Application of a Variance-Based Importance Analysis to a Complex Probabilistic Performance Assessment

Abbreviated Title:

App Variance Importance

by

Randall D. Manteufel

Senior Research Engineer

Center for Nuclear Waste Regulatory Analyses

Southwest Research Institute

6220 Culebra Road

San Antonio, TX 78238-5166

210/522-5250

Abstract

The most important input parameters in a complex probabilistic performance assessment are identified using a variance-based method and compared with those identified using a regression-based method. The variance-based method has the advantage of not requiring assumptions about the functional relationship between input and output parameters. However, it has the drawback of requiring heuristic assessments of threshold variance ratios above which a parameter is considered important, and it also requires numerous executions of the computer program, which may be computationally expensive. Both methods identified the same top 5 and 7 of the top 10 most important parameters for a system having 195 inputs. Although no distinct advantage for the variance-based approach was identified, the ideas which motivate the new approach are sound and suggest new avenues for exploring the relationships between the inputs and the output of a complex system.

Key Words

Sensitivity/uncertainty analysis, importance analysis, probabilistic performance assessment.

Introduction

A probabilistic risk or performance assessment is often used to evaluate complex systems such as a nuclear power plant during accident scenarios[1] or a nuclear high-level waste repository over long time periods[2,3]. A computer program is used to predict the response of the system, y, is dictated by the input parameters, $y=y(x_1, x_2, \ldots x_n)$. Each input parameter has some uncertainty or range of plausible values which can be described using a probability distribution function (PDF). The PDF for each input can be sampled to obtain discrete values for all inputs, and the output can then be calculated using the computer program. Typically, this process is performed in a probabilistic manner in which the x's are selected randomly from their PDFs and used to generate a set of y predictions.

For complex systems, it is rarely obvious which parameters are most important. In many cases a complex system may require hundreds of input parameters, yet only a relatively small subset strongly influences the output. It is often the analyst's objective to identify these. There are many reasons for seeking the most important parameters, including comparing results between programs modeling the same problem, guiding the design of the system being modeled to reduce the vulnerability to particular phenomena, and guiding future work in specific areas to better quantify important yet poorly characterized phenomena.

A distinction is made in this work between uncertainty, sensitivity, and importance. The uncertainty is the variability or range of an input or output parameter. The sensitivity is the rate of change in the output as an input changes. Importance is the degree to which the output changes in response to the uncertainty in one of the inputs. It is helpful to associate input

uncertainty with a range in the parameter (e.g., Δx_i for the i^{th} variable), and sensitivity with the change in the output per change in an input (e.g., $\partial y/\partial x_i$). The product of output sensitivity and input uncertainty is a measure of output uncertainty attributed to a particular input parameter (e.g., $\Delta y_i = (\partial y/\partial x_i)\,\Delta x_i$). Importance is a relative term where the output uncertainty attributed to each input parameter is compared with the overall range of the output (e.g., $\Delta y/\Delta y_i$) For an input parameter to be important, the output needs to be sensitive to the input and the input parameter needs to have some uncertainty (or range of plausible values if it is a design controlled parameter).

Overall, there are two main approaches to identifying important parameters—regression- and variance-based methods. The regression-based methods have been used extensively throughout the literature[4,5,6] while variance-based methods are relatively new[7]. In general, the regression methods are based on establishing functional relationships between the input and output parameters. Frequently, linear relationships are sought, yet nonlinear and coupled relationships can also be used. For linear relationships, the slope between each input and the output is used to gauge the importance of each input parameter. The slope can be multiplied by the range, or standard deviation, of the input and divided by the overall output variability, or standard deviation. This ratio equals the standardized regression coefficient (SRC)[4]. An input parameter is important if it has a large SRC, hence, is responsible for a large variability of the output.

One drawback of regression-based methods is that an algebraic function relating inputs and output must be developed. This approach is very successful when couplings and interactions

between inputs are minimal, and the inputs are independently and linearly related to the output. As interactions become stronger, it becomes more critical to specify an appropriate functional form relating inputs and output. In an attempt to relax the need to specify the form of the relationship, McKay[7] has developed a variance-based importance method which does not require assumptions about the type of relationship between input and output. In addition, it is integrated with the Latin Hypercube Sampling (LHS) scheme[8,9] and readily quantifies the effects of inputs on the output. A potential drawback of the variance-based method, however, is that it requires numerous executions of the computer program.

Application Problem

Both the variance-based and regression-based methods are evaluated using a code that simulates the release and transport of radionuclides from a proposed high-level waste (HLW) repository at Yucca Mountain, Nevada[3,10]. The code uses the LHS scheme with a total of 195 input parameters which are described using PDFs. The output is the summed normalized release of radionuclides over the next 10,000 yr across a compliance boundary at 5 km. The details of the code are not of primary interest in this paper and are documented elsewhere[3]. The primary objective of this study is to apply the variance-based method and compare the results with regression-based results.

Introduction to Variance-Based Method

The main objective of the variance-based approach is to identify a subset of input parameters which most strongly drive the output. In the top of Figure 1, the spread of the output is characterized using either the PDF, Cumulative Distribution Function (CDF), or

Complementary Cumulative Distribution Function (CCDF). The CCDF is used frequently in this paper. As the range of the output narrows, the slopes of the CDF and CCDF steepen, and this will be used later to gauge the importance of a parameter. The idea behind variance-based importance is to identify the parameters that, when held constant, significantly reduce the spread of the output or steepen the CDF/CCDF curves. In the bottom of Figure 1, a hypothetical distribution of outputs is shown with smaller variance, hence narrower PDF and steeper CDF and CCDF. If the variability of an important parameter is reduced, then the variability of the output will be significantly reduced. If a parameter is identified as being important based on output variability, then it follows that significant changes in its mean value will lead to significant changes in the mean of the output. For analysis purposes, it is convenient to focus on output variability, noting that important parameters strongly affect both the mean and variance of the output.

Scatter Plots

Scatter plots are a simple and informative tool to investigate visually the relationship between any one input parameter and the output. Typically, the analysis will have generated a large set of valid inputs which are used to compute an equal number of outputs. A plot is constructed of the output y versus a single input parameter x_i. Each model evaluation or computer code run is represented as a single point. Depending on the distribution of points, one may identify a relationship between the input parameter and the output. If no pattern exists, then the scatter plot will appear as a cloud of uncorrelated points.

In Figure 2, two scatter plots are shown from this work. The two parameters, infil and

akr2, are plotted against the output, normalized release. Each scatter plot contains 2,000 points representing 40 distinct LHS-50 runs. The term LHS-50 describes how each input parameter is discretized into 50 equal probability bins, and the mean value in each bin is used once and only once in a LHS run. In total, 50 executions of the computer program (runs) are performed for an LHS-50, regardless of the number of input parameters. Each run is based on a vector of inputs (equal in length to the number of input parameters) determined by randomly matching binned input parameters. In this work, 40 different matchings of input parameters were used, hence, 40 distinct LHS-50 runs. As a result, the scatter plots consist of 40 points along 50 vertical lines. The first parameter, infil, describes the deep percolation of infiltrating meteoric water at the repository site. The second parameter, akr2, describes the gaseous fracture permeability of one of the hydrostratigraphic units at Yucca Mountain. The phenomenon being described by each parameter is not important for this paper, only that infil has a visual correlation with normalized release while akr2 does not. One also notices that the normalized release is plotted on a log scale because it ranges over 5 orders-of-magnitude. A few of the outputs had zero value, and are set equal to 2×10^{-4} in order to be plotted.

Because the output values ranged over 5 orders-of-magnitude and contained a few zero values, the output was transformed to perform either a variance- or regression-based analysis[11]. A linear rank transformation was used in which the outputs from all of the runs are ordered and replaced by a set of steadily increasing values. In this problem, the 40 LHS-50 generated 2,000 outputs, and the lowest output was transformed to 1/2,000, the second lowest to 2/2,000, and so forth up to the highest output being transformed to 1. This generated a uniform distribution in

the output values of y.

In Figure 3(a), the scatter plot for the rank transformed output and infil are shown. The rank transformed data continue to show visually the trend in the data. The mean of the 40 values for each of the 50 bins is shown in Figure 3(b), as well as the standard deviations about the means in (c). The bin means show much more clearly the correlation between input and output. Figure 3 is used to introduce the variance-based method.

Variance-Based Importance

The main goal behind the variance-based method is to determine what portion of the total output variance is explained by a trend through the means or is unexplained due to residual uncertainty attributed to other parameters. This approach is consistent with the idea of correlation and regression. A well-known variance identity used in analysis of variance relates the total variability of y to the variability between the bin means (explained) and variability within the bins (unexplained)[7,12,13,14,15]. For this work, the identity is expressed as:

$$\sum_{j=1}^{Nrep}\sum_{i=1}^{Nvec}\left(y_{ij}-\bar{y}_{..}\right)^2 = Nrep\sum_{i=1}^{Nvec}\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2 + \sum_{i=1}^{Nvec}\sum_{j=1}^{Nrep}\left(y_{ij}-\bar{y}_{i.}\right)^2 \qquad (1)$$

where Nvec=number of LHS vectors or bins (=50 in this work), and Nrep=number of repetitions (=40 in this work). In words, Equation 1 states that the total variability equals the variability of bin means plus the mean of bin variabilities.

The average of all output y's is used in Equation 1 and defined as:

$$\bar{y}_{..} = \frac{\sum_{j=1}^{Nrep} \sum_{i=1}^{Nvec} y_{ij}}{Nrep \; Nvec} \tag{2}$$

The average of all y's in one LHS bin is used in Equation 1 and defined as:

$$\bar{y}_{i.} = \frac{\sum_{j=1}^{Nrep} y_{ij}}{Nrep} \tag{3}$$

This variance identity is used to develop importance indices.

Importance Indices: R^2, R_a^2, F

If the input parameter has a negligible influence on the output, then the variability of bin means will be relatively small and the mean bin variability will be relatively large. Conversely, if the input has a strong influence on the output, then the variability of bin means will be relatively large. This suggests an importance index which is the ratio of variances:

$$R^2 = \frac{Nrep \sum_{i=1}^{Nvec} \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2}{\sum_{j=1}^{Nrep} \sum_{i=1}^{Nvec} \left(y_{ij} - \bar{y}_{..}\right)^2} \tag{4}$$

This ratio is analogous to the coefficient of determination which is commonly denoted as R^2, except that Equation 4 is derived from an analysis of variance instead of a continuous regression-based estimate. The theoretical underpinnings of an analysis of variance can be related to a regression analysis[16]; however, it is preferable to distinguish between variance- and regression-based methods. The primary difference is that the variance-based method is indifferent

to the ordering of bins, whereas the ordering is important for a regression (or curve-fitting) approach. In Figure 3(b), a trend is noticeable between the input and output, hence the variability of bin means will be large. As the trend becomes more pronounced, then R^2 will increase in magnitude, regardless of the shape of the trend.

If no trend exists, then the bin means will uniformly lie about a mean of one-half. A result of the central limit theorem in statistics[13], is that the variance of bin means is equal to the total variance divided by Nrep, provided each bin contains random samples from the same population (i.e., no bin-to-bin variation for either the mean or variance). The explained variance is slightly over-estimated because it is based on a finite number of repetitions. Hence, an improved importance index, expressed as an alternative variance ratio has been suggested[7]

$$R_a^2 = \frac{\text{Nrep} \sum_{i=1}^{\text{Nvec}} \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 - \frac{\sum_{i=1}^{\text{Nvec}} \sum_{j=1}^{\text{Nrep}} \left(y_{ij} - \bar{y}_{i.}\right)^2}{\text{Nrep}}}{\sum_{j=1}^{\text{Nrep}} \sum_{i=1}^{\text{Nvec}} \left(y_{ij} - \bar{y}_{..}\right)^2} \tag{5}$$

Finally, the F-statistic can be used as an importance index[14,15]. The 50 LHS bins act as distinct levels within which 40 samples are collected. The analysis determines if there is a statistically significant difference between the means of the distributions in each of the bins. In statistical terms, a null hypothesis is formed stating that the bin means are equal. If the null hypothesis is true, then the input has a negligible effect on the output. Alternatively, data may indicate that all of the bin means are not equal, hence, the null hypothesis is rejected and the input is identified as being important. The F-statistic for this application is defined as:

$$F = \frac{\dfrac{\text{Nrep}\sum_{i=1}^{\text{Nvec}}\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2}{(\text{Nvec}-1)}}{\dfrac{\sum_{i=1}^{\text{Nvec}}\sum_{j=1}^{\text{Nrep}}\left(y_{ij}-\bar{y}_{i.}\right)^2}{\text{Nvec}\,(\text{Nrep}-1)}} \tag{6}$$

Both the numerator and denominator of the F-statistic are estimates of the total variance of y, assuming no trend exists between input and output. If a trend exists, then the numerator will overestimate the variance. Hence, the magnitude of F will increase.

Comparing Importance Indices

The first step was to complete a set of LHS-50 runs. The number of LHS-50 runs needed to provide good statistics was determined by increasing the total number of runs from 8 to 16, 24, 32, and finally 40. As more runs were completed, the importance indices were computed using all of the available data for all of the 195 input parameters. It was determined that enough data was collected from 40 runs to identify statistically important parameters.

In Figure 4, the importance indices are compared for the top ten parameters. In each of the three plots, the abscissa represents the quantity of output information and the ordinate represents the value of the importance index for various input parameters. The importance of a parameter is judged by a large value of the importance index, typically one exceeding a threshold. As the amount of output information increases, one expects the magnitude of the importance index to increase in exceedance of a cutoff, thereby indicating increased statistical confidence that a parameter is truly important. Only one importance index has this property, the F-statistic.

Both the R^2 and R_a^2 decrease with increasing output information, which is not a desirable attribute of an importance index. Hence, the F-statistic was used as the importance index in this work.

A heuristic F-statistic cutoff value of 3.0 was adopted in this work primarily because a natural break was observed in the first set of 40 LHS-50 runs with 5 variables performing noticeably different than the other 190 variables. Heuristic selection of a cutoff value is suggested by McKay[7]. In the traditional one-way analysis of variance[13,14,15], a theoretically derivable cutoff for a 1 or 5 percent level of significance can be derived (in this work it would be based on Nvec-1 and Nvec(Nrep-1) degrees of freedom). Embedded in the significance levels associated with the F-statistic are assumptions that all measurements are independent, random samples, drawn from normal distributions that have equal variances. The assumption of independent random samples is not true for this work. For a independent random sampling, sometimes called pure Monte Carlo, the y's are independent. However, for an LHS scheme, stratified sampling without replacement, the set of x's are not independent; hence, the y's are not completely independent (see Appendix of Reference No. 8). Because we are using the LHS scheme, the data points are not completely independent and significance levels normally associated with the F-statistic do not apply.

Top Ten Parameters

Based on the initial 40 LHS-50 runs, five parameters were identified as being important. These parameters are the top five identified in Figure 4(c). In Figure 5, rank ordered plots show the F-statistic computed for each parameter. The parameters are sorted by F-statistic so that the

first has the largest F. The F-statistic is plotted against the sorted (or ranked) order. These plots help identify individual or groups of important parameters. In Figure 5(a), the infil parameter is clearly important, as noted by the large F-statistic which confirms the trend observed in the scatter plots in Figure 3.

After identifying at least one important parameter, an iterative approach is used to identify additional parameters. The previously identified parameters are set to fixed values. In general, they can be set to more than one value, but a new set of runs would need to be completed for each unique set. Because our program requires a long time to complete calculations, it was prohibitive to explore more than the mean of the previously selected parameters.

In Figure 5(b), the results of the second set of runs indicated that three more parameters are important. For the last set of runs, eight parameters were set to their mean values and a new set of runs initiated. Based on our experience, if the F-statistic of a parameter exceeded a cutoff, then as Nrep increased it would only continue to increase in exceedance of the cutoff [see Figure 4(c)]. Hence, the last set of runs was terminated as soon as two parameters exceeded the cutoff so that ten parameters were identified as important. These parameters are: infil, forwar1, ecorr6, ecorr7, rdiff11, ecorr3, ecorr2, ecorr8, ecorr5, sol4Am. These parameters are described elsewhere[3] and only briefly described here: infil = infiltration rate, forwar1=UO_2 alteration rate , ecorr6=crevice corrosion potential, ecorr7=pitting corrosion potential, rdiff11=diffusion coefficient in the nearfield, ecorr3=temperature effect on ambient corrosion potential, ecorr2=temperature effect in corrosion model, ecorr8=rate of localized corrosion, ecorr5=decay rate for gamma emitters, and sol4Am=Americium solubility.

In Figure 6, conditional CCDFs are plotted where the output has been conditioned by holding either 0, 5, or 10 input parameters fixed. Each curve represents the output from an equivalent LHS-400 run. As expected, Figure 6 shows that the variance of the output is reduced as the important parameters are fixed.

The reduction in output variability can also be quantified. For each set of five equivalent LHS-400 curves, a single equivalent LHS-2000 curve was constructed. Based on this curve, the output value was determined for 95 percent of the distribution. Thus the range of output was measured between the 2.5 and 97.5 percent probabilities. For the case of no fixed parameters, the CCDF ranged from 0.02 to 18.0 which is a factor of 890. By selecting and fixing the top 5 parameters, the range was decreased so that the output varied from 0.27 to 5.94 which is a factor of 22.0. Fixing the top ten parameters yielded an output range from 0.29 to 1.86 which is a factor of 6.5. By fixing the ten most important parameters to their mean values, the output variability was reduced by over two orders-of-magnitude.

Comparison with Regression Results

A multilinear regression analysis was completed using the original rank transformed 40 LHS-50 runs with no parameters fixed. A stepwise procedure was employed where each parameter was independently regressed with the output data[17]. The parameter with the largest coefficient of determination, R^2, was selected as an important parameter. The next step added one previously unimportant parameter and regressed. The set of parameters with the largest R^2 is selected as the new set of important parameters. The p-value for each regression coefficient is checked to determine if it is above a 5 percent threshold. The p-value is the probability of

observing a nonzero regression coefficient from the finite sample when the true (population) coefficient is actually zero. If the p-value exceeds a significance level of 5 percent, then the parameter was excluded from the important set. In each step, one parameter is added to the set of important parameters. The process is stopped for one of a number of reasons: (i) no additional statistically significant parameters can be identified, or (ii) the addition of parameters yields a minimal improvement in R^2, thus indicating an overfitting of the output data. In this work, the stepwise addition of parameters was stopped when the overall R^2 changed by less that 0.01.

In Table I, the results of the regression analysis are presented. Fortuitously, the regression analysis terminated with ten parameters due to small incremental improvements in R^2 with the addition of new parameters. The regression results are very similar to the variance-based results. The top five parameters in both methods are the same, and seven of the parameters are the same in both sets. This agreement between sets of important parameters significantly increases confidence in both methods.

A comparison was made to determine why the variance- and regression-based methods differed in three parameters. It should be noted that the top five parameters control a significant amount of output variability and that other parameters are progressively less important. Scatter plots for the three parameters selected by the regression, yet missed by the variance-based method, are shown in Figure 7. A trend in any of the scatter plots is difficult to detect. One can detect a trend in the akr3 scatter plot [Figure 7(a)]. After further investigation, it was concluded that a combination of low akr3 and low infil leads to low outputs.

It is interesting that akr3 was the next most significant parameter identified in the

variance-based method after the first set of runs [see Figure 5(a)]. If the cutoff value for the F-statistic were lower, then akr3 would have been selected. In comparison, ecorr8 was also nearly selected to be important in the variance-based method in the initial set of 40 LHS-50 runs. Although not selected in the first set of runs, ecorr8 was selected in the second set of runs. An explanation for why akr3 was not selected is that it is important only at low values of infil. In the second step of the variance-based method, infil was set to its mean value. Thus, the importance of akr3 was diminished in subsequent runs. Other than this observation, the distinction between the parameters selected by the regression- and variance-based methods appears minimal because both methods selected the same top five parameters.

Verification Runs

In Figures 12 and 13, conditional CCDFs are shown which serve as a check on selected important parameters. The verification runs are performed to determine if the set of ten parameters truly control the output. If the most important parameters control the output, then one would expect that fixing them to high or low values would strongly affect the location of the outputs yet the range of outputs remains narrow for any single run. Alternatively, fixing the less important parameters to different values would not be expected to affect significantly the range or distribution of the output.

In Figure 8(a), a coarse LHS-5 was applied to the most important parameters and a fine LHS-400 was applied to the less important parameters. The LHS-400 was accomplished by combining eight independent LHS-50 runs. The conditional CCDFs (dashed lines) are compared with the original CCDF (solid line). The original CCDF is a based on a combination of the original

40 LHS-50 runs. We note that the conditional CCDFs are rather steep, indicating a narrow range of output values. This is because only the less important parameters are being varied. The locations of the conditional CCDFs are dictated by the specific values and combinations of the most important parameters. The most important parameters are noted to significantly affect the location where the conditional CCDFs break.

In Figure 8(b), similar calculations are presented. Here, the important parameters are varied while the less important parameters are fixed. A fine LHS-400 is used for the most important parameters and a coarse LHS-5 is used for the less important parameters. The conditional CCDFs lie near the original CCDF, indicating that the range and distribution of the output is being controlled by the most important parameters. One of the conditional CCDFs, however, does vary significantly from the cluster of other curves. The main deviation is due to a number of very low outputs, where 15 percent of the outputs had values smaller than 0.01. After reviewing the input, we attributed this to the parameter akr3 being small. Because our interests are more for higher values of the output, this effect at the low values was not explored further. Overall, the conditional CCDFs enhance confidence that the dominant parameters were identified in the variance-based method.

Conclusions

Both a variance- and regression-based method were applied to identify ten important parameters from a total of 195 input parameters for a complex system. Both methods agreed on five of the top five, and seven of the top ten parameters. After reviewing scatter plots, an explanation was developed for why the variance-based method missed one apparently important

parameter. This was attributed more to the application of the method than to a deficiency of the method.

The variance-based method has the potential benefit of not requiring any assumptions about the functional relationship between input parameters and the output. Two significant drawbacks of the method are the need to heuristically select critical (cutoff) values of an importance measure and the potentially prohibitive costs associated with the repetitive execution of the computer program.

The choice between variance- and regression-based methods is most probably influenced by the cost associated with computing time. The variance-based method requires many more computer runs than a regression-based method, hence this may prohibit its use in certain cases. Overall, the ideas which motivate the variance-based approach are sound and suggest new avenues for exploring the relationship between the input parameters and output for complex systems.

Acknowledgments

This paper was prepared to document work performed by the Center for Nuclear Waste Regulatory Analyses (CNWRA) for the Nuclear Regulatory Commission (NRC) under Contract No. NRC-02-93-005. The activities reported here were performed on behalf of the NRC Office of Nuclear Regulatory Research, Division of Regulatory Applications. The report is an independent product of the CNWRA and does not necessarily reflect the views or regulatory position of the NRC.

References

1. Nuclear Regulatory Commission. 1990. *Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants*. NUREG-1150. Washington, DC: Nuclear Regulatory Commission.

2. Helton, J.C., J.W. Garner, R.D. McCurley, and D.K. Rudeen. 1991. *Sensitivity Analysis Techniques and Results for Performance Assessment at the Waste Isolation Pilot Plant*. SAND90-7103. Albuquerque, NM: Sandia National Laboratories.

3. Nuclear Regulatory Commission. 1995. *Phase 2 Demonstration of the NRC's Capability to Conduct a Performance Assessment for a High-Level Waste Repository*. NUREG-1464. Washington, DC: Nuclear Regulatory Commission.

4. Iman, R.L., and J.C. Helton. 1985. *A Comparison of Uncertainty and Sensitivity Analysis Techniques for Computer Models*. NUREG/CR-3904. Washington, DC: Nuclear Regulatory Commission.

5. Iman, R.L., and J.C. Helton. 1991. The repeatability of uncertainty and sensitivity analyses for complex probabilistic risk assessments. *Risk Analysis* 21(4): 591–606.

6. Wu, Y.T., A.G. Journel, L.R. Abramson, and P.K. Nair. 1991. *Uncertainty Evaluation Methods for Waste Package Performance Assessment*. NUREG/CR-5639. Washington, DC: Nuclear Regulatory Commission.

7. McKay, M.D. 1995. *Evaluating Prediction Uncertainty*. NUREG/CR-6311. Washington, DC: Nuclear Regulatory Commission.

8. McKay, M.D., W.J. Conover, and R.J. Beckman. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21: 239–245.

9. Iman, R.L., and M.J. Shortencarier. 1984. *A FORTRAN 77 Program and User's Guide for the Generation of Latin Hypercube and Random Samples for Use with Computer Models*. NUREG/CR-3624. Washington, DC: Nuclear Regulatory Commission.

10. Sagar, B., and R.W. Janetzke. 1993. *Total-System Performance Assessment (TPA) Computer Code: Description of Executive Module, Version 2.0*. CNWRA 93-017. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

11. Iman, R.L., and W.J. Conover. 1979. The use of rank transform in regression. *Technometrics* 21: 499–509.

12. Draper, N.R., and H. Smith. 1966. *Applied Regression Analysis*. New York, NY: John Wiley and Sons.

13. Mendenhall, W., R.L. Scheaffer, and D.D. Wackerly. 1981. *Mathematical Statistics with Applications*. Boston, MA: Duxbury Press.

14. Dunn, O.J., and V.A. Clark. 1987. *Applied Statistics: Analysis of Variance and Regression*. New York, NY: John Wiley and Sons.

15. Rice, J.A. 1988. *Mathematical Statistics and Data Analysis*. Pacific Grove, CA: Wadsworth & Brooks.

16. Montgomery, D.C. 1991. *Design and Analysis of Experiments*. New York, NY: John Wiley & Sons.

17. Iman, R.L., J.M. Davenport, E.L. Frost, M.J. Shortencarier. 1980. *Stepwise Regression with PRESS and Rank Regression*. SAND79-1472. Albuquerque, NM: Sandia National Laboratories.

Table I. Stepwise multilinear regression analysis results using the original 40 LHS-50 runs.

Parameter	SRC	R^2
infil	0.527	0.28
forwar1	0.283	0.35
rdiff11	0.260	0.42
ecorr7	-0.239	0.48
ecorr6	-0.236	0.54
akr3	0.173	0.57
ecorr8	0.150	0.59
retard3	-0.124	0.61
retard1	-0.111	0.62
ecorr2	0.098	0.63

SRC = standardized regression coefficient based on final stepwise multilinear regression.

R^2 = coefficient of determination based on regression with all previous parameters.

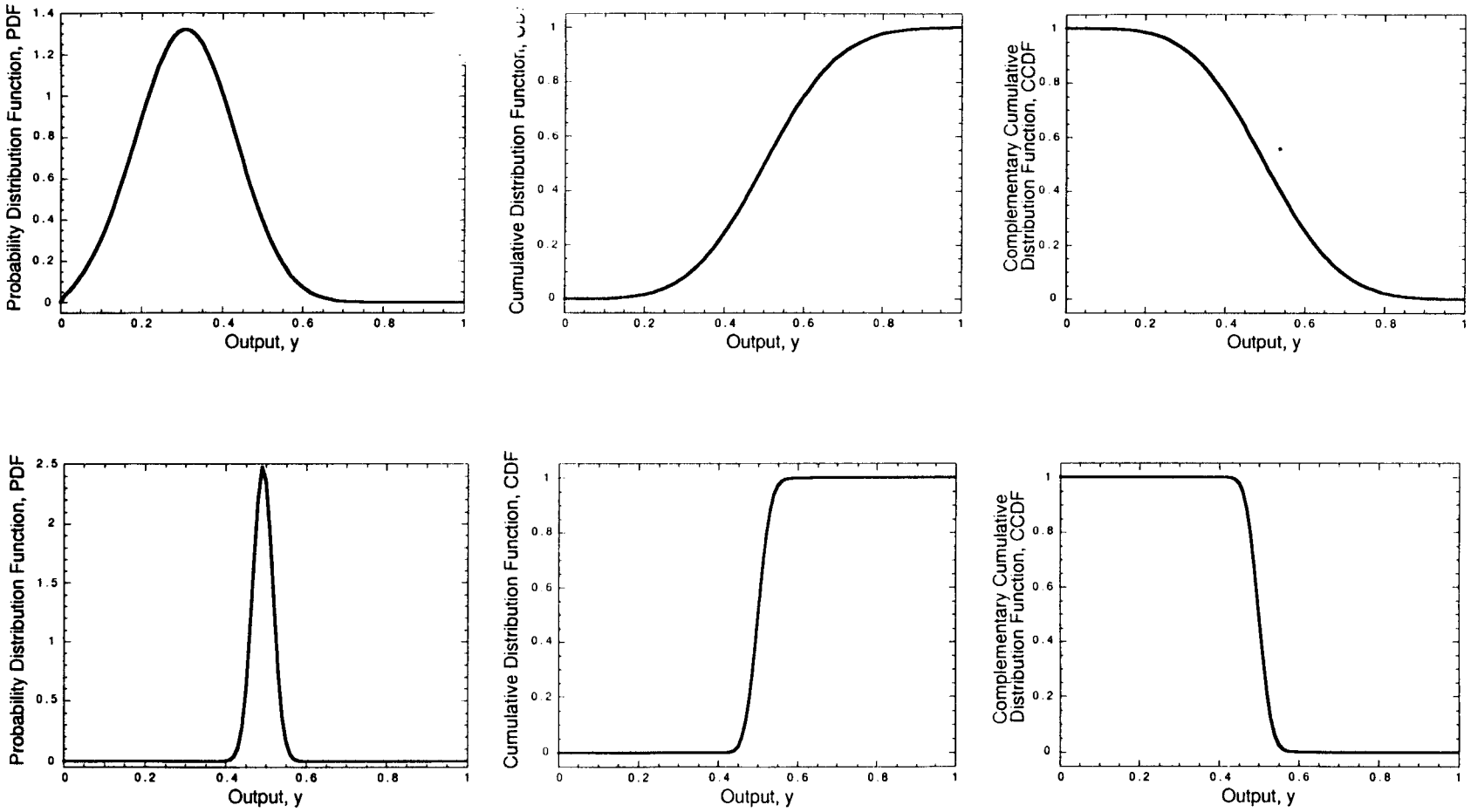
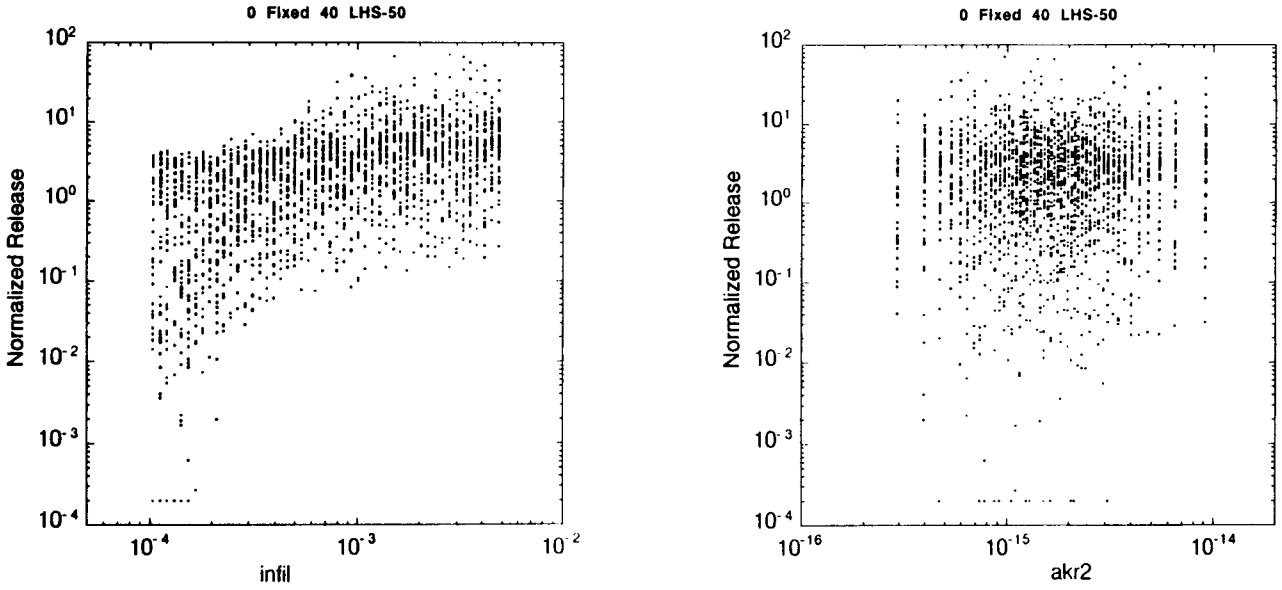
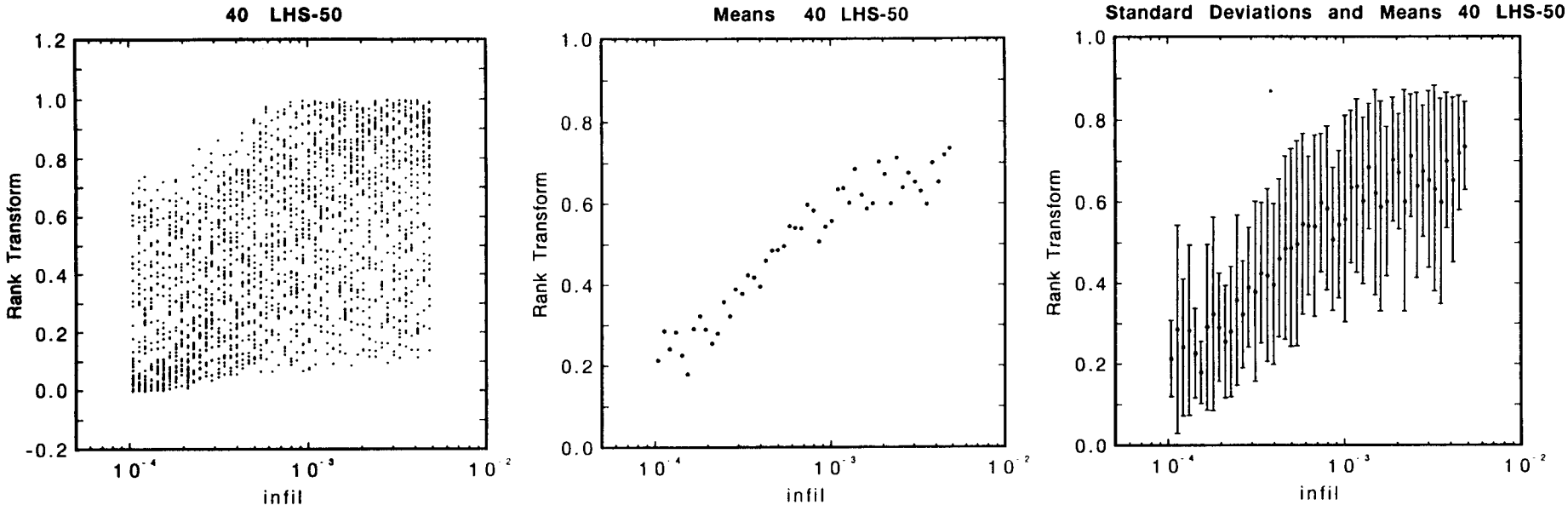


Figure 1: Due to input variability, the output has a distribution described using either a PDF, CDF, or CCDF (top). As the most important input variables are fixed, the range of output decreases (bottom).



(a) (b)

Figure 2: Scatter plots showing strong (a) and weak (b) correlations between inputs and the output.



(a) (b) (c)

Figure 3: Scatter plot using (a) rank transformed output, (b) means, and (c) standard devations within each of the 50 LHS bins.

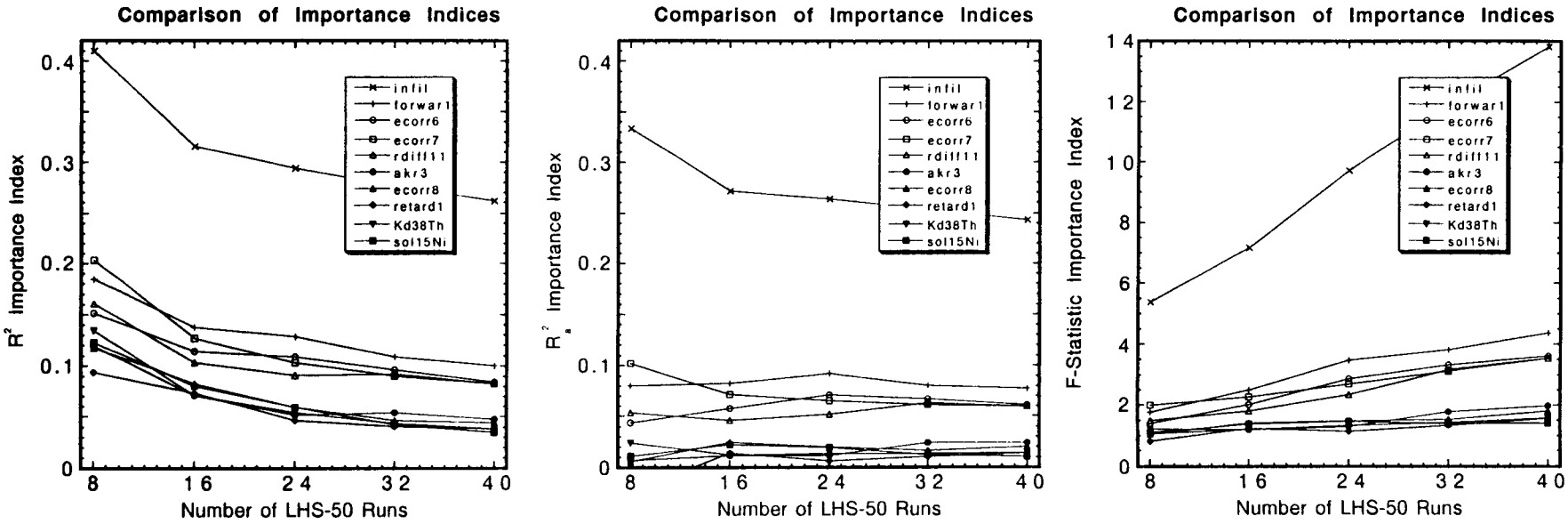
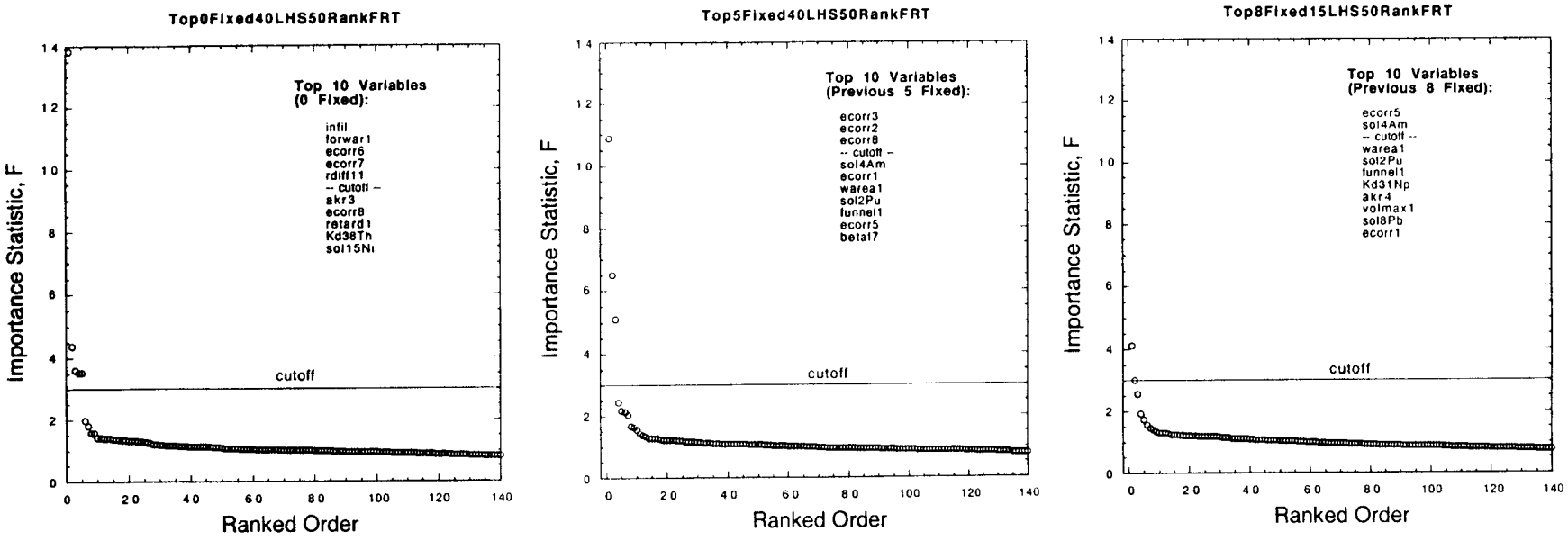


Figure 4: Comparison of convergence properties of R^2, R_a^2, and F importance indices.



(a) (b) (c)

Figure 5: Ranked order plots used to select top ten variables.

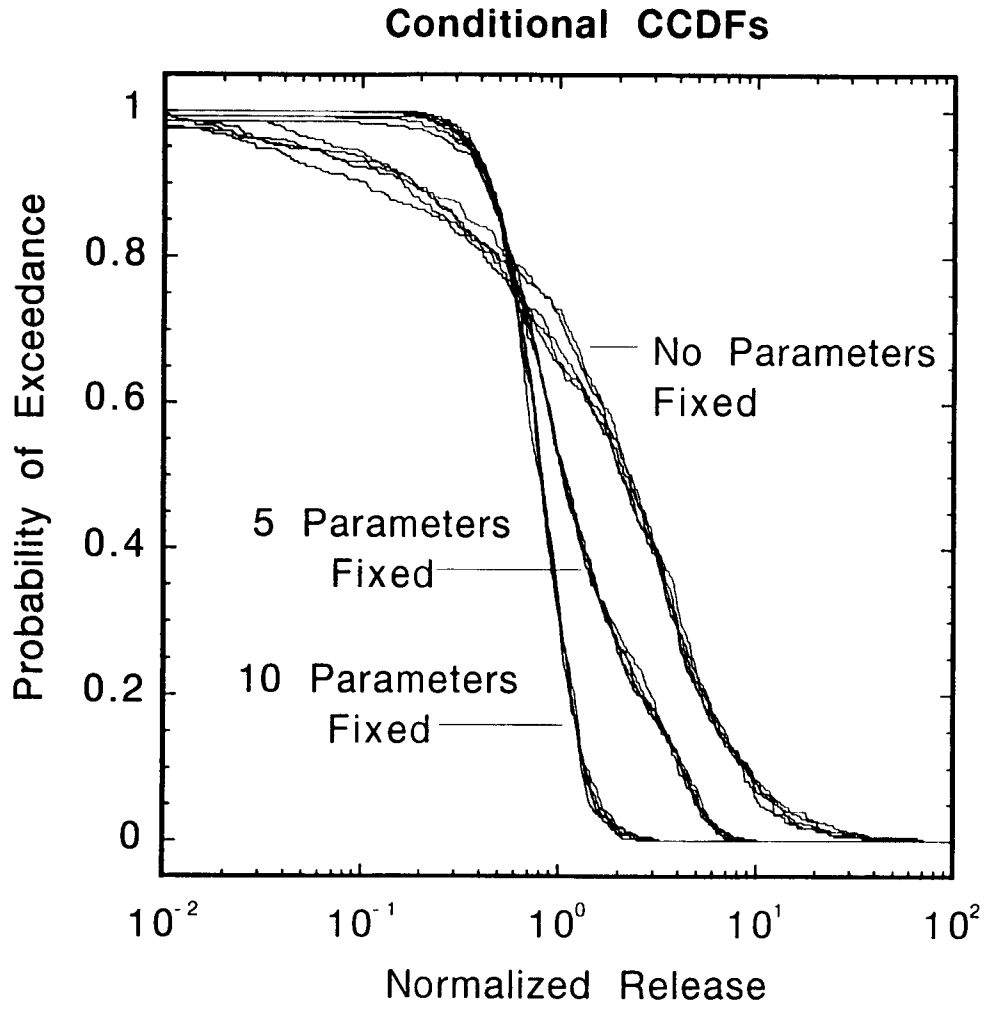
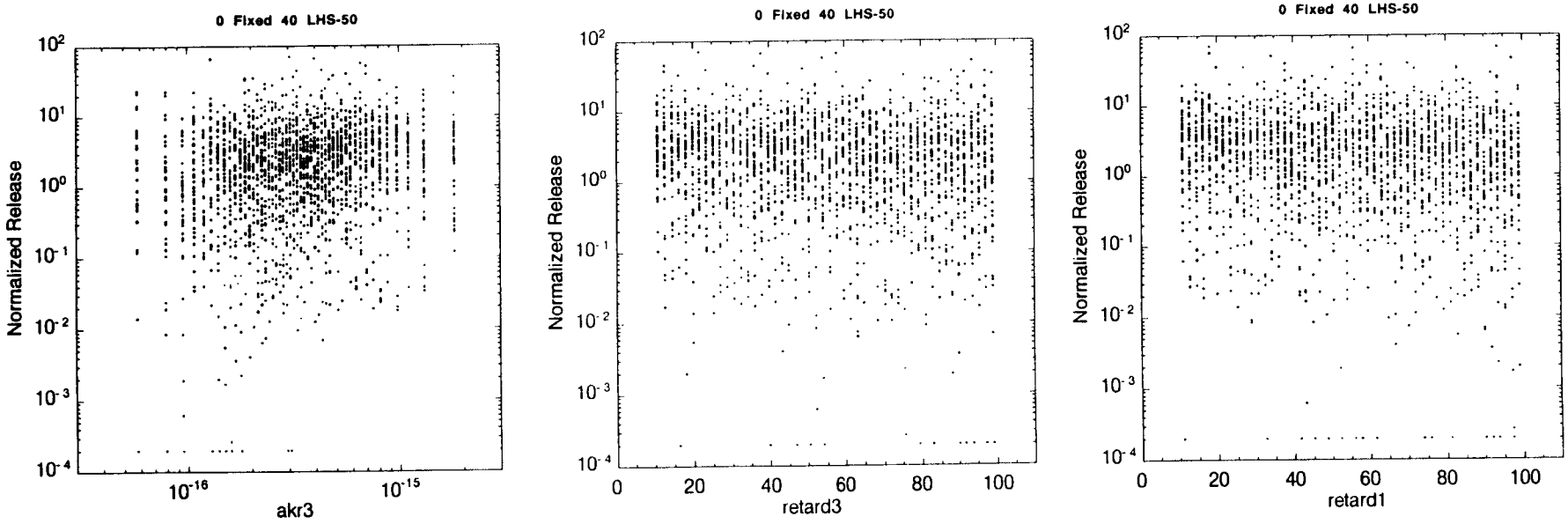
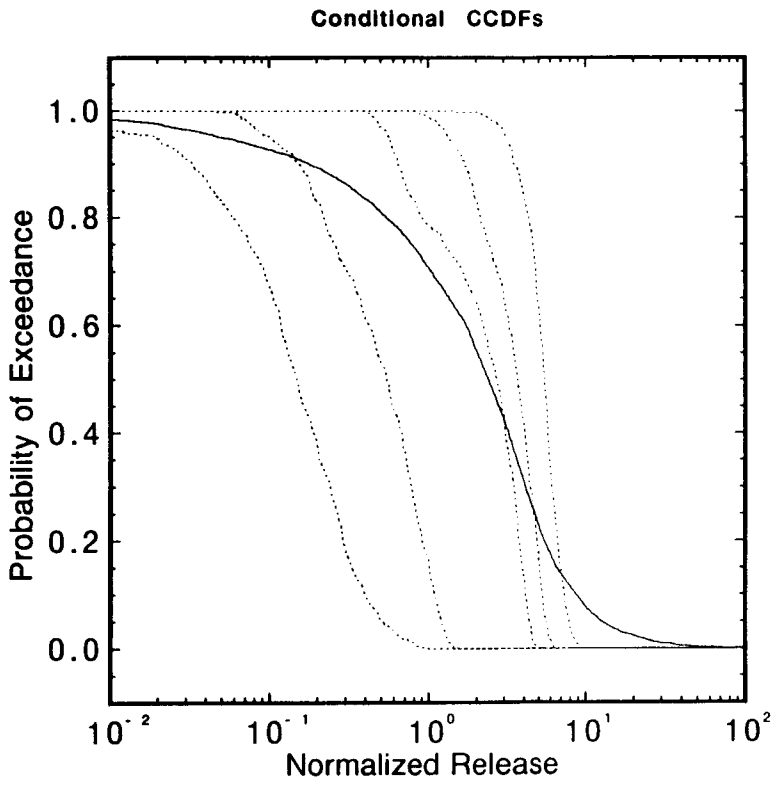


Figure 6: Conditional CCDFs showing how the output variability is reduced by fixing important input parameters.

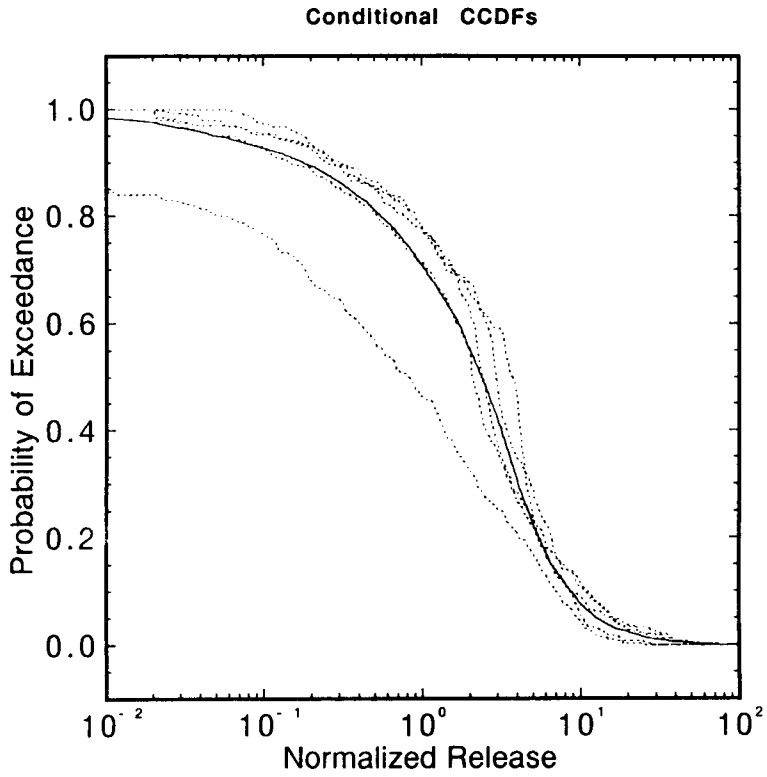


(a) (b) (c)

Figure 7: Scatter plots for parameters identified in the stepwise multilinear regression, yet missed in the variance-based method.



(a)



(b)

Figure 8: Original CCDF (solid line) compared with conditional CCDFs (dashed lines) using (a) coarse LHS-5 on most important parameters and fine LHS-400 on less important parameters, and (b) fine LHS-400 on most important parameters and coarse LHS-5 on less important parameters.