Large Simulation Models: Calibration, Uniqueness and Goodness of Fit

Robert C. Spear
Environmental Engineering and Health Sciences Laboratory
University of California
Berkeley, California

Abstract Large simulation models of environmental systems which are based on biological and physical mechanisms are useful because of their ability to integrate diverse types of information relevant to the problem under analysis. Inherent in such models is a high degree of both structural and parametric complexity. In a number of studies using such models, which have also used the Regional Sensitivity Analysis concept, it has been found that there are many parameter sets which produce good fits to calibration data. This lack of uniqueness requires a different perspective on parameter estimation which can be usefully addressed employing computer-intensive methods of multivariable statistical analysis.

1. Introduction

In 1977 I came to Australia to spend a sabbatical year in Peter Young's group at the Centre for Resource and Environmental Studies of the ANU. There were a variety of projects going on at CRES at that time that afforded interesting opportunities for visitors like myself as well as an atmosphere of activity and vigorous discussion. Not surprisingly, a topic of central interest concerned the appropriate perspectives and methods to bring to the analysis of environmental systems. Given Peter Young's interests, there was a focus on time series methods and something of a philosophical aversion to large simulation models. To some extent this stemmed from the perception that the degrees of freedom available to fit the field or experimental data were, in some sense, excessive. As I recall, most of us subscribed to the view that, given a large model with lots of parameters, it was possible to fit any data set free of pathology with a little judicious fiddling.

George Hornberger was also at CRES on sabbatical that year and, after some particularly vigorous discussion which I now only vaguely recall, he and I decided to seek a counter example and select, out of one of the then current CRES studies, a problem that might be elucidated through the use of a simulation model based on physical, chemical and biological principals. A perfect opportunity was presented by a CRES project which involved data integration and analysis from several other groups studying a eutrophication problem in the Peel-Harvey Estuary of Western Australia. At that time the Peel Inlet study was coming towards the end of the first phase of collection of field data including hydological data, information on the nuisance algae, data on nutrient sources, and surveillance data on the levels of nutrients in the inlet as well as algal biomass and phytoplankton levels, etc. The planning and resource allocation issue that was on the immediate horizon concerned identifying where the remaining data gaps might be, and conversely, in identifying those areas where enough was already known.

I review this history here because the Peel Inlet example contains most of the generic issues that one must deal with in the analysis of a large class of environmental problems and, of course, that investigation has conditioned how I continue to regard these issues. First and foremost is the role of the model as information integrator. The great attraction of the kind of simulation model based on the physics, chemistry and/or biology of the problem under investigation is that it can be used to integrate three quite different types of information. The first of these is the set of causal hypotheses that describe our current understanding of how different processes and variables are inter-related. For example, the Michaelis-Menton formulation of enzyme kinetics has often been used to describe the

uptake by algae of nutrients from water and their conversion into biomass. Hence, the structure of a simulation model using such elements is a synthesis of a set of mathematical descriptions of how the system is assumed to function and how one part depends on or influences the other.

Secondly, simulation models can incorporate existing information, often from the literature, on the range of values of parameters that may be relevant to the current application. This is because, for the most part, these parameters have a clear experimental interpretation. Indeed, many of the parameters in these models are exactly the parameters that are measured and reported in the literature of the various scientific specialties that underpin environmental modeling studies.

Finally, this class of simulation models can integrate what might be termed macroscopic data on the behavior or performance of the system under study. I think about this in the sense of state variables or that observable set of measures of the state of the system that exemplify the behavior that one is trying to understand or control. It is at this level that classical parameter estimation operates i.e. given a structure and a set of input-output data, find a set of parameter estimates which minimize some error criteria. In some contexts, this step is called "calibration."

2. Regional Sensitivity Analysis: a Different Approach to Calibration and Goodness-of Fit

In addressing ourselves to the Peel Inlet problem, Hornberger and I quickly came to understand that there was more information available to us in the first two of the above categories than in the third. That is, the literature contains a great deal of information on the causal relationships underlying algal proliferation and accumulation, nutrient limitations, light extinction coefficients and the myriad other factors and processes related to the problem under study. Similarly, the literature was rich in data directly or indirectly relating to the values of the parameters of models that were very similar in structure to those which appeared suitable for our purposes. Conversely, when one began to accumulate data collected from or directly related to the Peel Harvey system itself, they were a good deal more limited. Moreover, these data were very diverse in format. There was good time series data on tidal fluctuations in water level in the inlet, river hydrographs were available, and good data existed on incident solar radiation and meteorological variables. On nutrient levels in the water column or in the benthos, algal and phytoplankton biomass, and data on the growth characteristics of the particular alga, data was much more limited, both spatially and temporally [Hornberger and Spear, 1980].

It became clear that a good way to incorporate all of this information was first to develop one or more models based, in this case, on prior assumptions regarding the identity of the limiting nutrient. We called each of these a scenario and devoted our initial attention to phosphorous. We chose to use a lumped parameter model specific to the area we defined as the "growth area" which turned out to be comprised of 5 nonlinear, ordinary differential equations with 19 parameters. Because of the nature of the data from the literature on parameter values it was obvious that point estimates were not defensible and the uncertainly and variability in the parameter values was best described by statistical distribution functions. This was not a new idea. We also incorporated the data on solar radiation and similar time varying inputs in a straightforward way. The new idea came in the means we chose to incorporate the sparse field data relating to the seasonal variation in algal biomass, phytoplankton levels, and nutrient concentrations, those variables which defined the eutrophication problem. We felt that the spatial and temporal sampling patterns were simply not sufficiently dense to provide reliable data points to which we could curve-fit the model output in any defensible way. Rather, we discussed with various of the field scientists what it was about this system that characterized its problem behavior. This evolved into a set of six conditions on the state variables of the system that allowed us to classify any simulation, with parameters randomly drawn from the prior distributions, as mimicking the behavior of the system or not doing so. This classification was also assigned to the parameter vector which gave rise to it. This was our version of a generalized goodness-of-fit criterion to be used in site-specific calibration [Spear and Hornberger, 1980]. That is, we specified a model structure and associated parameters which pertained to a large class of eutrophication problems which were then made specific in the context of the data from a particular site.

After collecting a large number of vectors, each appropriately classified, we then carried out a posterior analysis which was focused on identifying the subset of parameters which appeared to be responsible for achieving good simulations. This procedure has come to be

called Regional Sensitivity Analysis (RSA). The power of the method arises from the classification notion. Visualize the situation geometrically by considering each element of the p-dimensional vector of allowable parameter values to be independently and uniformly distributed over the interval [0,1], that is, the prior parameter space can be defined as the unit hypercube without loss of generality. Any point within the hypercube can be identified as leading to a good or bad simulation by running the model with the corresponding parameter vector and applying the classification criteria to the output so produced. Hence, the model and the classification criterion provide a means of dividing the hypercube into two regions, one associated with good simulations and the other with bad. The information available from the approach is contained in the problem-specific interpretation of the observable features of these two regions.

In the Peel Inlet study and in most applications of the RSA concept, the posterior analysis has been confined to viewing the good/bad results from the perspective of the univariate marginal distributions. For example, if $F(x_i)$ is the prior distribution of the ith element of the parameter vector, x_i, then one asks if $F(x_i|G) = F(x_i|B)$, that is, do the conditional distributions show any difference under the good/bad mapping. If such difference can be discerned by an appropriate statistical test applied to the sample distribution functions, then evidence exists that x_i is an "important" or a "sensitive" parameter. However, it was realized from the outset that this index of sensitivity is a sufficient, but not necessary, condition for sensitivity. One can envision various types of parametric interactions that would not be observable from the univariate marginal distributions, a fact that motivated multivariable analysis as early as the Peel Inlet study. However, neither in that study nor since has conventional multivariable analysis of the parameter vectors been particularly revealing.

3. Uniqueness

An important difference between the Peel Inlet study and all of the other applications of RSA that our group has carried out, concerns the fraction of the total number of simulations that result in good outcomes or "passes". In the Peel Inlet case this fraction was about 45%. In all subsequent studies, it has been a to struggle to achieve numbers as high as 5%. The worst case we have encountered was 20 passes in 2.6 million simulations. In all of these subsequent cases there was some minor pruning of the range used in the very first runs, but for the most part these ad hoc adjustments, which were always based on the univariate marginal distributions, were very modest, e.g. 10% to 20% reductions in the range of one or two distributions out of 20 to 25. The interesting observation has been that, even before the pruning, the univariate marginal distributions of passing parameters extended over the entire allowable range for almost all parameters and, after the prune, across the entire range of all parameters. We have never seen evidence that the passing parameters occupied a single well-defined region internal to the hypercube which might arise, for example, from a multidimensional normal distribution centered at some interior point. However, we have made explorations of the connectedness of the pass region using nearest neighbor metrics of several sorts which suggest that the pass region is generally a single connected region. The general observation that "good" simulations can be found over almost the entire range of any single parameter clearly implies a lack of uniqueness. That is, there is not a single point in the parameter space associated with good simulations, indeed there generally is not even a well-defined region in the sense of a compact region interior to the prior parameter space. The two obvious responses to this observation are either to reiterate that this is the generic problem with big simulation models or, conversely, that the definition of a good simulation is too loose. We carried out a variety of studies in the late 80's that bear on the "goodness-of-fit" issue in the context of more traditional parameter estimation procedures. One provides a particularly straightforward illustration of the uniqueness issue.

In the field of environmental toxicology a problem of much contemporary interest concerns the relationship between external measures of human or animal exposure, like concentration of the chemical in the breathing zone, and the dose delivered to the internal site of toxic action, termed the receptor. It is common to address the issue of distribution and metabolism using what are called physiologically-based, pharmacokinetic models, PBPK models for short. These are sets of coupled ordinary differential equations, usually with a few nonlinear terms in which each state variable corresponds to the concentration of the chemical or a metabolite in some body region, e.g. "poorly perfused tissue" as shown in Figure 1. The attraction of the PBPK model is that the parameters correspond to physiologically meaningful quantities, for example, blood flows

to the various compartments like the liver or the bone marrow or blood-air partition coefficients for the chemicals in question. Hence, there exist sets of semi-standard parameters for various animal species and humans although within-species variability in these parameters is usually acknowledged, if seldom addressed.

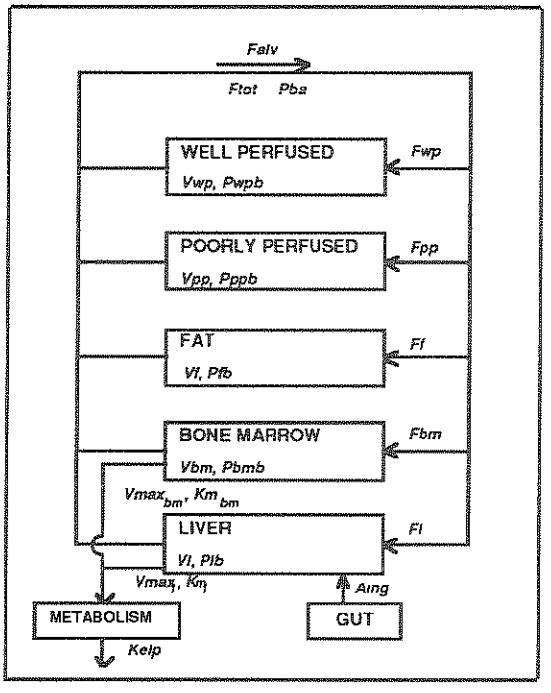

Figure 1: Schematic representation of the physiologic model used to simulate the distribution of benzene. V, volumes; F, flows; P, partition coefficients.

Because benzene is a chemical of much interest amongst my toxicological colleagues at Berkeley, our group developed and explored the usefulness of PBPK models to aid in the elucidation of the mechanisms by which benzene causes leukemia in humans. Following toxicological tradition, we began with a rat model because of the extensive data on distribution and metabolism in that species. Animal data of this sort usually relates to the concentration of benzene in each compartment at various times as a result of a pre-determined exposure pattern which may range from single injections to inhalation exposures over periods of hours to days. A small number of animals are sacrificed at each sampling period and tissue and blood levels measured. Hence, each data point is the average level in a small number of animals, typically three to five, and as such subject to significant variability. Typical data and some simulation results are shown in Figure 2.

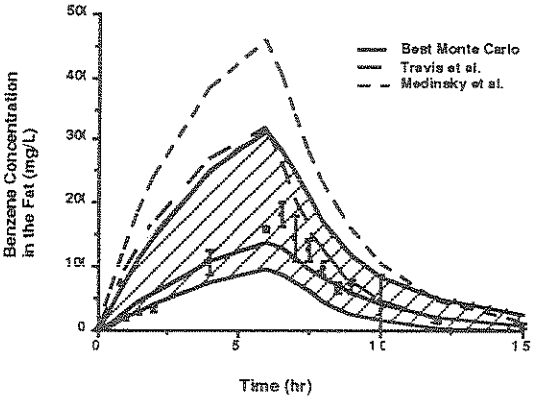

Figure 2: Benzene concentration in the fat of Fischer 344 rats during and after a 6 hour exposure to 490 ppm benzene in air.

Our benzene distribution model was comprised of 5 compartments and 24 parameters with several nonlinear elements [Bois, et al., 1991]. Two other groups had also developed PBPK benzene models of similar structure and with similar parameters [Medinsky et al., 1989, Travis et al., 1990]. However, their method of parameterization followed standard PBPK practice which involved fixing all parameters except those of the nonlinear elements governing metabolism in the liver which were varied until the best fit was achieved. (We were all using the same published experimental data sets on benzene distribution in Fischer rats.) In contrast, we followed the initial steps of the RSA procedure and developed biochemically and physiologically plausible ranges for each of the parameters. We then decided to contrast the fit obtained by the other investigators with what could be obtained from a Monte Carlo search over the ranges of all 24 parameters. A log-likelihood index of goodness-of-fit was used.

We ran 1000 Monte Carlo simulations of which 200 had better fits, as measured by the log-likelihood index, than those obtained using the parameter values of the other investigators. Any of these 200 Monte Carlo parameter sets yielded an acceptable fit to the experimental data by

conventional standards. Moreover, as we had seen in other applications, the 200 values leading to acceptable fits extended over the entire range of each of the parameters. For a number of parameters this simply implies a lack of sensitivity and for others, the presence of a strong covariance structure.

The issue of the covariance structure among the elements of parameter vectors leading to acceptable simulations is a very interesting issue. This is particularly true in light of the previous observation that it is usually difficult to obtain a high fraction of acceptable simulations. For example, in a recent application of RSA to mosquito population dynamics related to arboviral disease transmission, Eisenberg et al. [1995a] obtained only about 1% passes or acceptable simulations. Returning to the unit hypercube as the prior parameter space, we may interpret the fraction of good simulations to be the volume of the hypercube occupied by acceptable parameter vectors, V_g. Then, 1- V_g is a measure of the information gained between the prior and the posterior spaces, although it is not a measure independent of the dimension of the parameter space. A unique parameter set has a V_g of zero and complete information has been gained on the parameterization of the model.

4. Describing the Space of Acceptable Fits

The challenge, of course, is to describe the parameter space leading to acceptable fits to the output data. Whatever the criterion of acceptable simulations may be, our work over the years has shown the space of good parameter vectors to be very complex and not usefully described by traditional statistical methods. That is, we seem not to be dealing with hyper-ellipsoids or other macro-geometries of the sort which underlie principal components analysis or other standard multivariate statistical procedures. Our first exploration of new computer intensive methods was an application of the CART technique in our benzene work [Spear, et al., 1990]. CART is a non-analytic, computer intensive procedure which leads to classification rules based on inequality constraints applied to individual parameter values or to linear combinations of parameters. The issue in the benzene application was to see if calibration of the model separately to each of three different experimental data sets would be associated with the same region of the parameter space. If so, the model and the associated parameter space had captured the content of all of the data available to us in this form of meta-analysis.

The application of CART involved asking it to analyze the three sets of good parameter vectors, one set from each calibration experiment, and attempt to find rules that would allow one to discern which experiment a particular vector came from. CART found this to be a trivial challenge. Figure 3 shows the CART tree which resulted. On the basis of only three parameters it was able to discern which experiment a parameter set came from with very low misclassification error. One of these parameters was alveolar ventilation rate, which was not too surprising, in retrospect, since the two inhalation experiments used different methods of benzene administration and there was also a substantial altitude difference between the two laboratories at which the work was carried out.

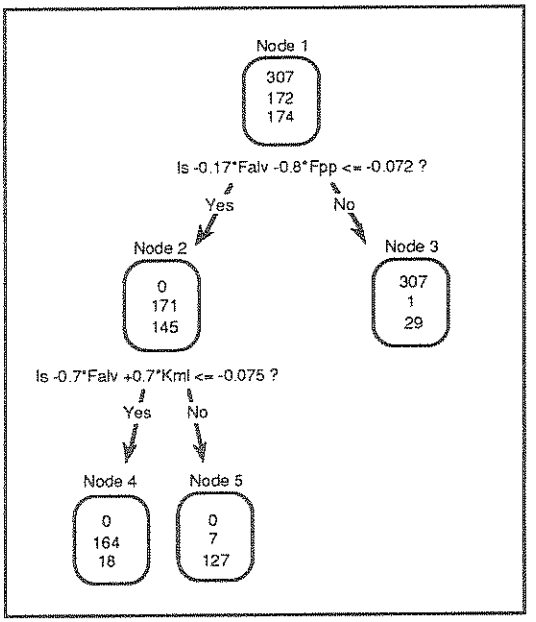

Figure 3: CART diagram from three experiments. Node 1 contains only passes, 307 from experiment 1, 172 from experiment 2, and 174 from experiment 3. F_{alv} is air flow to the alveolar spaces, F_{pp} blood flow to the poorly perfused tissue and K_{ml} the Michaelis-Menton parameter for liver metabolism.

We have recently completed another study using CART as the principal analytical tool [Eisenberg, et al., 1995b] This investigation

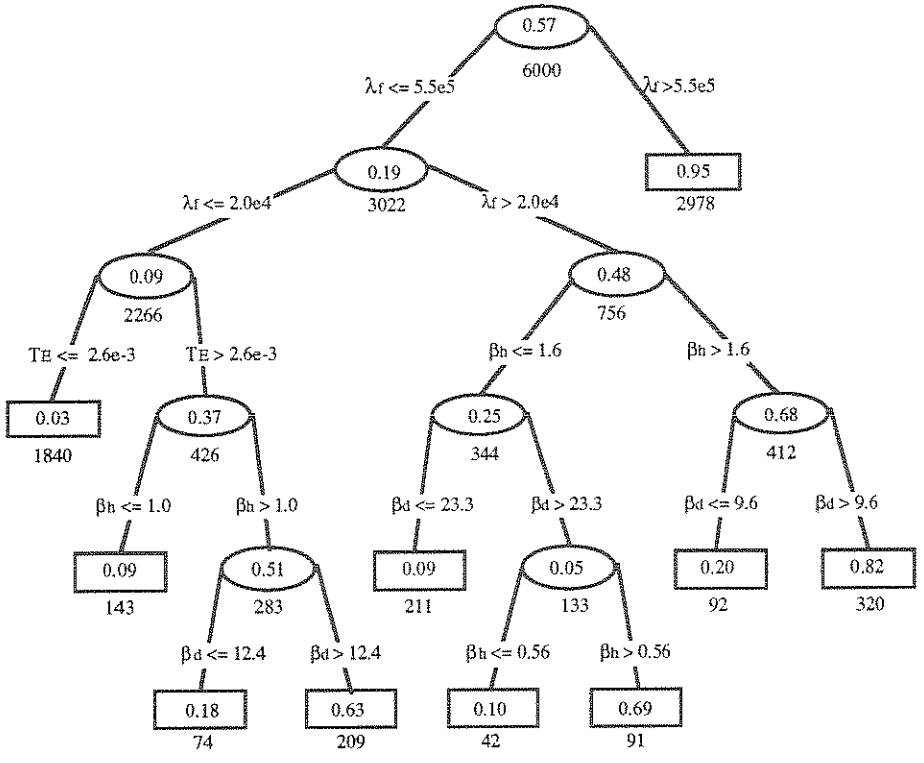

Figure 4: CART diagram for the giardia analysis. The number below the oval is the number of parameter vectors routed to that node. The number within the oval is the proportion of these vectors associated with ourbreak conditions.

involved characterizing the risk of infection from enteric pathogens associated with swimming in impoundments which receive treated municipal wastewater. The model is epidemiological in structure and is comprised of sub-populations of susceptible, immune, infected, and clinically ill individuals. In this case, calibration utilized epidemiological data on the occurrence of giardiasis under background or non-outbreak conditions where waterborne exposure was not present. Parameter vectors consistent with the background constraints were then used in a full simulation with waterborne exposure. The task presented to CART was to determine the parameters which were principally responsible for outbreaks of the disease under conditions of waterborne exposure. Figure 4 shows the resulting CART tree. Here λ_f is a parameter associated with the shedding of cysts by infected asymptomatic swimmers, β_h and β_d parameters related to the intensity of water contact, and T_E a treatment-related parameter. The outcome of the analysis indicated that the risk of infection was dominated by the direct exposure of swimmers to pathogens shed by other swimmers and that the water reclamation pathway was of concern in only very unusual conditions. For present purposes, however, the CART applications provided a glimpse of the potential power of recent advances in computer-based multivariate analysis.

5. Tree-Structured Density Estimation

In the context of the RSA procedure, Beck [1987] pointed out in his encyclopedic review of uncertainty in water quality modeling that a disadvantage of the approach was that" the interpretation of the derived *a posteriori* parameter distributions becomes more difficult as

the dimension of the parameter vector increases, and for all practical purposes, it seems probable that any conclusions will have to be restricted to the properties of the univariate and bivariate marginal distributions." This has certainly been true in all of the RSA work conducted by our group until the CART experience led us to invest more heavily in acquiring greater expertise in the area of computer-based multivariate analysis. The methodological innovations I will now describe are due to Dr. Nong Shang whose doctoral work involved CART and related methodologies under one of its originators, Leo Breiman at Berkeley.

Shang has argued that that the challenge of describing the pass region beyond what can be observed from the univariate marginal distributions is best approached as a problem in multivariate density estimation. The underlying concept of his method involves the extension of the concept of the variable length histogram to multiple dimensions [Shang, 1993]. That is, the complex density of passes in the parameter space is approximated by uniform densities in local regions, just as the histogram with variable length does in a single dimension. The idea is based on the following philosophy: if the density is uniformly distributed in a local region (including the extreme cases where the density is equal to 0 to 1), then no further information about the parameter and parameter interactions can be extracted from the local region. The first challenge, however, is to find the local regions, determine their extent, and arrange them in some systematic way.

Our first applications of this methodology concerned the groundwater pathway of a multimedia fate, transport and exposure model called MMSOILS developed by the U.S. EPA [Spear, et al., 1994]. The issue was to determine which parameters controlled predictions of unacceptable levels of benzene in a drinking water well 75 years in the future due to current

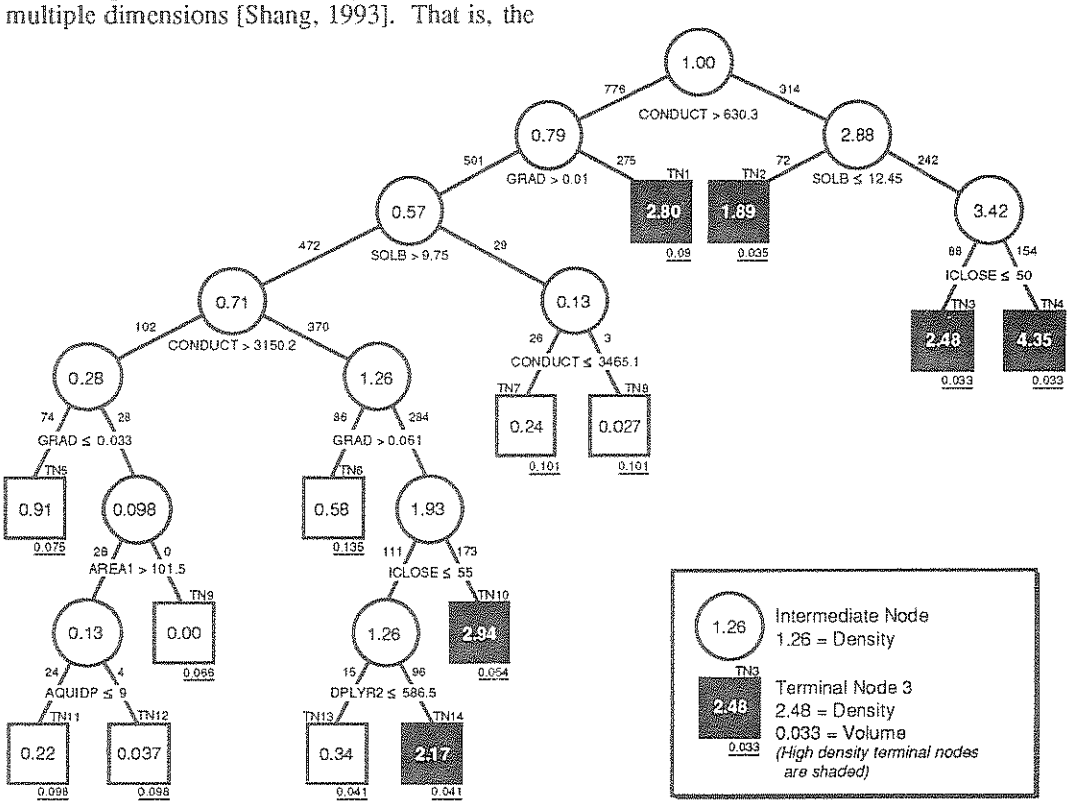

Figure 5 shows a tree from the MMSOILS study. The number in the circle is the density of points in the space normalized to 1.00 which corresponds to the actual density determined by the total number of points divided by the volume of the original space. The splitting condition is shown under each circle and the number on each line is the number of simulations sent to the next node. The squares are the terminal nodes which are the subspaces where the algorithm has decided that the points within it appear homogeneous and that it does not have sufficient resolution to proceed further.

site contamination. This study was more an exploration of the method than the model, and it was successful in that sensible results were obtained and our experience showed that theresults are amenable to interpretation in practice. Of particular note is that the trees that describe the results can be used to direct subsequent investigations of interesting parts of the parameter space.

To explore the directed search notion, we carried out a further set of simulations targeted on terminal node 5 on the extreme left of the tree because that node has a relative density of 0.91, almost equal to that of the original space. As intuition would suggest, a greater number of sample points in this subspace allowed the algorithm to discern finer structure and several new parameters appeared as important for the first time.

6. Conclusions and Recommendations

Although much work remains to be done to make the tree-structured density estimation procedure a generally available tool, it appears that it does have the properties necessary to describe the pass region in a practically useful way. While its application in Monte Carlo analyses grew out investigations using the RSA procedure, I suggest that the perspective on uniqueness that it offers is more generally applicable. In dealing with the parameterization of large simulation models, and by large I mean with roughly 10 parameters or more, it seems very likely that for almost any reasonable index of goodness-of-fit, discrete or continuous, there will be many parameter sets that give rise to fits that are practically indistinguishable. Further, many of these "good" parameter sets, judged to be so on the basis of fit to the calibration data, will extend over the full range of plausible values of each of its individual elements. To the extent that this contention is true, there are clear limitations on how one might interpret the technical or scientific significance of any particular set of parameters that lead to a good fit. For large simulation models, it appears that we must alter our traditional view and think of the "best" parameter estimate as an extended and complex region in a high dimensional space.

7. References

Beck, M.B., Water quality modeling: A review of the analysis of uncertainty, *Water Resour. Res.*, 23(8), 1393-1442. 1987.

Bois, F.Y., T.J. Woodruff and R.C. Spear, Comparison of three physiologically based pharmacokinetic models of benzene disposition, *Toxicol. Appl. Pharmacol.* 110 , 79-88 , 1991.

Eisenberg, J.N., W.K. Reisen and R.C. Spear, Sensitivity analysis of a dynamic model comparing the bionomics of two isolated *Culex tarsalis* (Diptera: Culicidae) populations, *J. Med. Entomol.*, 32, 98-106, 1995a.

Eisenberg, J.N., E.Y.W. Seto, A.W. Olivieri and R.C. Spear, Quantifying water pathogen risk in an epidemiological framework, in review, 1995b

Hornberger, G.M. and R. C. Spear, Eutrophication in Peel Inlet, I, The problem defining behavior and a mathematical model for the phosphorus scenario, *Water Res.*, 14, 29-42, 1980.

Medinsky, M.A., P.J. Sabourin, G. Lucier, L.S. Birnbaum, and R.F. Henderson, A physiological model for simulation of benzene metabolism by rats and mice, *Toxicol. Appl. Pharmacol.* 99, 193-206,1989b.

Shang, N., New developments in tree-structured methodology, Ph.D. dissertation, Univ. of Calif., Berkeley, 1993.

Spear, R.C., F.Y. Bois, T.Woodruff, D. Auslander, J. Parker, and S.Selvin, Modeling benzene pharmaco-kinetics across three sets of animal data: Parametric sensitivity and risk implications, *Risk Anal.*, 11, 641-654, 1991

Spear, R.C., T.M. Grieb, and N. Shang, Parameter uncertainty and interaction in complex environmental models, *Water Resour. Res.*, 30(11), 3159-3169, 1994

Spear, R.C. and G.M. Hornberger, Eutrophication in Peel Inlet, II, Identification of critical uncertainties via generalized sensitivity analysis, *Water Res.*, 14, 43-49, 1980.

Travis, C.C., J.L. Quillen and A.D. Arms, Pharmacokinetics of benzene, *Toxicol. Appl. Pharmacol.* 102, 400-420, 1990.