity in aquifer	0.10333
Seepage velocity	0.10111
Redundant Parameters (significance level>0.1):	
Hydraulic gradient	0.08111
Porosity in UZ	0.07556
Normal distribution coefficient	0.07444
Aquifer porosity	0.07333
Temperature in aquifer	0.07111
Aquifer thickness	0.06556
Bulk density in UZ	0.06111
Retardation coefficient in aquifer	0.05667
pH in aquifer	0.05556
Bulk density in aquifer	0.05445
Distributed coefficient	0.05111
Recharge rate	0.04889
Organic carbon content in aquifer	0.04889
Saturated conductivity in UZ	0.04889
Depth of UZ	0.04444
Particle diameter in aquifer	0.04222
Residual water content	0.03889
Longitudinal dispersivity in UZ	0.03556
Beta	0.03222
Alpha	0.00222

of dispersion in the saturated zone are dependent upon the distance to the receptor site, so that the importance of the latter may be merely an artifact of this relationship. When the dispersivities are assumed known with certainty, i.e., fixed, the corresponding results of

Table 5-2 are obtained (those relating to the parameters of the unsaturated zone have been omitted from the Table, since they play no apparently vital role in the model for this particular test). Identification of the distance to the receptor site as the sole, key parameter confirms its crucial importance in predicting high-end exposure concentrations. It also indicates the relatively high trustworthiness of the model in predicting high-end exposure concentrations, since the distance to the receptor site, above all the other model parameters, ought to be a relatively well known quantity.

Taking stock of the results of Table 5-1(a) it is apparent that: (i) all the parameters identified as key and important are associated with either the properties of the saturated zone (the aquifer) or the source of the leachate; (ii) all the parameters associated with the unsaturated zone are found to be redundant; and (iii) parameters associated with the adsorption of the contaminant are likewise redundant. The second observation is notably inconsistent with our expectations, quite possibly as a consequence of the steady-state form of the test conditions. If this is so, then the subsequent analysis of performance over the range of percentile exposure concentrations should likewise result in the redundancy of this group of parameters (which is not necessarily the same as confirming the test conditions to be the cause of this counter-intuitive result). The third concluding observation suggests that, as opposed to being entirely redundant, the effects of adsorption of the contaminant are, in fact, dominated by the effects of dispersion. In this respect, it is significant that when the coefficients of dispersion are removed from the analysis, the organic carbon content of the aquifer -- upon which the capacity for contaminant adsorption depends -- is identified as an important parameter (compare the results of Table 5-2 with those of Table 5-1(a)).

It is possible to detect a more subtle feature in the comparative results of Tables 5-1(a) and 5-2. Exclusion of the effects of dispersion from the analysis (as in Table 5-2), which leaves adsorption as the only mechanism of contaminant attenuation under investigation (other than dilution), gives rise to quite different rankings of the parameters in the important and redundant classes. It would appear that the form of the contaminant attenuation mechanism, i.e., other than simple dilution, may play a crucial role in the model's achievement of its task (a conclusion already foreshadowed in the preceding analysis of Section 4). When Aroclor 1248, a slowly biodegradable contaminant with a substantially higher capacity for adsorption, was substituted for benzene in the analysis, the results of Table 5-3 were obtained. Again, the rankings of the parameters (relative to both Tables 5-1(a) and 5-2) have changed materially, including those now identified within the key category. Here, the parameters associated with the unsaturated zone are seen to play a key role in achieving the predictive task, since it is predominantly in this zone that biodegradation and adsorption occur. It also followed that evaluation of the parameters of the saturated zone ceased to be of significance, since so little of the

Table 5-2. Classification of Model Sensitivity
at 95 Percentile with the Dispersivity in Aquifer Fixed

Key Parameters (significance level=0.001, $d_{m,n}$=0.2636):	d_{max}=
Receptor from site (distance)	0.7874
Important Parameters (significance level=0.1, $d_{m,n}$=0.1655):	
Hydraulic gradient	0.2168
Aquifer porosity	0.1895
Seepage velocity	0.1768
Organic carbon content in aquifer	0.1684
Redundant Parameters (significance level>0.1):	
Aquifer thickness	0.1516
Source penetration depth in aquifer	0.1474
Hydraulic conductivity in aquifer	0.1368
Temperature in aquifer	0.1200
Recharge rate	0.1116
Normal distribution coefficient	0.1095
Retardation coefficient in aquifer	0.1011
Bulk density in aquifer	0.0968
pH in aquifer	0.0926
Infiltration rate	0.0695
Distributed coefficient	0.0842
Particle diameter in aquifer	0.0632

Table 5-3. Classification of Model Sensitivity
at 95 Percentile with Consideration of Biodegradation

Key Parameters (significance level=0.001, $d_{m,n}$=0.1864):	d_{max}=
Biodegradation rate in UZ	0.8094
Percent organic matter in UZ	0.1958
Important Parameters (significance level=0.1, $d_{m,n}$=0.1170):	
Residual water content	0.1432
Hydraulic gradient	0.1200
Source penetration depth in aquifer	0.1179
Redundant Parameters (significance level>0.1):	
Recharge rate	0.1105
Retardation coefficient in aquifer	0.1084
Organic carbon content in aquifer	0.1021
Infiltration rate	0.0937
Distributed coefficient	0.0873
Normal distribution coefficient	0.0853
Seepage velocity	0.0779
Bulk density in UZ	0.0768
Depth of UZ	0.0758
Hydraulic conductivity in aquifer	0.0747
Longitudinal dispersivity in aquifer	0.0737
Particle diameter in aquifer	0.0726
Temperature in aquifer	0.0726
Aquifer porosity	0.0716
Vertical dispersivity in aquifer	0.0695
Receptor from site (distance)	0.0684
Aquifer thickness	0.0684
Bulk density in aquifer	0.0674
Transverse dispersivity in aquifer	0.0653
Beta	0.0590
Biodegradation rate in aquifer	0.0579
Saturated conductivity in UZ	0.0558
Alpha	0.0558
pH in aquifer	0.0495
Porosity in UZ	0.0484

contaminant ever penetrated to that sector of the subsurface environment. Were the leachate infiltration rate to be very high, however, the same conclusion may not be tenable.

For the task of predicting high-end exposure concentrations in a screening-level analysis, knowledge of how to parameterise the contaminant attenuation mechanisms -- other than dilution -- is the most significant item of quantitative information for site characterisation (and thus a priority for the allocation of funds to any further fact-finding).

Results: Performance Over the Entire Distribution

In general, the relative degrees of significance of the model's parameters in discriminating between target and non-target performances at various other percentile contaminant concentrations do not differ greatly from those at the high-end exposure. This is easily seen from Figure 5-1, where the numbering of each significance "path" denotes the parameters as so numbered, i.e., ranked, in Table 5-1(f) for the 10th-percentile analysis. A natural means of grouping the parameters of EPAMMM is apparent from Figure 5-1, as follows:

Group I: The three coefficients of longitudinal, transversal and vertical dispersion in the saturated zone, and the distance from the source to the receptor site (upper "bundle" of paths in Figure 5-1(a)).

Group II: The source penetration depth (*H*); the rate of leachate infiltration; the hydraulic conductivity of the saturated zone; the seepage velocity, i.e., the rate of vertical movement of water downwards through the unsaturated zone; and the rate of recharge of the saturated zone (lower "bundle" of paths in Figure 5-1(a)).

Group III: All remaining parameters (Figure 5-1(b)).

Bearing in mind the fact that the following comments refer strictly to the case of contaminants that are non-biodegradable and have a low adsorption capacity, the Group I parameters are clearly key parameters whatever the percentile concentration, including the high-end (95th-percentile) exposure. These parameters are important in determining not only the overall degree of contaminant attenuation along its flow-path, but also the uncertainty attached to the resulting residual exposure concentration. For example, a comparison of the central tendencies of the distributions of Figures 4-1 and 4-2 shows that increasing the distance to the receptor site from 50-100m to 1000m decreases the exposure concentration from about 0.0500 to 0.0010 g/m^3 (for the loam soil), and from about 0.0650 to 0.0015 g/m^3 (for the sandy clay-loam). Further, Figure 5-2 shows that

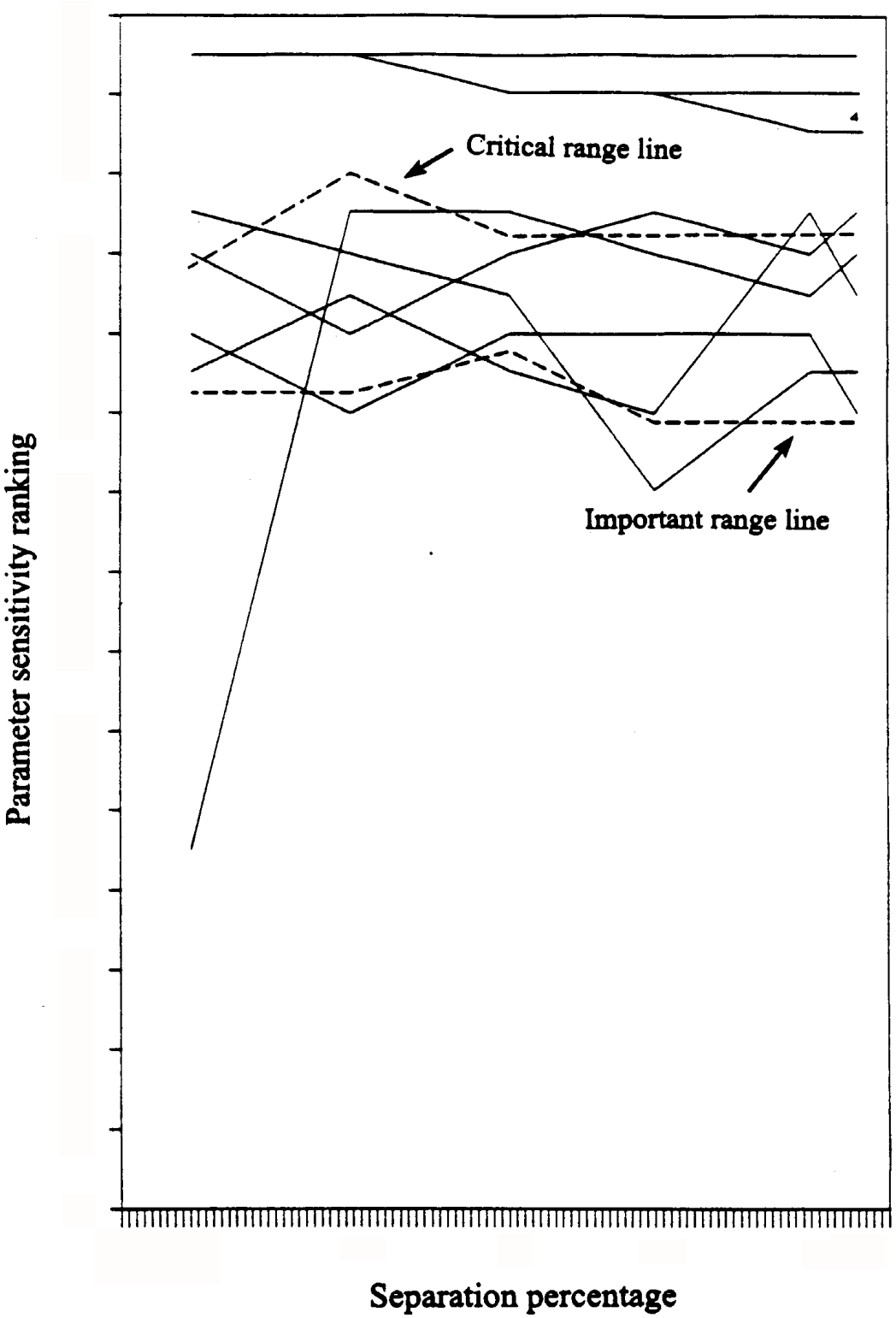


Figure 5-1 (a) Ranking of parameter sensitivity; numbers identifying parameters reflect the ordering of parameters in table 5-1 (f).

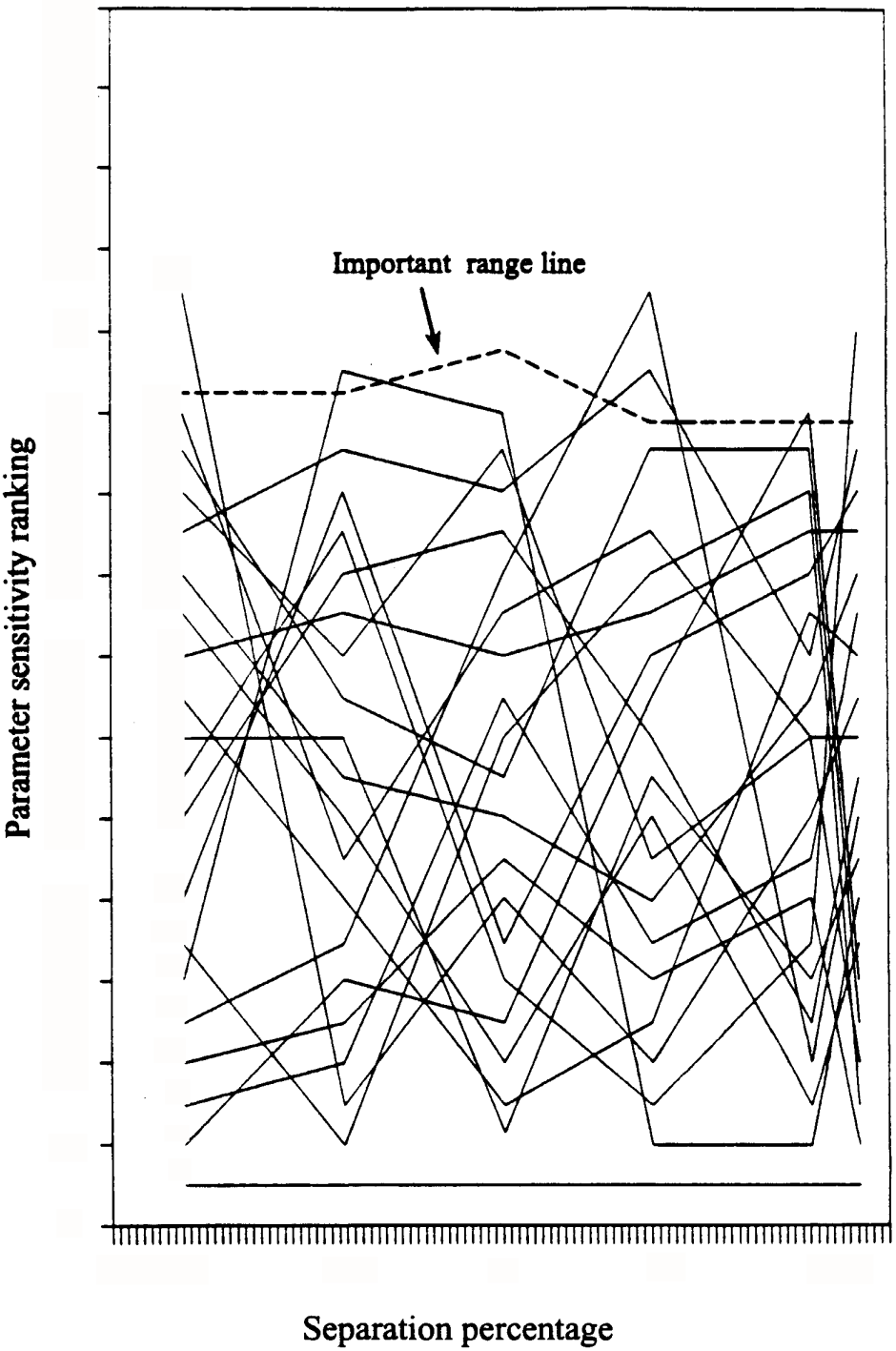


Figure 5-1 (b) Ranking of parameter sensitivity; numbers identifying parameters reflect the ordering of parameters in Table 5-1 (f).

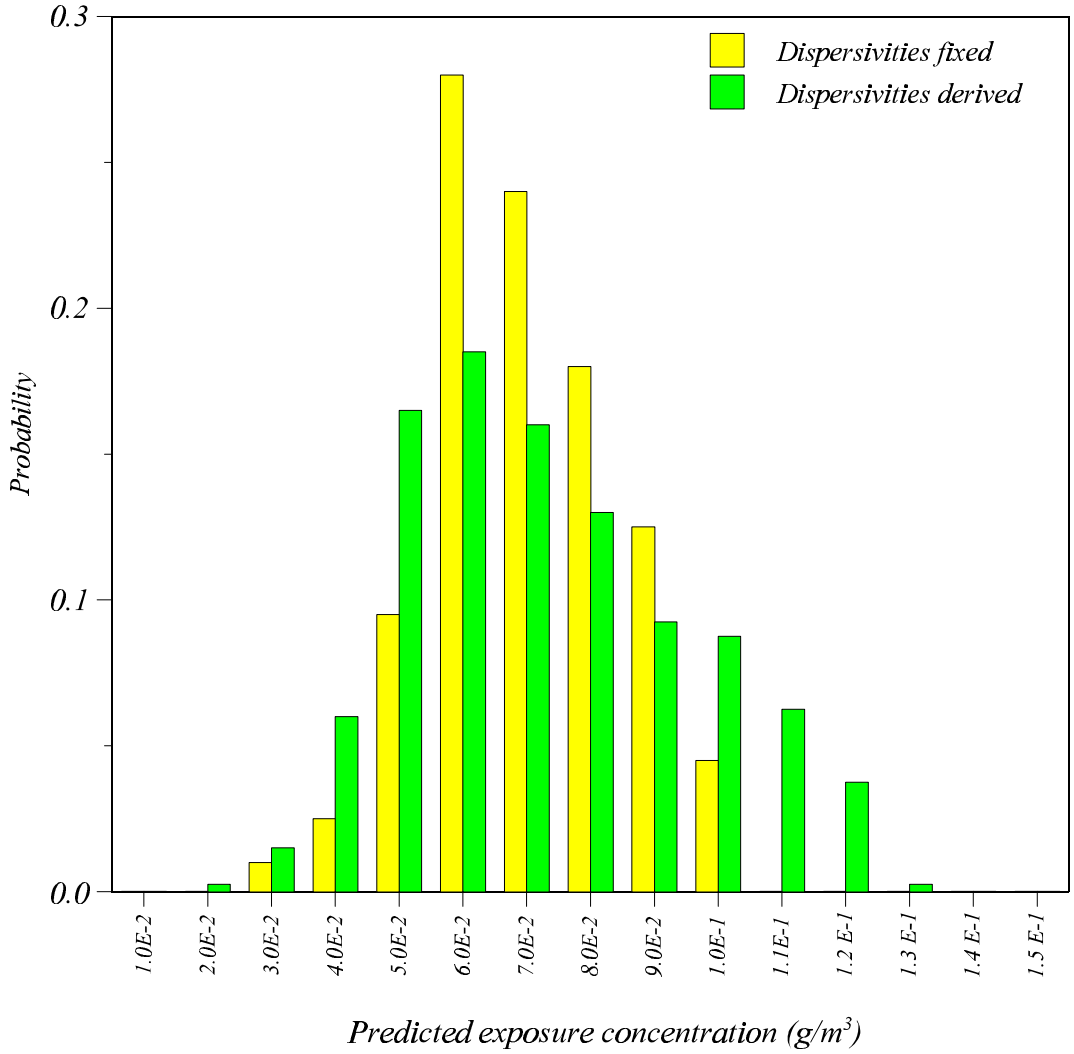


Figure 5-2 Exposure concentration distributions when dispersivities are given or derived.

when the coefficients of dispersion are assumed known with certainty, a significantly narrower band of exposure concentrations results. We will resume this form of analysis in the following section. Reflection on how well the elements of the Group I parameters might be known bears both encouraging and discouraging insights: whereas the distance to the receptor site might be well known, the same is not true for the coefficients of dispersion which, still less encouragingly, may be scale-dependent properties.

The significance of the Group II parameters is much less than that of the Group I parameters; in fact, it is more akin to that of the Group III (redundant) parameters. This is true even for parameter (5), the source penetration depth, which is a key parameter at the lowest percentile, but declines in significance towards the higher percentiles, parallel with the trend in the coefficient of vertical dispersion (parameter (4) in Figure 5-1(a)). The two are, indeed, related, source penetration being a function of the square-root of the coefficient of vertical dispersion, so that when the latter is assumed known with certainty, the significance of the source penetration depth disappears (as in the results of Table 5-2). The dramatic, yet opposite, differences in significance of parameters (7) and (21), i.e., between the 10th-percentile and all other percentiles, are a salient feature of Figures 5-1(a) and 5-1(b). These parameters are, respectively, the percentage of organic matter in the unsaturated zone and the rate of recharge of the saturated zone. If mechanisms of attenuation in the unsaturated zone are important at the lowest residual contaminant concentrations -- such as adsorption, which is a function of the organic matter content -- further dilution, as afforded by the rate of recharge of the saturated zone, is unlikely to play a key role. The results, then, have a measure of self-consistency, in the sense that a plausible explanation drawn from prior beliefs can be advanced to explain them.

Since the Group III parameters are largely redundant, a discussion thereof might likewise seem redundant. However, the significance of many of the parameters associated with the unsaturated zone fall into this third category, a result foreshadowed in the analysis of discriminating power in Section 4.1. Therefore, it is natural to enquire whether inclusion of the unsaturated zone module in EPAMMM is really essential, when the model is to be applied to a Subtitle D facility with respect to a **conservative** contaminant. Figure 5-3 shows the results of a test in which the module for the unsaturated zone has been excluded altogether. Under these specific conditions, the conclusion must be that the unsaturated zone plays no significant role and the associated module of software might be omitted, with significant consequential computational savings.

Taking stock of these results as a whole, let us recall that the trustworthiness of the model has been assessed with respect to a number of minor variations on the basic task of predicting a particular percentile value of the residual exposure concentration of the contaminant. Is it, therefore, especially well suited to one or the other of these minor task variations? Does this form of "wind-tunnel" testing of the current design of the model

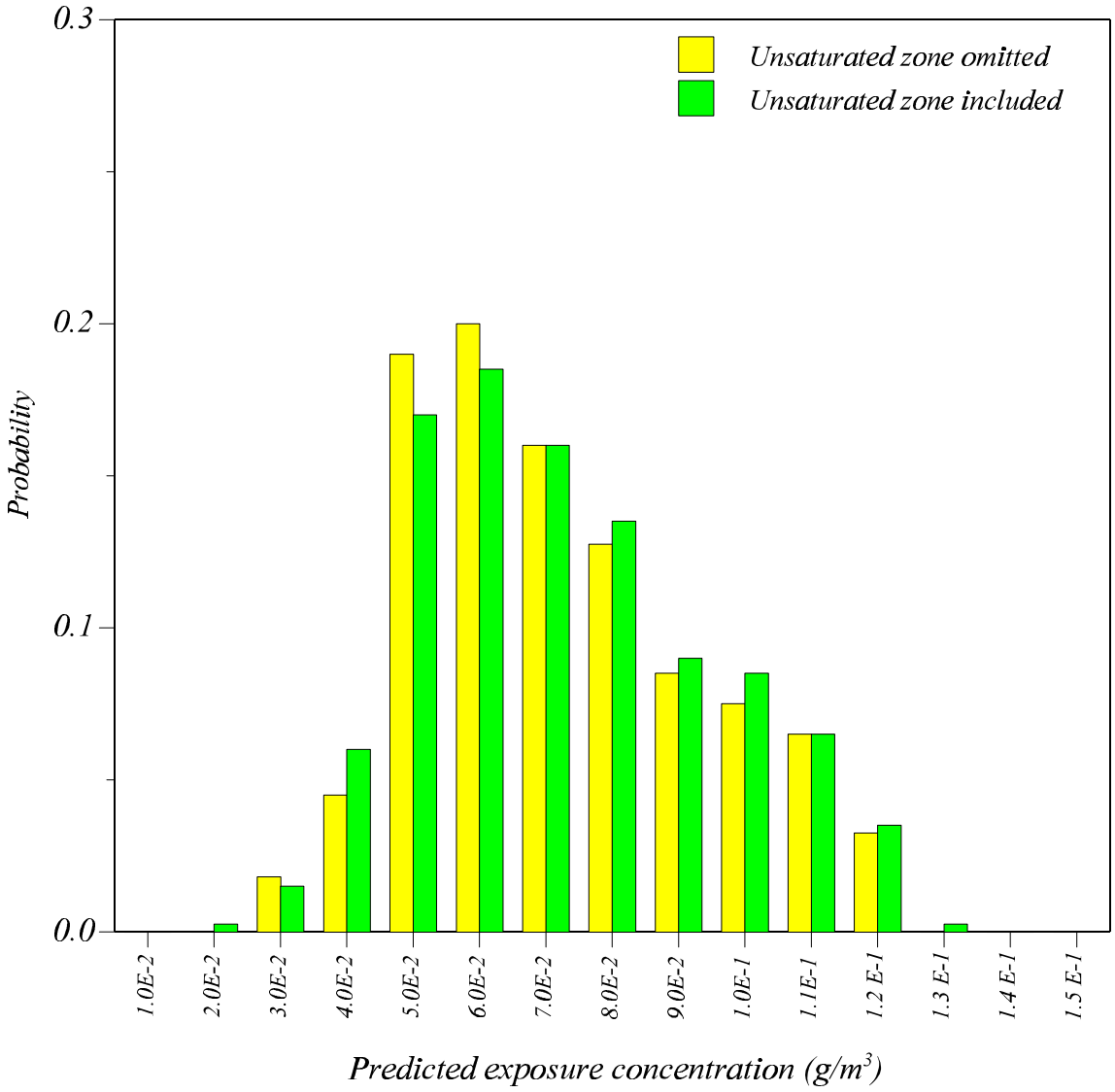


Figure 5-3 Exposure concentration distributions when the unsaturated zone is omitted.

indicate that it would be relatively more appropriate for achieving one form of "eventual flight" as opposed to another? Is there any summary, quantitative means of making such a judgement?

It has been argued elsewhere (Beck *et al* (1995; 1997)) that some numerical function of the key and redundant parameters, for example, simply the ratio of (key/total) numbers of parameters, might offer a means of judging the validity of a candidate model design. Or, where just a single model is being considered for several predictive tasks (as here), it could be suggested that the model will be more relevant (better suited) to performing one task than another when the maximum number of its constituent parameters are key to the performance of the given task. From Tables 5-1(a) through 5-1(f), however, we find that the number of key parameters varies between 4 and 6, out of a total of 29 parameters (in this particular application of the model). This small variation hardly seems a promisingly sensitive discriminant of whether a model should (or should not) be used for the given predictive task. Alternatively, a good model design for performing a predictive task might

be one in which all the constituent parameters have an important role to play with few redundant elements, a measure that might be constructed as some function of the Kolmogorov-Smirnov statistic (d_{max}) in Tables 5-1(a) through 5-1(f). This is a highly speculative assertion and we shall defer further discussion of it until Section 7.

6 Towards a Global Form of Sensitivity Analysis

The identification of which (key) constituent model parameters we would most want to have good knowledge of, as in the foregoing, is clearly important, but perhaps not sufficient. There may also be a need to determine just how well the key parameters should be known in order to perform the given task. This latter -- in fact, an inverted form of it -- is the question to which we now turn. In this last section of the present analysis our concern is to establish which of the key parameters identified above, if known better, would be of the greatest significance in reducing the uncertainty of the model's predictions? To be still more precise: to what extent would perfect knowledge of a key parameter induce a shift in the central tendency and/or the spread of the distribution of the predicted exposure concentrations?

In order to answer this question, a number of comparative sets of Monte Carlo simulation may be undertaken as follows, for each key parameter identified. The first set is generated in a manner identical with that already used in Section 4, i.e., with all the parameters of the model ($\boldsymbol{\theta}$) sampled from within their respective ranges of plausible values (as given in Table 4-1). This will be denoted as the reference case and the cumulative distribution function so obtained for the residual concentration (y) will be denoted as F. The other sets of simulations, for each (key) parameter, $\boldsymbol{\theta}_i$, are then generated according to the following procedure:

(i) *n* discrete values of the parameter θ_i are selected, uniformly spaced over the range of plausible values allowed for this parameter, i.e., θ_{i1}, θ_{i2}, ..., θ_{in};

(ii) a Monte Carlo simulation is conducted for **each** θ_{ik} ($k = 1, 2, ..., n$), in which all other parameters are assumed **not** to be known with certainty and are sampled across ranges identical with those of the reference case, thus yielding an accompanying cumulative distribution function F_{ik} for the exposure concentration (y).

The means, standard deviations and 95th-percentile values of F and F_{ik} are then compared. In practice n is chosen as five or six. Of the various key parameters previously identified, four were selected for this analysis: the distance between the source and receptor sites; the rate of leachate infiltration into the unsaturated zone; the hydraulic conductivity of the saturated zone; and the rate of recharge of the saturated zone.

Results

From Figure 6-1 we conclude that perfect knowledge of the hydraulic conductivity of the saturated zone is of marginal value: in sum, it has hardly any bearing on the properties of the distribution of the exposure concentration.

Quite the opposite is the conclusion in the case of the distance between the source and receptor sites. The distributions F_{ik} are entirely altered in the light of perfect knowledge of this parameter (Figure 6-2). Whether the distance is 50m or 100m has a significant bearing on the mean of the predicted exposure concentration (Figure 6-2(a)). Moreover, across the range of its values (θ_{ik}) the spread of the predicted distribution is substantially narrowed (according to both the 95th-percentile relative to the mean (Figure 6-2(a)) and the standard deviation of Figure 6-2(b)). In other words, providing the source-receptor distance is well known, and irrespective of its actual value, the residual exposure concentrations can be relatively "tightly" predicted.

Perfect knowledge of the rate of recharge of the saturated zone also has significant consequences, in the sense of shifting the whole of the distribution of exposure concentration as a function of the particular value assigned to this parameter (Figure 6-3(a)). The effect of dilution with increasing recharge rate is clearly evident in reducing the exposure concentration. However, any investment in acquiring perfect knowledge of this parameter would not bring benefits -- in terms of reducing the spread of the predicted distribution -- that are comparable with those deriving from perfect knowledge of the source-receptor distance, since there was hardly any diminution in the standard deviation of the F_{ik} relative to that of F (Figure 6-3(b)).

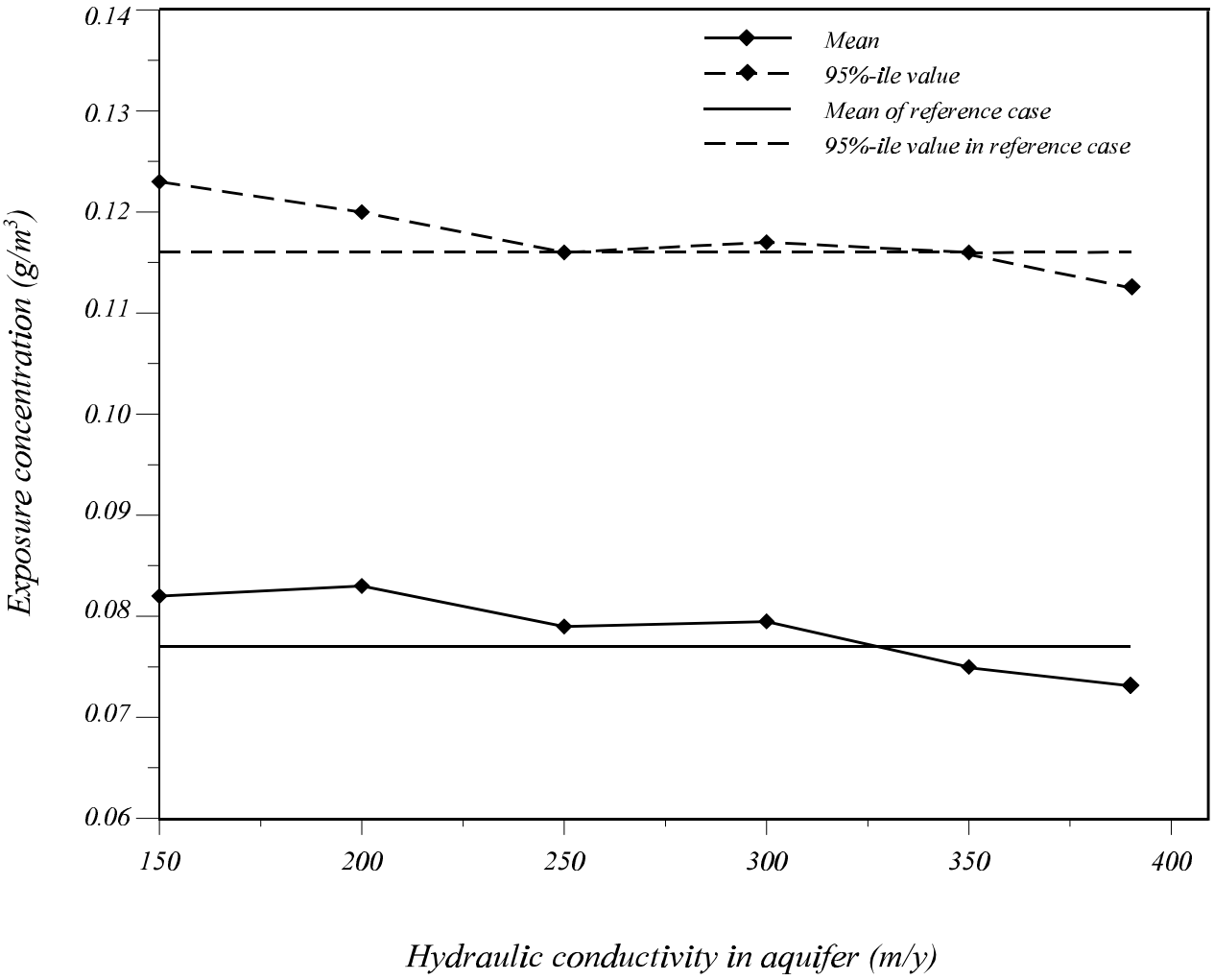


Figure 6-1 Behaviour of hydraulic conductivity in aquifer under global sensitivity analysis.

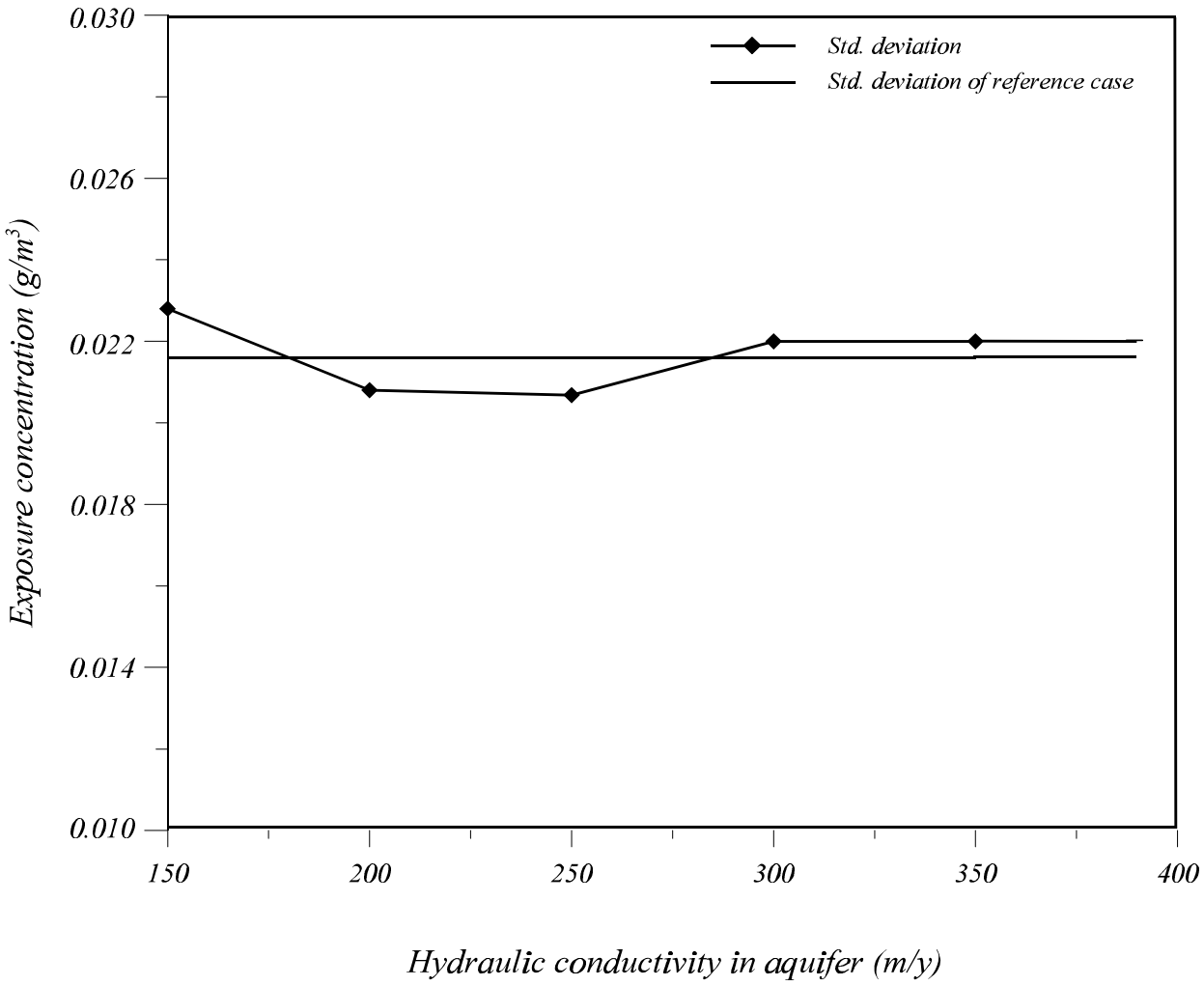


Figure 6-1 (b) Behaviour of hydraulic conductivity in aquifer under global sensitivity analysis.

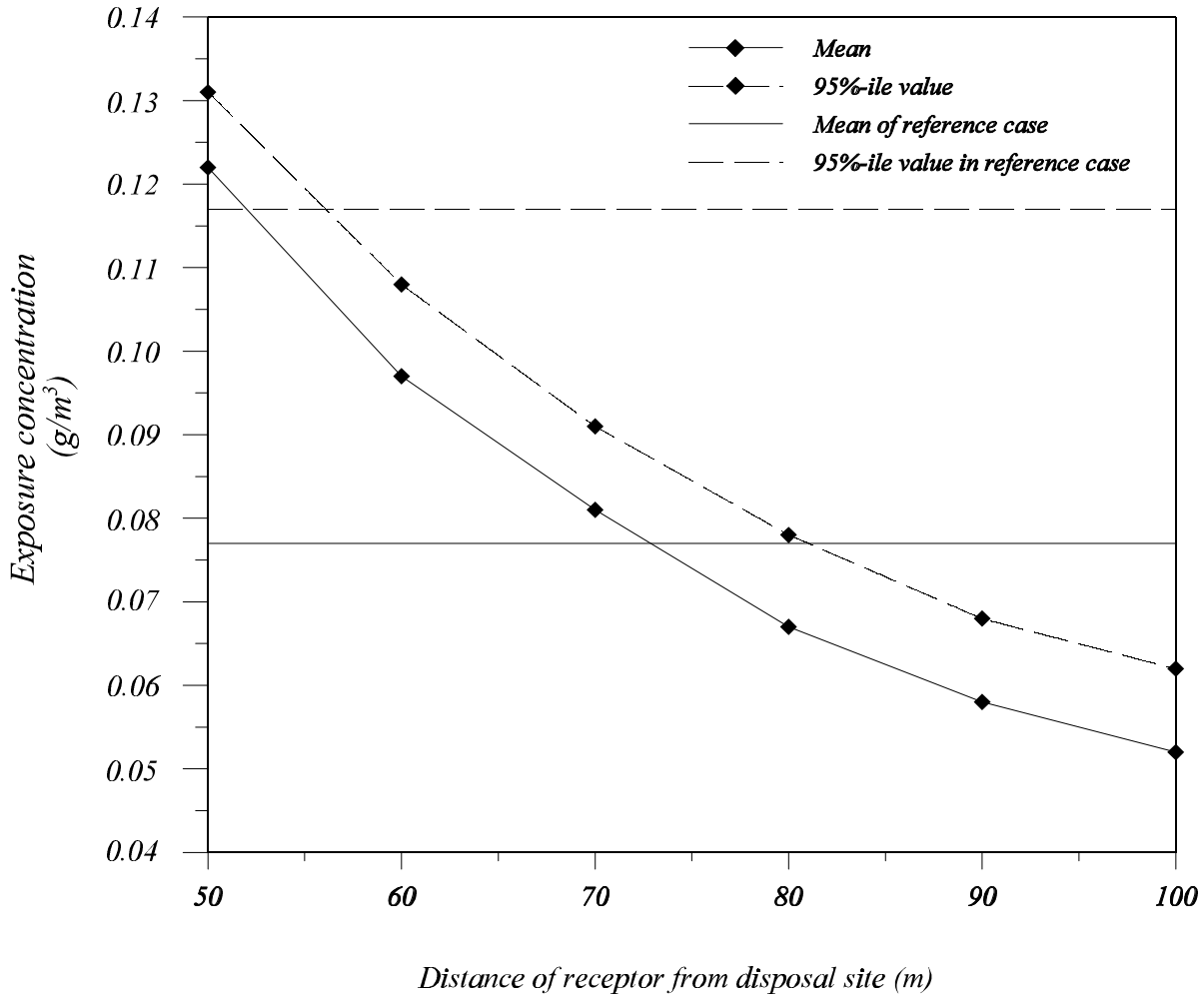


Figure 6-2 (a) The behaviour of the distance of receptor from disposal site under global sensitivity analysis.

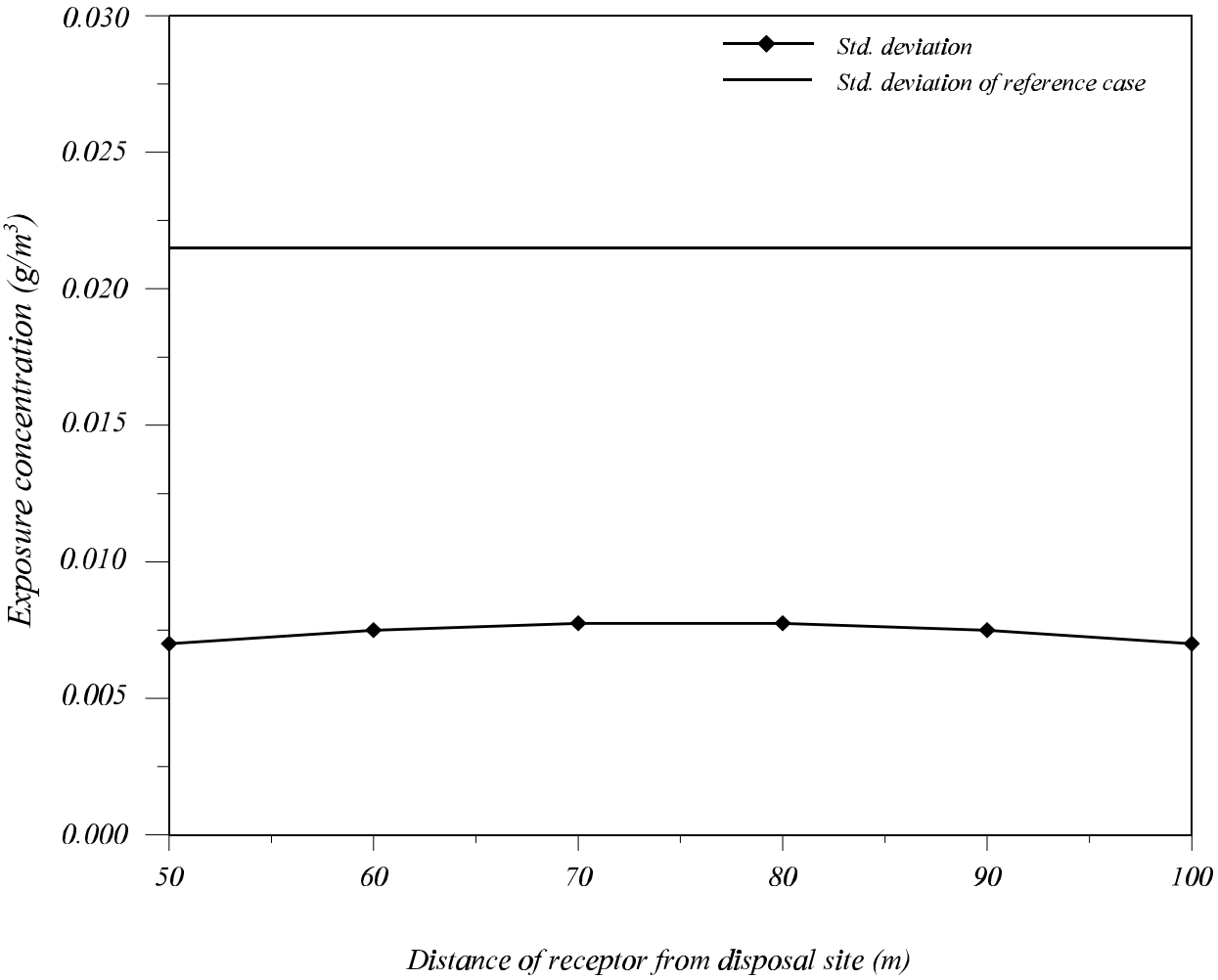


Figure 6-2 (b) The behaviour of the distance of receptor from disposal site under global sensitivity analysis.

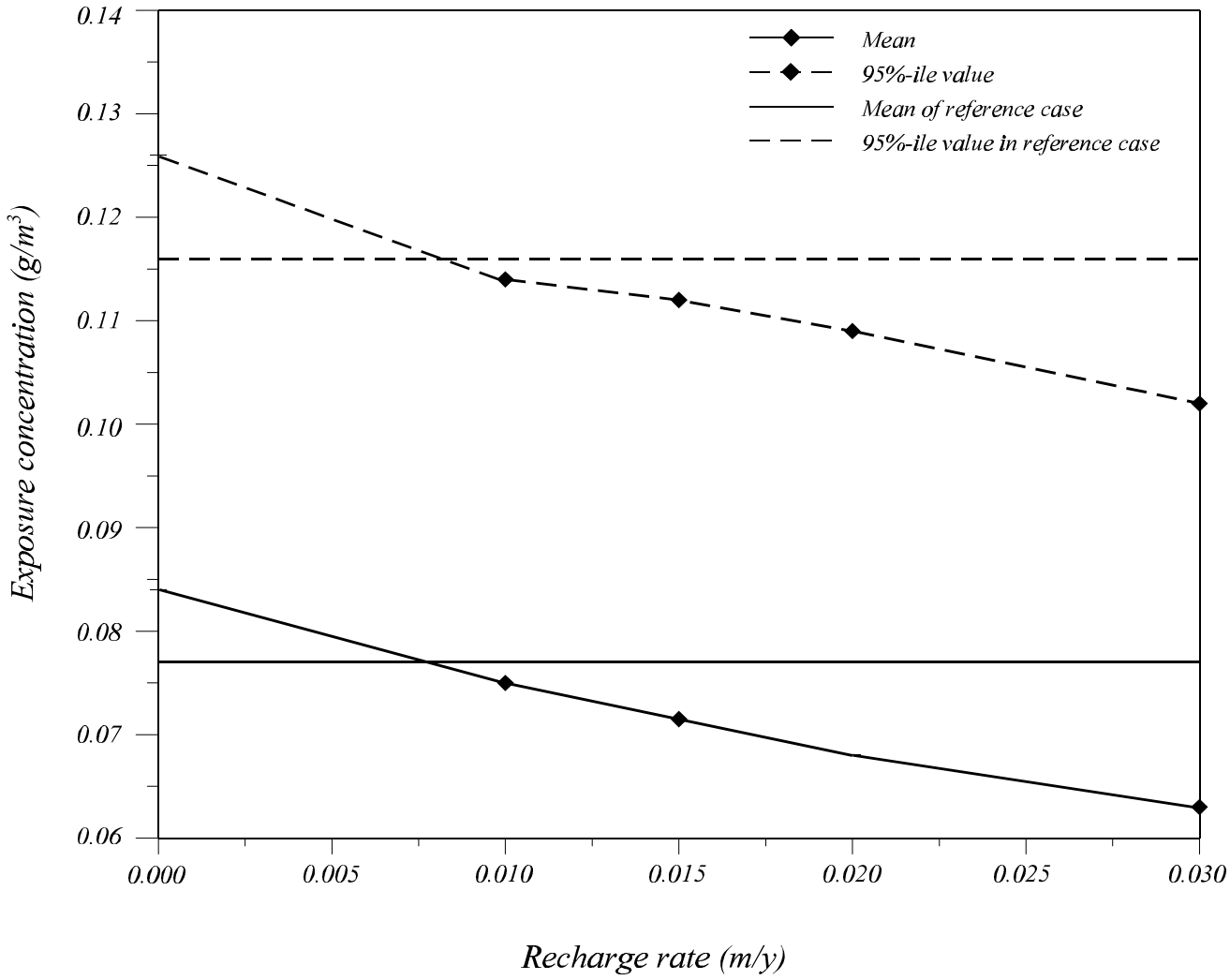


Figure 6-3 (a) *The behaviour of recharge rate under global sensitivity analysis.*

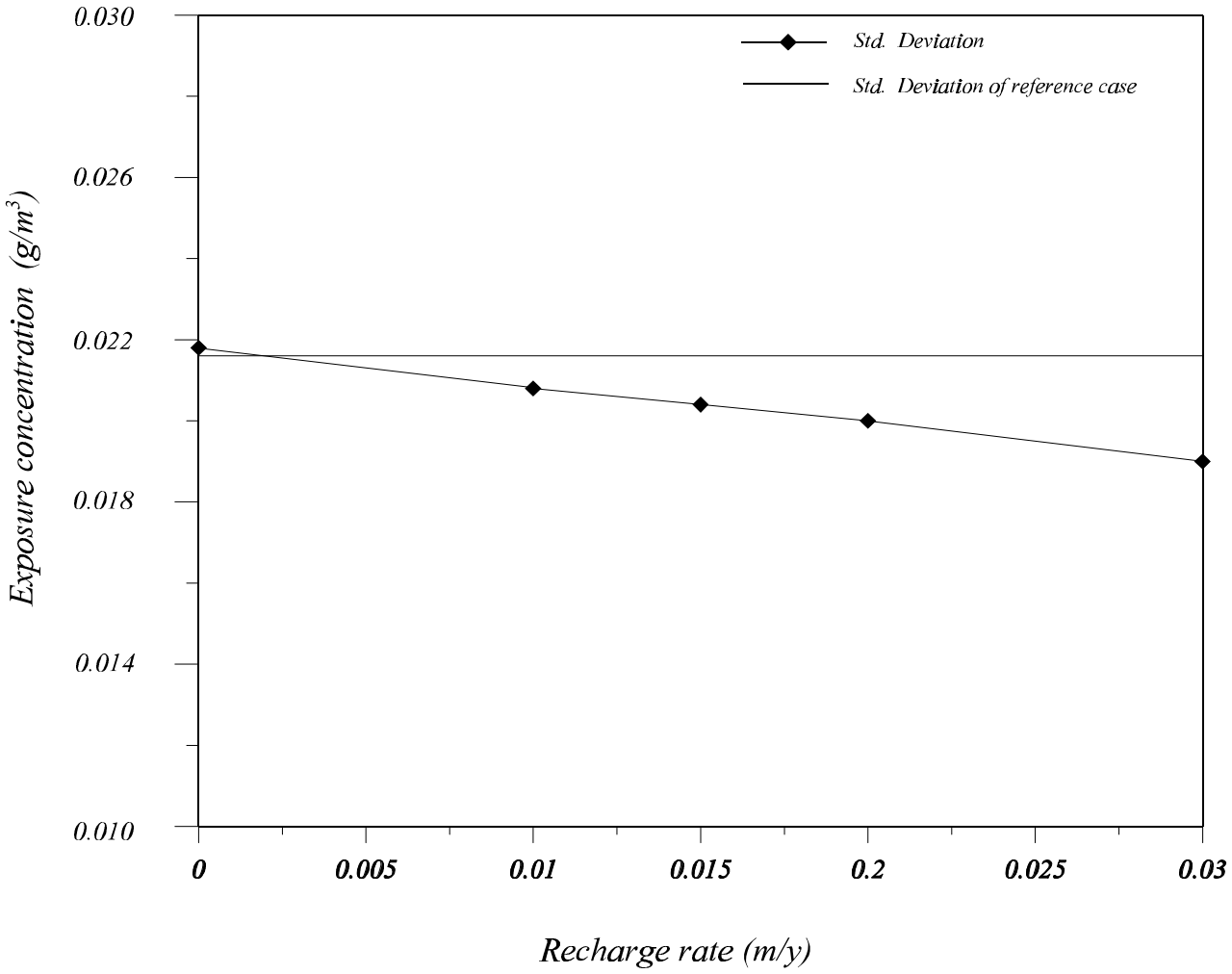


Figure 6-3 (b) The behaviour of recharge rate under global sensitivity analysis.

The results for the rate of infiltration of the leachate into the unsaturated zone are similar in significance to those of the rate of recharge of the saturated zone, although opposite in sense (Figure 6-4(a)). There is also evidence of a nonlinearity in the relationship between the infiltration rate parameter and the predicted exposure concentration. Above $0.3my^{-1}$ the predicted exposure concentration may exhibit a threshold effect (Figure 6-4(a)). Furthermore, this threshold value is predicted with notably less uncertainty than for the remainder of the ranges of prediction (as gauged by the decline in the standard deviation of the distribution in Figure 6-4(b)).

If further field observations were thought desirable, our results allow us to specify some priorities: one should seek first to improve knowledge of the source-receptor distance, then of either the recharge rate or leachate infiltration rate and, last of all, of the hydraulic conductivity (which is so often the primary target of aquifer characterisation). If one were to define the conditions most conducive to the making of relatively "reliable" predictions from EPAMMM for a Subtitle D facility, the site should have a relatively high hydraulic conductivity, a high infiltration rate, a high recharge rate, and a source-receptor distance that can be measured accurately. These conclusions, however, are conditional upon the particular nature of the task specification of Section 5, which relates to a non-biodegradable contaminant with low adsorption capacity.

7 Coming to a Judgement on the Trustworthiness of a Model

Our analysis has addressed the issue of establishing how well the Environmental Protection Agency's Multi-Media Model (EPAMMM) performs as a tool for discriminating between sites that are of concern, and those that are of no concern, with respect to the risk of contamination of the sub-surface environment. Throughout, we have sought to project the terms of the debate on model validation into a domain somewhat broader than the classical notion of demonstrating the match of the model with observed history (Konikow and Bredehoeft, 1992; Beck *et al*, 1997). In particular, in Section 5, we introduced the idea of judging the trustworthiness of the model according to a task specification -- prediction of a high-end or low-end exposure level -- and as a function of the internal attributes of the model (numbers of key and redundant parameters), as opposed to features associated with its output responses. In the conventional terms of matching history, quantification of the goodness of a model's performance is well known and straightforward: it is epitomised by the residual errors of mismatch between observed and simulated behaviour. Furthermore, the distinction between good and poor performance is obvious; it resides in the difference between the smallness or largeness of the residual errors. It is not nearly as immediately obvious how one would judge the quality of a model designed to fulfil a predictive task, since there are no histories of observed behaviour. In this section, we conjecture on the terms in which such a judgement might be developed. Our purpose is strictly to provide a framework within which to open a debate.

6-4 (a) Behaviour of infiltration rate under global sensitivity analysis.....missing, from available electronic records used to create this reprint.

6-4 (b) Behaviour of infiltration rate under global sensitivity analysis.....missing, from available electronic records used to create this reprint..

In the present analysis "goodness of performance" has been described somewhat informally in the following three ways:

(i) A well-performing model should generate predicted receptor site concentrations of contaminants that are distinctly different (in a statistical sense) for distinctly different site, field, and contaminant characteristics. If our prior (subjective) expectations are consistent with the predicted result -- in the sense that what we believe to be quite different sites **are** associated with quite different predictions of contaminant concentrations -- this has a

self-reinforcing effect. We are comfortable with the model and our prior expectations have been confirmed. If the expectation and outcome are inconsistent, either further investigation will be necessary in order to reconcile the two (by a change of expectation or model or both), or one could accept the expectation and the model as conditionally valid, but merely served by "insufficiently certain" knowledge of the model's parameters and inputs.

(ii) Given a specified task of prediction, such as the task of predicting a high-end exposure concentration at the receptor site, a well-performing model will contain a relatively large proportion of **key** parameters, i.e., parameters for which the choice of a particular value (from within the range of possibilities) is critical to discriminating between whether any given exposure concentration is predicted or not. Moreover, there will be greater confidence in the model the better these key parameters are believed to be known (or identifiable from some prior exercise in model calibration). The result that all (or a majority of) the model's key parameters are believed to be the least well quantified, for example, would be most disquieting for any judgement about the reliability of such a model in performing a predictive task. This outcome of the analysis will be all the more disconcerting if these key parameters can only be better quantified as a consequence of further highly costly experimentation and field monitoring.

(iii) An ill-performing model may be capable of being improved, as the foregoing suggests (and as widely appreciated, e.g., Beck, 1987; Janssen, 1994), providing the uncertainty attached to the knowledge of the model's key parameters can be reduced. A good model is, therefore, one in which perfect knowledge of the values of key parameters would permit (in principle) substantial reductions in the uncertainty attached to the model's predictions.

In none of these three elements is there a presumption of the availability of actually

observed (past) performance, i.e., measured contaminant concentrations at the receptor site against which to evaluate the trustworthiness of the model. In a screening-level analysis, the predominant concern lies **not** so much in demonstrating that history has been matched (Konikow and Bredehoeft, 1992) but in calibrating the performance of the model with reference to some predictive task (Beck *et al*, 1997). A good design of the model for the purpose of performing this task is what is sought. In the above, element (ii), in particular, is oriented towards an assessment of this feature and we now explore what might be revealed through an analysis of the d_{max} statistics of Tables 5-1.

An Indicator of the Quality of the Model's Design

For illustrative purposes in presenting the potential power of a novel way of assessing whether a model is (relatively) better or worse suited to performing a given predictive task, we focus on Tables 5-1(a) and 5-1(d), which give values of d_{max} for the tasks, respectively, of predicting the 95th- and 50th-percentiles of the exposure concentration at the receptor site.

Figure 7-1 shows the (normalised) frequency **distributions** of a statistic, $[(d_{max}(i)/d^{*})-1]$, we have developed for comparing the performances of the model against the two predictive task specifications. As before, for any parameter θ_i, $d_{max}(i)$ is the maximum separation of the cumulative distributions of the "target behaviour-giving" values $\{\theta_i(T)\}$ and the "not-target behaviour-giving" values $\{\theta_i(\tilde{T})\}$. d^{*} is a value of the Kolmogorov-Smirnov statistic chosen to discriminate with a given degree of confidence between significant and insignificant maximum separations of these two cumulative distributions. d^{*} has, therefore, the same role as $d_{m,n}$ previously, yet unlike $d_{m,n}$ will be invariant and independent of the differing magnitudes of m and n arising from tests against the various task specifications. The $d_{m,n}$ values for a given level of confidence can be seen to vary somewhat in Tables 5-1(a) through 5-1(f), for example, as a function of the varying m and n associated with varying numbers of random candidate parameterisations classified as behaviour- and not-behaviour-giving for the different task specifications. To summarise, d^{*} has here the function of a normalising parameter, allowing different **distributions** of the model's "parametric significances" -- with respect to discriminating whether the task specification is matched or not -- to be compared on a consistent basis, irrespective of the given task. In principle, our statistic may be used to compare either the performance of the same model against different tasks (as herein) or the performances of different models, with different numbers of parameters, against the same task specification. Once normalised as a plot of the relative frequency distributions for $[(d_{max}(i)/d^{*})-1]$ (the parametric significances), the number of parameters in any model (indexed here through i) can to some extent be abstracted from the judgement on the quality of model performance. The effect of d^{*} is to scale the plot of the distribution of parametric significances so that 0.0 separates insignificance (redundancy) from significance (the results of Figure 7-1 are based on a value of d^{*} reflecting a level of

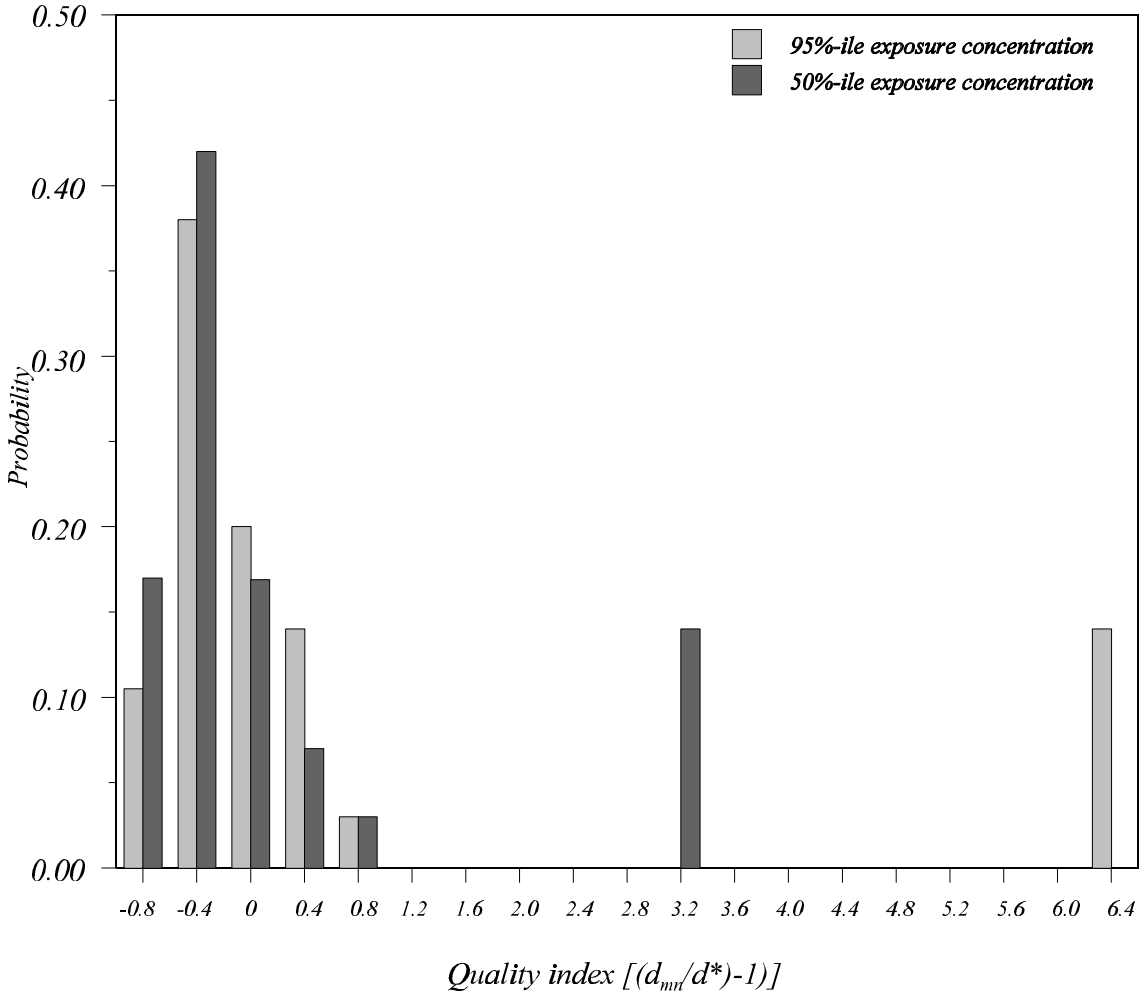


Figure 7-1 Probability distribution of the index of the quality of model design.

confidence of 0.001). In general, if the distribution of the parametric significances were skewed towards the right, this would suggest the model contains a relatively large number of parameters that are key in the performance of the specified task. If the distribution were skewed towards the left, the model might be said to be suffering from a preponderance of redundant parameters, relative to the task at hand.

Figure 7-1 shows that in performing predictive tasks, a majority of the constituent parameters of EPAMMM appear to be redundant, especially so in the case of predictions required to discriminate exposures above and below the 50th-percentile. At the same time, a small number of the model's parameters (13%) are critical to the performance of the two tasks, notably more so in the case of the 95th-percentile task. In very general terms we might be tempted to conclude from Figure 7-1 that EPAMMM is **better** suited to performing the task of predicting high-end exposures **relative** to the prediction of mean exposure concentrations[2]. Put in more familiar terms, we might say that EPAMMM is a better-designed tool for predicting high-end as opposed to mean exposure concentrations. This conclusion would be subject, of course, to the qualifying statement that the parameters associated with the highly positive values for the statistic in Figure 7-1 are relatively well known.

Having been tempted to draw such conclusions, however, it must be noted that the validity of computing and using the distribution of our proposed statistic (or index) has yet to be fully evaluated by much more extensive analyses. For instance, the legitimacy of using a single value for d^*, when $d_{m,n}$ varies significantly as a function of small magnitudes for either m or n for some of the assessments of the model's performance against the various task specifications, has to be established. Similarly, judgements about the character of the relative frequency distribution of our statistic may be compromised in cases where the number of parameters in the model is very small. We also note that use of the Kolmogorov-Smirnov statistic may have its limitations and that assessment of the model as a function of attributes of its **individual** parameters, as opposed to key and redundant **clusters** of parameters, should be interpreted with great care (Spear ***et al***, 1994). Nevertheless, here -- in Figure 7-1 -- is a quantitative measure of the quality of a model conditioned upon how that model performs a task of prediction, including projection into utterly novel conditions, **not** upon how the model matches observed past behaviour. This measure is cast in terms of the internal features of the model itself. Further, we might imagine that a particular shape of the distribution of $[(d_{max}(i)/d^*)-1]$ could be attached to the concept of a well-designed model. For example, this might be a distribution in which there are not too many redundant parameters (associated with the left tail of the distribution) nor a few excessively key parameters, affecting the distribution towards its right tail. In these respects, we argue that our new statistic and its associated analysis have, in principle, great appeal. This appeal,

[2] Note that this is not a statement of an absolute property but rather one of a more "relativistic" character.

moreover, is tied closely to the practical purpose of decision-making for which the model was designed and developed.

Complexity in the Type and Weight of Evidence

Put briefly, given a predictive task, should we use a given model? If the answer is in the affirmative, according to what evidence have we arrived at this judgement? If the answer is in the negative, why should this be so, which parts of the model are defective, and what do we need to know in order to remedy these defects?

What, then, can we say of the performance of EPAMMM as a result of its being subjected to the above battery of tests relative to its intended application in screening a large number of storage (Subtitle D) facilities for the potential off-site exposures and risks due to contamination arising therefrom?

EPAMMM does not discriminate strongly among predictions of receptor-site contaminant concentrations for different soil types, a result inconsistent with prior expectation. If we wish to leave this expectation unchallenged, we could ascribe poor model performance to the use of easily available soil characteristics as surrogates for the not-so-easily available, but more vitally important, hydrological characteristics of the sites. For a given site and soil type, however, the model discriminates well among different residual concentrations arising from contaminants differing strongly in their migration and attenuation mechanisms (biodegradation, hydrolysis, and adsorption). This seems at first site an attractive property. Yet, upon brief reflection, it indicates that the discriminating power of the model is a function of parameters that are notoriously difficult to quantify under field conditions.

When given the task of predicting high-end exposure concentrations, knowledge of how to parameterise the contaminant attenuation mechanisms -- other than attenuation by dilution -- is the most significant item of quantitative information needed for good performance. This conclusion seems obvious, until one recalls the crucial role that "dilution as the solution to pollution" has played in so much of our decision-making in the past. The same conclusion is robust across a range of other tasks, i.e., across the prediction of other (low, moderate) exposure concentrations, but the quality of the model's performance shows diverging features. According to the distributions of our proposed measure of the quality of a candidate model (a manipulation of the Kolmogorov-Smirnov test statistic), EPAMMM appears to be well suited to the task of predicting the 95th-percentile concentration yet ill suited to that of predicting the 50th-percentile concentration.

If further field observation and site characterisation were thought desirable, one should seek first to improve knowledge of the source-receptor distance, then of either the

recharge rate or leachate infiltration rate and, last of all, of the hydraulic conductivity (which is so often the primary target of aquifer characterisation). In this sense -- of a capacity to perform well when given appropriately good knowledge of the system under investigation -- EPAMMM shows much promise as a model of quality.

All of these conclusions, however, are based on a complex assembly of evidence, in which there is a danger of overlooking some of the subtleties of the way in which the battery of tests has been constructed and applied in order to arrive at such judgements on the quality of the model. Interpretation of the results of the tests, moreover, is far from straightforward and in fact requires a fairly intimate knowledge of the inner workings of the model. This is troubling, for EPAMMM is not a notably complex, high-order model. It is also troubling because one is seeking to distil all of the complexity and subtlety down to the essential simplicity of a choice between just two alternatives: to trust or not to trust the use of the model in performing a predictive task.

8 Conclusions

Discriminating which might be the more problematic sites for the storage of hazardous materials from those that are unlikely to be problematic -- in order to set priorities for allocating the scarce resources for remediation -- is a task of considerable current interest. It is beset with great uncertainty. There are many such sites whose performance has not been well monitored; and the liquid contaminants from most of these sites are likely to have their greatest impact on the subsurface aquatic environment, whose properties are intrinsically more difficult to characterise than those of the surface water environment. If a model is to be used to support the decisions on what is, and what is not, to be a site requiring remediation, it is highly pertinent to ask whether the performance of the model in fulfilling this task is rendered ineffective by the uncertainty and, more broadly, for what predictive tasks would the model be well or ill suited. Both avenues of enquiry have been the subject of this paper, in the specific context of using the US Environmental Protection Agency's Multi-Media model (EPAMMM) for Subtitle D storage facilities.

Our first conclusion is that, for the conditions of the tests constructed herein, characterisation of a site's subsurface hydrological behaviour on the basis of the more readily available soil-type parameters and knowledge of the often-sought hydraulic conductivity of the saturated zone are, in general, not critical to the discriminating power of EPAMMM. In fact, for the predictive power of the model to rise above the obscuring effects of all the uncertainties, good knowledge of the source-receptor distance and chemical and biological (as opposed to physical) mechanisms of contaminant attenuation would appear to be most vital. Such a conclusion has an element of counter-intuition in it, an element of promise (successful application of the model may depend on something that should be easily quantifiable), but also the detraction of that successful application

being reliant upon properties that are notoriously difficult to estimate in the field. In order to arrive at this kind of conclusion, which -- in the absence of the many qualifying conditions of the model tests -- is clearly much simplified, we have introduced a more **globally** applicable form of senstivity analysis than the now well known **regionalised** sensitivity analysis of Hornberger, Spear and Young.

Our second conclusion is the speculation that EPAMMM is better suited to the prediction of high-end exposure concentrations (95%-ile) than average exposure concentrations (50%-ile) at the defined receptor site. Such a statement is speculative because it is based on but a preliminary analysis using a novel measure of the quality of a model in performing a predictive task, **without** recourse to any quantification of the extent to which the model can match an observed historical record. This is therefore directly in line with recent statements on a means of escape from the conventional impasse of procedures for model validation (Beck ***et al***, 1995; 1997): of having to assess the trustworthiness of a model for projection into the unknown as a function solely of its consistency with past observed (and possibly irrelevant) conditions. The index is based on a manipulation of the Kolmogorov-Smirnov statistic for comparing sample distributions of the model's parameters, a feature which also lies at the core of a regionalised sensitivity analysis. This index therefore gauges the quality of the model (in performing a predictive task) in terms of attributes of its parameters, i.e., in terms of its internal structure, as opposed to attributes of its outputs.

Last, the battery of tests applied to EPAMMM in this paper, together with the results of others of a more conventional nature (when applicable), could form the basis of a systematic protocol for model validation. However, the experience of this prototypical case study gives us cause for concern, on the following account. Interpretation of the test results is not at all straightforward. It demands very careful attention to the precise details of the conditions assumed for the tests and a rather comprehensive mental model of the inner workings of the mathematical model. Quite apart from the fact that this mental model may actually be defective, it is not hard to imagine that the outcome of applying the mooted protocol would be a statement circumscribed by so many restrictive qualifications as to render the judgement "this model is valid for its given task" almost without import.

References

Armstrong, A, Addiscott, T and Leeds-Harrison, P (1995), "Methods for Modelling Solute Movement in Structured Soils", in *Solute Modelling in Catchment Systems*, (S T Trudgill, ed), Wiley, Chichester, UK, pp 133-161.

Beck, M B (1987), "Water Quality Modeling: A Review of the Analysis of Uncertainty", *Water Resources Research*, **23**(8), pp 1393-1442.

Beck, M B (1994), "Understanding Uncertain Environmental Systems", in *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, (J Grasman and G van Straten, eds), Kluwer, Dordrecht, pp 294-311.

Beck, M B and Halfon, E (1991), "Uncertainty, Identifiability and the Propagation of Prediction Errors: A Case Study of Lake Ontario", *J Forecasting*, **10**(1&2), pp 135-161.

Beck, M B, Jakeman, A J and McAleer, M J (1993), "Construction and Evaluation of Models of Environmental Systems", in *Modelling Change in Environmental Systems*, (A J Jakeman, M B Beck, and M J McAleer, eds), Wiley, Chichester, UK, pp 3-35.

Beck, M B, Mulkey, L A, Barnwell, T O and Ravetz, J R (1995), "Model Validation for Predictive Exposure Assessments", in *Proceedings 1995 International Environmental Conference*, TAPPI, Atlanta, Georgia, pp 973-980.

Beck, M B, Ravetz, J R, Mulkey, L A and Barnwell, T O (1997), "On the Problem of Model Validation for Predictive Exposure Assessments", *Stochastic Hydrology and Hydraulics*, **11**(3), pp 229-254.

Blackburn, J W (1989), "Is There an 'Uncertainty Principle' in Microbial Waste Treatment?", in *Biotreatment of Agricultural Wastewater*, (M Hurtley, ed), CRC Press, Boca Raton, Louisiana, pp 149-161.

Carsel, R F and Parrish, R S (1988), "Developing Joint Probability Distributions of Soil Water Retention Characteristics", *Water Resources Research*, **24**(5), pp 755-769.

Chen, J (1993), "Modelling and Control of the Activated Sludge Process: Towards a Systematic Framework", *PhD Thesis*, Imperial College of Science, Technology, and Medicine, London.

Davis, P A, Zach, R, Stephens, M E, Amiro, B D, Reid, J A K, Sheppard, M I, Sheppard,

S C and Stephenson M (1993), "The Disposal of Canada's Nuclear Fuel Waste: The Biosphere Model, BIOTRAC, for Postclosure Assessment", *Report AECL-10720, COG-93-10*, Atomic Energy of Canada Limited, Chalk River, Ontario.

Dougherty, D E and Bagtzoglou, A C (1993), "A Caution on the Regulatory Use of Numerical Solute Transport Models", *Ground Water*, **31**(6), pp 1007-1010.

Environmental Protection Agency (1991), "Guidelines for Exposure Assessment", *Science Advisory Board Draft Final*, United States Environmental Protection Agency, Washington, D C (June).

Ewen, J (1995), "Contaminant Transport Component of the Catchment Modelling System SHETRAN", in *Solute Modelling in Catchment Systems*, (S T Trudgill, ed), Wiley, Chichester, UK, pp 417-441.

Gee, G W, Kincaid, C T, Lenhard, R J and Simmons, C S (1991), "Recent Studies of Flow and Transport in the Vadose Zone", *Reviews of Geophysics*, **29**, pp 227-239.

Goodrich, M T and McCord, J T (1995), "Quantification of Uncertainty in Exposure Assessments at Hazardous Waste Sites", *Ground Water*, **33**(5), 727-732.

Hornberger, G M and Spear, R C (1980), "Eutrophication in Peel Inlet, I. Problem-defining Behaviour and a Mathematical Model for the Phosphorus Scenario", *Water Research*, **14**, pp 29-42.

Janssen, P H M (1994), "Assessing Sensitivities and Uncertainties in Models: A Critical Evaluation", in *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, (J Grasman and G van Straten, eds), Kluwer, Dordrecht, pp 344-361.

Jury, W A (1982), "Simulation of Solute Transport Using a Transfer Function Model", *Water Resources Research*, **18**(2), pp 363-368.

Kendall, M G and Stuart, A (1961), "*The Advanced Theory of Statistics*", Griffin, London.

Konikow, L F and Bredehoeft, J D (1992), "Ground-water Models Cannot Be Validated", *Advances in Water Resources*, **15**(1), pp 75-83.

McLaughlin, D, Kinzelbach, W and Ghassemi, F (1993), "Modelling Subsurface Flow and Transport", in *Modelling Change in Environmental Systems*, (A J Jakeman, M B Beck, and M J McAleer, eds), Wiley, Chichester, UK, pp 133-161.

National Research Council (1990), "*Ground Water Models. Scientific and Regulatory*

Applications", National Academy Press, Washington, D C.

Onishi, Y, Shuyler, L and Cohen, Y (1990), "Multimedia Modeling of Toxic Chemicals", in *Proceedings of the International Symposium on Water Quality Modeling of Agricultural Non-point Sources, Part 2*, (D G DeCoursey, ed), **Report ARS-81**, Agricultural Research Service, United States Department of Agriculture, pp 479-502.

Oreskes, N, Shrader-Frechette, K and Belitz, K (1994), "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences", *Science*, **263**, 4 February, pp 641-646.

Salhotra, A M, Sharp-Hansen, S and Allison, T (1990), "Multimedia Exposure Assessment Model (Multimed) for Evaluating the Land Disposal of Wastes - Model Theory", *Report* (Contract Numbers 68-03-3513 and 68-03-6304), Environmental Research Laboratory, United States Environmental Protection Agency, Athens, Georgia.

Schanz, R W and Salhotra, A M (1990), "Estimating Cleanup Levels at Hazardous Waste Sites", in *Superfund '90*, Proceedings of the 11th National Conference, Washington D C, pp 157-160.

Schnoor, J L, Sato, C, McKechnie, D and Sahoo, D (1987), "Processes, Coefficients, and Models for Simulating Toxic Organics and Heavy Metals in Surface Waters", *Report EPA/600/3-87/015*, Environmental Research Laboratory, United States Environmental Protection Agency, Athens, Georgia.

Sharp-Hansen, S, Travers, C and Allison, T (1990), "Subtitle D Landfill Application Manual for the Multimedia Exposure Assessment Model (Multimed)", *Report* (Contract Number 68-03-3513), Environmental Research Laboratory, United States Environmental Protection Agency, Athens, Georgia.

Skiles, J L, Redfearn, A and White, R K (1991), "Determining the Number of Samples Required for Decisions Concerning Remedial Actions at Hazardous Waste Sites", *J Environmental Engineering and Management*, **1**(3), pp 57-61.

Smith, R E (1992), "OPUS: An Integrated Simulation Model for Transport of Nonpoint-source Pollutants at the Field Scale", *Report ARS-98*, Agricultural Research Service, United States Department of Agriculture, Fort Collins, Colorado.

Soil Conservation Service (1972), "Hydrology", Section 4, *SCS National Engineering Handbook*, NEH-Notice 4-102, United States Department of Agriculture, Washington, DC.

Spear, R C and Hornberger, G M (1980), "Eutrophication in Peel Inlet, II. Identification of Critical Uncertainties Via Generalised Sensitivity Analysis", *Water Research*, **14**, pp 43-49.

Spear, R C, Grieb, T M and Shang, N (1994), "Parameter Uncertainty and Interaction in Complex Environmental Models", *Water Resources Research*, **30**(11), pp 3159-3169.

van Genuchten, M T (1976), "A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils", *Soil Science Society Journal*, **4**, pp 892-898.

Varis, O (1995), "Belief Networks for Modelling and Assessment of Environmental Change", *Environmetrics*, **6**, pp 439-444.

Young, P C, Hornberger, G M and Spear, R C (1978), "Modelling Badly Defined Systems - Some Further Thoughts", in *Proceedings SIMSIG Simulation Conference*, Australian National University, Canberra, pp 24-32.