Available online at www.sciencedirect.com
ScienceDirect
Reliability Engineering and System Safety 93 (2008) 1563–1573
RELIABILITY ENGINEERING & SYSTEM SAFETY
www.elsevier.com/locate/ress

Uncertainty and sensitivity analysis for models with correlated parameters

Chonggang Xu, George Zdzislaw Gertner*

Department of Natural Resources and Environmental Sciences, University of Illinois, 1102 South Goodwin Avenue, Urbana, IL 61801, USA

Received 15 February 2006; received in revised form 25 April 2007; accepted 7 June 2007
Available online 26 June 2007

Abstract

When conducting sensitivity and uncertainty analysis, most of the global sensitivity techniques assume parameter independence. However, it is common that the parameters are correlated with each other. For models with correlated inputs, we propose that the contribution of uncertainty to model output by an individual parameter be divided into two parts: the correlated contribution (by the correlated variations, i.e. variations of a parameter which are correlated with other parameters) and the uncorrelated contribution (by the uncorrelated variations, i.e. the unique variations of a parameter which cannot be explained by any other parameters). So far, only a few studies have been conducted to obtain the sensitivity index for a model with correlated input. But these studies do not distinguish between the correlated and uncorrelated contribution of a parameter. In this study, we propose a regression-based method to quantitatively decompose the total uncertainty in model output into partial variances contributed by the correlated variations and partial variances contributed by the uncorrelated variations. The proposed regression-based method is then applied in three test cases. Results show that the regression-based method can successfully measure the uncertainty contribution in the case where the relationship between response and parameters is approximately linear.

Keywords: Correlated parameters; Latin hypercube sampling; Linear regression; Sensitivity analysis; Uncertainty analysis

1. Introduction

Nowadays many complex models have been developed in physics, chemistry, environmental sciences and risk analysis. A consequence of model complexity is that the uncertainty in both model structure and parameter estimation has increased. Thus, the identification and representation of uncertainty is recognized as an essential component in model application [1–3]. In the study of uncertainty, we need to know how much uncertainty there is in the model output (uncertainty analysis) and where the uncertainty comes from (sensitivity analysis).

Many uncertainty and sensitivity analysis techniques are now available [4–16]. In the case of simple mathematical models, Taylor series expansions can be used to approximate the model and the analytical differential sensitivity index can be derived [5,8,17]. However, for complex models, it would be difficult to use the Taylor series approximation, and more advanced techniques are needed. McRae et al. [18] and Saltelli et al. [8] classified the sensitivity techniques into two groups: local sensitivity analysis methods and global sensitivity analysis methods. The local sensitivity analysis techniques examine the local response of the output(s) by varying input parameters one at a time by holding other parameters at central values. The global sensitivity techniques examine the global response (averaged over the variation of all the parameters) of the model output(s) by exploring a finite (or even an infinite) region. The local sensitivity analysis is easy to implement. However, local sensitivity analysis can only inspect one point at a time. Thus, more and more studies nowadays are using global sensitivity analysis methods instead of local sensitivity analysis. Many global sensitivity analysis techniques are available, such as Fourier amplitude sensitivity test (FAST) [18–24]; fractional factorial

*Corresponding author. Tel.: +1 217 333 9346; fax: +1 217 244 3219.
E-mail address: gertner@uiuc.edu (G.Z. Gertner).

0951-8320/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2007.06.003

design method [25–27]; Plackett–Burman technique [28]; Morris method [29]; sampling-based methods [3,10,11,16,17]; Sobol's method [30]; and McKay's method based on an one-way ANOVA [31].

When conducting sensitivity and uncertainty analysis, most of the global sensitivity techniques such as FAST, Sobol's method and sampling-based methods rely on the assumption of parameter independence. However, in many cases, the parameters are correlated with one another. For example, in meteorology, the central pressure of the storm is correlated with the radius of the maximum wind [1]. For models with correlated inputs, we propose that the contribution of uncertainty to model output by an individual parameter be divided into two parts: the correlated contribution (by the correlated variations, i.e. the variations of a parameter which are correlated with other parameters) and the uncorrelated contribution (by the uncorrelated variations, i.e. the unique variations of a parameter which cannot be explained by any other parameters). This distinction between correlated and uncorrelated contribution of uncertainty for an individual parameter is very important, since it can help us decide if we need to focus on the correlated variations among specific parameters (if the correlated contribution dominates) or the parameter itself (if the uncorrelated contribution dominates).

By now, only a few studies have been conducted to obtain the sensitivity index for models with correlated input [1,32–37]. Iman et al. [38] proposed the partial correlation coefficient (a correlation between model output and parameter that remains after adjusting for other parameters) as a measure of parameter sensitivity for models with correlated input based on Latin hypercube sampling [1,38]. Bedford [34] proposed a Gram–Schmidt orthogonalization to obtain the first-order Sobol's indices for correlated input. Saltelli et al. [33,39] proposed a correlation ratio method based on McKay's method. Fang et al. [32] proposed sequential sampling to approximate the differential sensitivity index. However, all methods except for the partial correlation coefficient method only provide an overall sensitivity index of one parameter, which does not distinguish the correlated or uncorrelated contribution of one parameter. The partial correlation coefficient method only provides a relative sensitivity index for the uncorrelated contribution but not for the correlated contribution. If two variables are highly correlated, then neither one will show up as being important in a sensitivity analysis based on the partial correlation coefficient. In addition, if the model is close to linear for each variable, then each variable will have a partial correlation coefficient close to 1 in absolute value even though their effect on the uncertainty in model results may be very different.

In this study, we use a regression-based method to quantitatively decompose the variances in the model output into partial variances contributed by the correlated and uncorrelated variations of parameters.

2. Methods

2.1. Variance decomposition by regression with independent input

For a model $y = f(x)$ $[x = (x_1, x_2 \ldots, x_i, \ldots, x_K)]$, if the effect of each parameter x_i on model output is linear and we only care about the main effects of each parameter, the model can be simplified as follows:

$$y = \beta_0 + \sum_{i=1}^{K} \beta_i x_i + e, \tag{1}$$

where $\beta_0 \ldots \beta_k$ are regression coefficients and e is the error. We can use the regression based on Eq. (1) to obtain the sensitivity index of each parameter given that the model at hand is approximately linear. If the parameters are mutually independent and are independent of e, based on Eq. (1), the variance in model output can be decomposed as follows:

$$V = \sum_{i=1}^{K} V_i + V_e, \tag{2}$$

where V_e is the variance contributed by the error in Eq. (1) and V_i is the variance contributed by parameter x_i. V_e is the variance in model output unaccounted for after removing the linear main effects of all the parameters. It can include non-linear effects or higher order interaction effects which can be present in the model.

Latin hypercube sampling can be used to explore the sensitivity of parameters based on the regression using Eq. (1). The Latin hypercube sampling method is an efficient sampling method compared to the simple random sampling method [16,40]. Latin hypercube sampling first stratifies the range of each variable (i.e. x_i) into N disjoint intervals of equal probability and then one random value is drawn from each interval. Thus, each parameter will have N random values. The sampled random values of all parameters are randomly permuted to form a sample of size N,

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NK} \end{bmatrix}, \tag{3}$$

where the ith column of the matrix represents the N permuted random sample points for x_i. The model is then run on the sample and N response values of y are generated. Namely,

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{bmatrix} f(x_{11}, \ldots, x_{1i}, .., x_{1K}) \\ \vdots \\ f(x_{N1}, \ldots, x_{Ni}, .., x_{NK}) \end{bmatrix}. \tag{4}$$

The partial variance (V_i) and total variance (V) can be estimated as follows:

$$\hat{V}_i = \hat{\beta}_i^2 \mathrm{var}(x_i) = \frac{1}{N-1}\hat{\beta}_i^2 \sum_{j=1}^{N}(x_{ji} - \bar{x}_i)^2,$$

$$\hat{V} = \mathrm{var}(y) = \frac{1}{N-1}\sum_{j=1}^{N}(y_j - \bar{y})^2, \quad (5)$$

where x_{ji} is the jth sample element for parameter x_i; $\bar{x}_i$ is the sample mean of parameter x_i from the Latin hypercube sample, $\bar{y}$ the mean response value based on the Latin hypercube sample; and $\hat{\beta}_i$ is the least-square estimate of β_i in the regression in the form of Eq. (1)

$$\hat{\beta} = (\hat{\beta}_0 \ldots \hat{\beta}_i \ldots \hat{\beta}_K)' = ([\underline{1}_N X]'[\underline{1}_N X])^{-1}[\underline{1}_N X]' Y, \quad (6)$$

where $\underline{1}_N$ is a N by 1 column vector of 1's and $[\underline{1}_N X]$ represents the N by $(K+1)$ matrix by inserting $\underline{1}_N$ to the left of the matrix X. Based on Eq. (5), the sensitivity of parameter x_i can then be calculated as

$$S_i = \frac{\hat{V}_i}{\hat{V}}. \quad (7)$$

The variance in model output that can be explained by the regression of Eq. (1) is

$$\hat{V}^{\mathrm{L}} = \frac{1}{N-1}\sum_{j=1}^{N}(\hat{y}_j - \bar{y})^2, \quad (8)$$

where $\hat{y}_j = \hat{\beta}_0 + \sum_{i=1}^{K}\hat{\beta}_i x_{ji}$. The statistic $R^2 = \hat{V}^{\mathrm{L}}/\hat{V}$, which is also termed as the multiple correlation coefficient in regression [41], can be used to indicate how well the regression in form of Eq. (1) approximates the model under study. The closer the statistic is to 1, the closer the approximation of Eq. (1) is to the model.

2.2. Variance decomposition by regression with correlated input

If the parameter input is correlated, we can also use the Latin hypercube sampling to generate a rank-correlated sample X for the model [42]. However, Eq. (2) may not hold, since V_i will include the contributions of other parameters due to the correlation [4]. We propose to decompose V_i into partial variance (V_i^{U}) contributed by the uncorrelated variations of parameter x_i (variations unique to x_i) and partial variance (V_i^{C}) contributed by the correlated variations of parameter (variations x_i correlated with other parameters), namely

$$V_i = V_i^{\mathrm{U}} + V_i^{\mathrm{C}}. \quad (9)$$

If the effect of the parameter on the model output is approximately linear, the partial variance in the model output contributed by x_i, V_i, can be derived by regressing y on only x_i, that is,

$$y = \theta_0 + \theta_i x_i + e, \quad (10)$$

where θ_0 and θ_i are the regression coefficients and e represents the error. The partial variance V_i can be estimated as follows:

$$\hat{V}_i = \frac{1}{N-1}\sum_{j=1}^{N}(\hat{y}_j^{(i)} - \bar{y})^2, \quad (11)$$

where $\hat{y}_j^{(i)} = \hat{\theta}_0 + \hat{\theta}_i x_{ji}$. $\hat{\theta}_0$ and $\hat{\theta}_i$ are the least-square estimates of θ_0 and θ_i for linear regression of Eq. (10) [41],

$$\hat{\theta} = (\hat{\theta}_0 \hat{\theta}_i)' = ([\underline{1}_N X_{(i)}]'[\underline{1}_N X_{(i)}])^{-1}[\underline{1}_N X_{(i)}]' Y, \quad (12)$$

where $X_{(i)}$ is the ith column of the matrix of X. Since x_i contains both the correlated and uncorrelated variations, the partial variance calculated using Eq. (11) is the total partial variance, which can be explained by x_i.

The partial variance contributed by the variation of x_i uncorrelated with all other parameters, V_i^{U}, can be derived by the regression as follows:

$$y = r_0 + r_i \hat{z}_i + e, \quad (13)$$

where r_0 and r_i are the regression coefficients and $\hat{z}_i$ is the estimated residual from the regression of x_i over all other parameters $x_{j_s}(j_s \neq i)$

$$\hat{z}_i = x_i - \hat{x}_i = x_i - \left(\hat{\eta}_0 + \sum_{j_s \neq i}\hat{\eta}_{j_s} x_{j_s}\right), \quad (14)$$

where $\hat{\eta}_{j_s}$ $[s = 1, 2 \ldots K-1;\ (j_s \neq i)]$ is the least-square estimation obtained as follows [41]:

$$\hat{\eta} = (\hat{\eta}_0 \hat{\eta}_{j_1} \ldots \hat{\eta}_{j_s} \ldots \hat{\eta}_{j_{(K-1)}})' = ([\underline{1}_N X_{(-i)}]'[\underline{1}_N X_{(-i)}])^{-1} \times [\underline{1}_N X_{(-i)}]' X_{(i)}, \quad (15)$$

where $X_{(-i)}$ is the matrix X without the ith column.

For the regression of x_i over all other parameters $x_{j_s}(j_s \neq i)$, since the estimation space (the vector space where the estimation $\hat{x}_i$ exists) is perpendicular to the error space (the space where the residual $\hat{z}_i$ exists) [43], the variation in $\hat{z}_i$ is uncorrelated with all other parameters x_{j_s} $(j_s \neq i)$. Thus, the variance of y explained by $\hat{z}_i$ in regression of Eq. (13) is contributed by the uncorrelated variations of x_i. Finally, V_i^{U} can be derived as follows:

$$\hat{V}_i^{\mathrm{U}} = \frac{1}{N-1}\sum_{j=1}^{N}(\hat{y}_j^{(-i)} - \bar{y})^2, \quad (16)$$

where $\hat{y}_j^{(-i)} = \hat{r}_0 + \hat{r}_i \hat{z}_{ji}$. $\hat{r}_0$ and $\hat{r}_i$ are least-square estimates of r_0 and r_i in the regression of Eq. (13),

$$\hat{\gamma} = (\hat{r}_0 \hat{r}_i)' = ([\underline{1}_N \hat{Z}_i]'[\underline{1}_N \hat{Z}_i])^{-1}[\underline{1}_N \hat{Z}_i]' Y, \quad (17)$$

where

$$\hat{Z}_i = \begin{pmatrix} \hat{z}_{1i} \\ \vdots \\ \hat{z}_{Ni} \end{pmatrix} = \begin{pmatrix} x_{1i} - \left(\hat{\eta}_0 + \sum_{j_s \neq i} \hat{\eta}_{j_s} x_{1j_s}\right) \\ \vdots \\ x_{Ni} - \left(\hat{\eta}_0 + \sum_{j_s \neq i} \hat{\eta}_{j_s} x_{Nj_s}\right) \end{pmatrix}.$$

Based on Eqs. (11) and (16), we can obtain the partial variance in the model outputs contributed by the total and uncorrelated variations of parameter x_i. Based on Eq. (9), it becomes clear that we can estimate the partial variance contributed by the variations of x_i correlated with $(x_{j_1} \ldots x_{j_s} \ldots x_{j_{(K-1)}})$ by the following equation:

$$\hat{V}_i^{\mathrm{C}} = \hat{V}_i - \hat{V}_i^{\mathrm{U}}. \tag{18}$$

Finally, based on Eqs. (11), (16) and (18), we can get the partial variance in model output contributed by the uncorrelated and correlated variations of each parameter. Using the ratio between the partial variance and total variance, we can get the total (S_i), correlated (S_i^{C}), and uncorrelated (S_i^{U}) contribution of parameter x_i (namely, the first-order sensitivity indices):

$$S_i = \frac{\hat{V}_i}{\hat{V}}, \quad S_i^{\mathrm{U}} = \frac{\hat{V}_i^{\mathrm{U}}}{\hat{V}}, \quad S_i^{\mathrm{C}} = \frac{\hat{V}_i^{\mathrm{C}}}{\hat{V}}. \tag{19}$$

We treat the correlation among parameters and response as positive by default. But the correlation is not necessarily positive. However, in the case of negative correlation, it is possible that the correlated contribution becomes negative. For example, consider the model $y = x_1 + x_2$, then

$$\mathrm{var}(y) = \mathrm{var}(x_1) + \mathrm{var}(x_2) + 2\mathrm{cov}(x_1, x_2). \tag{20}$$

If the correlation between x_1 and x_2 is negative, then the correlated contribution may be negative. In this case, the total contribution of x_1 and x_2 are less than the uncorrelated contributions.

We need to mention that our method replies on the assumption of an approximately linear effect. For models with highly non-linear effect, refer to Section 4.3 for more details.

3. Application

In this section, we apply the regression-based method to three test cases using a Latin hypercube sample of size 1000. In order to test the reliability of the proposed method, we also conduct the correlation ratio method as a reference. The correlation ratio method for models with correlated parameters is proposed by Saltelli based on McKay's one-way ANOVA method [33]. It is based on the replicated Latin hypercube sampling and suitable for non-linear and non-monotonic models. However, it needs replicated samples, and thus a large sample size is required. In this paper, the correlation ratio method is based on the 100 replications with each replicate having a sample size of 500 (a total of 50 000 model runs to get the sensitivity indices for all parameters).

3.1. Test case—one

In the first test case, we use a very simple model $y = 2x_1 + 3x_2$, where x_1 and x_2 are standard normally distributed with a Pearson correlation coefficient of 0.7. Based on the analytical decomposition,

$$V = \mathrm{var}(y) = 4\mathrm{var}(x_1) + 9\mathrm{var}(x_2) + 12\mathrm{cov}(x_1, x_2). \tag{21}$$

Since x_1 includes both uncorrelated and correlated variations, the variance component $4\mathrm{var}(x_1)$ in the model output includes the contribution from both the uncorrelated and correlated contribution of x_1. The partial variance contributed by the uncorrelated variations of x_1 is the conditional variance of $2x_1$ given x_2

$$V_1^{\mathrm{U}} = \mathrm{var}(2x_1|x_2) = (1 - 0.7^2) \times 4\mathrm{var}(x_1). \tag{22}$$

Similarly, the partial variance contributed by the uncorrelated variations of x_2 is

$$V_2^{\mathrm{U}} = \mathrm{var}(3x_2|x_1) = (1 - 0.7^2) \times 9\mathrm{var}(x_2). \tag{23}$$

Finally, the correlated contribution of x_1 and x_2 to variances in the model output should be

$$V_1^{\mathrm{C}} = V_2^{\mathrm{C}} = 12\mathrm{cov}(x_1, x_2) + 0.7^2 \times 4\mathrm{var}(x_1) + 0.7^2 \times 9\mathrm{var}(x_2). \tag{24}$$

Based on Eqs. (21)–(24), we can analytically obtain the variance contribution by the correlated and uncorrelated

Table 1
Uncertainty decomposition for model $y = 2x_1 + 3x_2$

Statistics	Partial variance by regression	Variance contribution (%) by regression	Analytical partial variance	Analytical variance contribution (%)
Total variations of x_1	17.73	79.48	16.81	78.55
Uncorrelated variation of x_1	1.95	8.74	2.04	9.53
Correlated variations of x_1	15.78	70.75	14.77	69.02
Total variations of x_2	20.36	91.26	19.36	90.47
Uncorrelated variations of x_2	4.58	20.52	4.59	21.45
Correlated variations x_2	15.78	70.75	14.77	69.02

variations of x_1 and x_2 (Table 1). For the regression-based method, we can obtain the uncorrelated and correlated contribution of x_1 and x_2 based on the Latin hypercube sample using Eqs. (11), (16) and (18) (Table 1).

The results show that there is a good agreement between the analytical contribution and the regression-based contribution (Table 1). This suggests that the method we present in Section 2.2 can accurately quantify the uncorrelated and correlated contributions of linear model parameters. Also, with the correlation ratio method, the total contribution of x_1 (78.44%) and x_2 (90.48%) is close to the analytical contribution, which indicates that, as expected, the correlation ratio method can accurately measure the total contribution. However, our test case is a very simple linear model, for a complex linear model, the accuracy may decrease.

3.2. Test case—two

In this case, we test a non-linear model

$$y = \frac{x_1 x_2}{x_3}, \tag{25}$$

where x_1, x_2 and x_3 are uniformly distributed with a lower bound of 1 and upper bound of 10. A more practical measure of correlation among parameters with non-normal distributions is the rank correlation coefficient [42,44]. The rank correlation coefficient describes the rank dependency among variables. In this case, the rank correlation matrix for x_1, x_2, and x_3 is as follows:

$$\begin{bmatrix} 1 & 0.4 & 0.2 \\ 0.4 & 1 & 0.4 \\ 0.2 & 0.4 & 1 \end{bmatrix}. \tag{26}$$

Based on a Latin hypercube sample with size of 1000, a linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e \tag{27}$$

is built. The scatter plot matrix is shown in Fig. 1. Due to the interactions among parameters and non-linearity between x_3 and y (Fig. 1), the regression of Eq. (27) only has a multiple correlation coefficient of 0.743. However, the regression-based contribution is very close to that obtained with the correlation ratio method (Fig. 2). For x_3, due to the negative correlation between x_3 and y (Fig. 1), the correlated contribution is negative and the total

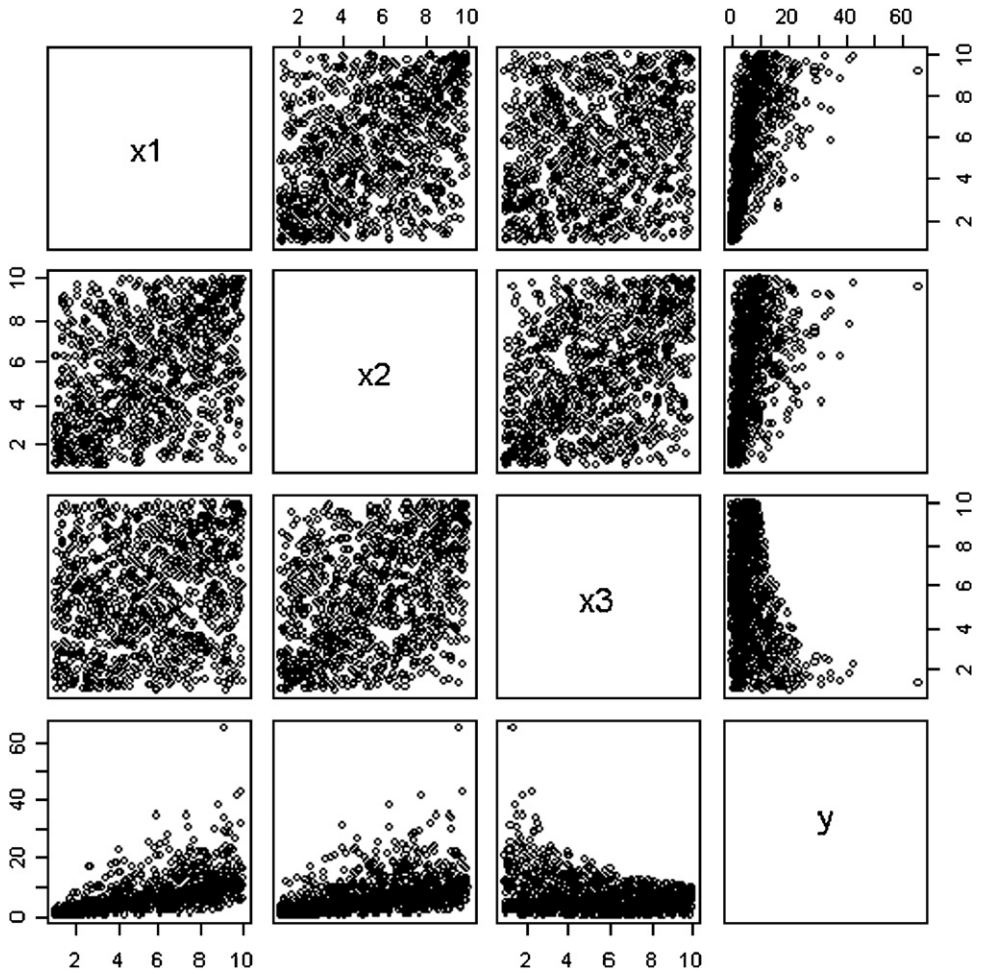

Fig. 1. Scatter plot matrix for test case two.

contribution is less than the uncorrelated contribution (Fig. 2).

As to the magnitude of the uncorrelated and correlated contributions, for x_1 and x_3, both uncorrelated and correlated contributions are important. However, for x_2, only the uncorrelated contribution dominates.

We need to point out that a log-transformation may transform the non-linear model to a linear model. For a sensitivity analysis conducted on the log-transformed data, refer Section 4.3.

3.3. Test case—three

In this section, we test a complex real model: World3. The World3 model is a well-known computer program to simulate the interactions among human population, industrial growth, food production and limits in the ecosystems of the earth [45]. The model consists of six main systems: food system, agriculture system, industrial system, population system, non-renewable resources system and the pollution system.

In this study, we are concerned with the industrial system, which can provide the products for the world population [45]. At the same time, this system creates pollution, which reduces the land productivity (for details, please refer to Meadows et al. [45]).

The seven parameters of interest are shown in Table 2. We assume the uniform distribution for each parameter. The bounds for each parameter are assumed to be a 10% deviation of the central value (Table 2). The version of the model that results in SCENARIO 10 specified in Meadows et al. [45] is used for the test case. We assume that there is a positive rank correlation of 0.6 between x_2 and a x_3 and positive rank correlation of 0.4 between x_4 and x_5. The output of interest is the world human population on a yearly basis.

The output for the world population based on 1000 runs is shown in Fig. 3. We illustrated the results of the correlation ratio method and regression-based method in Fig. 4. However, we only show the results for x_2, x_3, x_4 and x_6. All the other parameters having a contribution less than 2% are not illustrated in view of the fact that the small numbers are very susceptible to numerical error.

Results show that the total contribution of the parameters calculated by the regression-based method is in reasonably good agreement with that from the correlation ratio method before the year 2025 (Fig. 4). For x_2, the correlated and uncorrelated contributions are of the same order. Thus, both the correlated and uncorrelated variations of x_2 have important contributions to the uncertainty in model output. For x_3, the uncorrelated contribution is

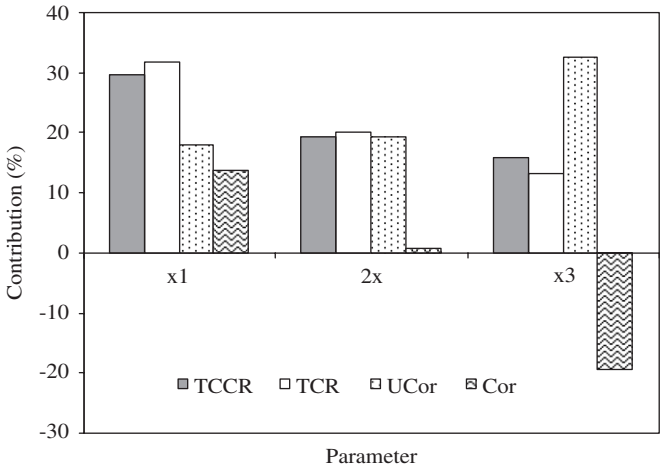

Fig. 2. Variance contribution by parameter for test case two. TCCR is the total contribution calculated by correlation ratio method; TCR is the total contribution calculated by regression-based method; Ucor is the uncorrelated contribution by regression-based method; and Cor is the correlated contribution by regression-based method.

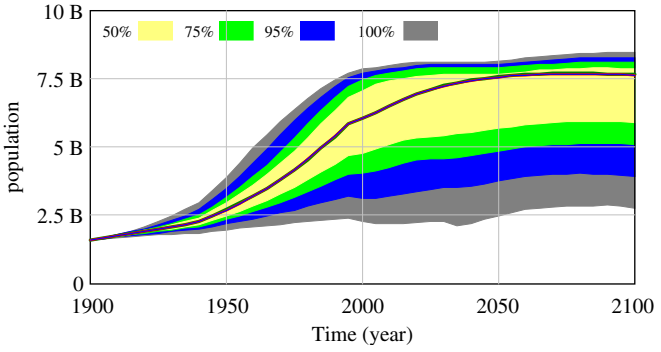

Fig. 3. World3 model output distribution for world population size in billions. The centerline is the output based on the central values of parameters. 50% (yellow) represents 50% confidence interval of the central line; 75% (green) represents 75% confidence interval of the central line; 95% (blue) represents 95% confidence interval of the central line; and 100% (blue) represents the output boundary.

Table 2
Parameter specification of uniform distribution for test case three

Parameter	Label	Central value	Lower bound	Higher bound
x_1	Industrial output per capita desired	350	315	385
x_2	Fraction of industrial output allocated to consumption before 1995	0.43	0.387	0.473
x_3	Fraction of industrial output allocated to consumption after 1995	0.43	0.387	0.473
x_4	Average life of industrial capital before 1995	14	12.6	15.4
x_5	Average life of industrial capital after 1995	18	16.2	19.8
x_6	Industrial capital output ratio before 1995	3	2.7	3.3
x_7	Initial industrial capital	$2.1(10^{+11})$	$1.89(10^{+11})$	$2.31(10^{+11})$

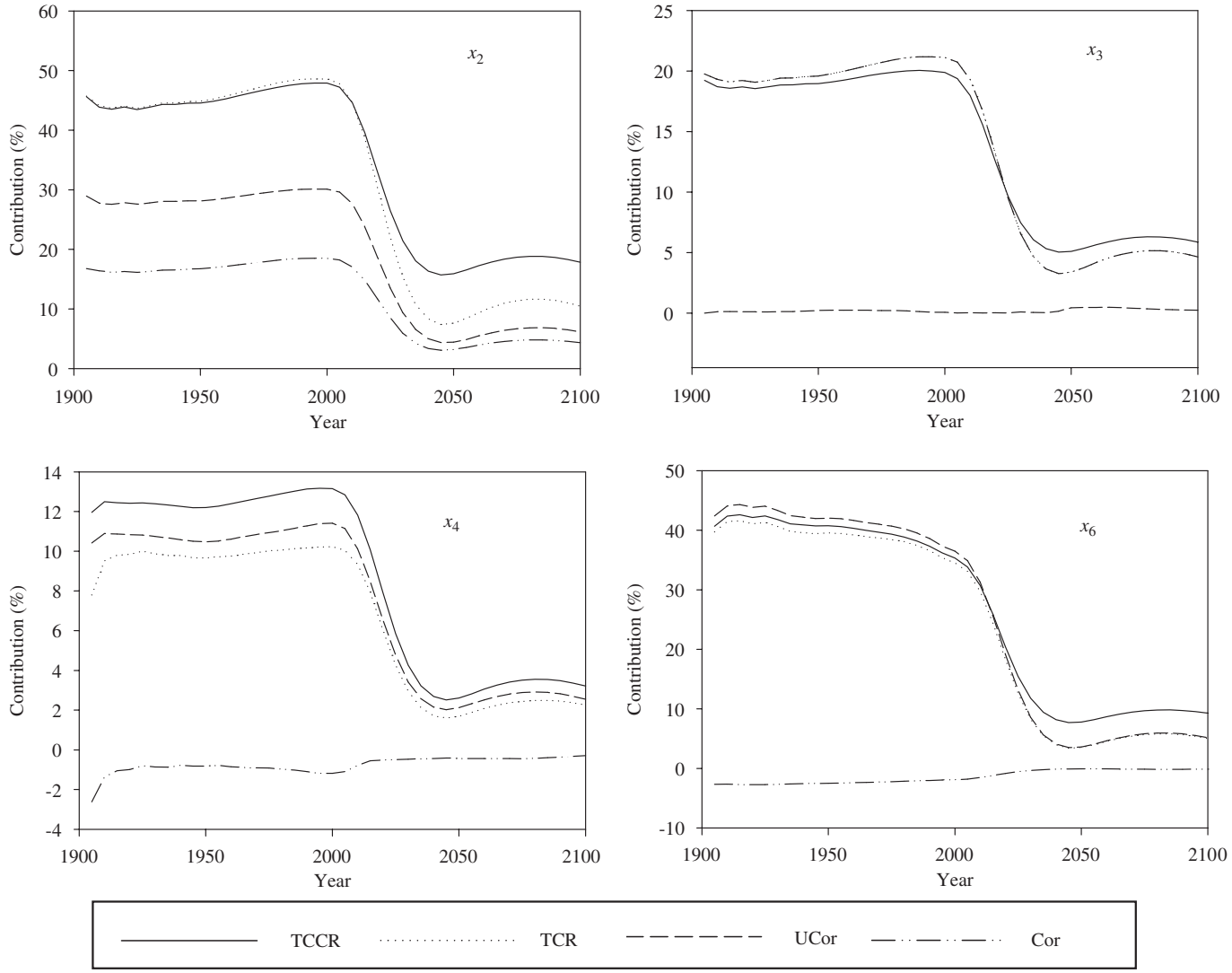

Fig. 4. Variance contribution by parameter for test case three. TCCR is the total contribution calculated by correlation ratio method; TCR is the total contribution calculated by regression-based method; UCor is the uncorrelated contribution; and Cor is the correlated contribution. For x_3, since uncorrelated contribution is close to 0, there are very little differences between TCR and Cor. Thus, they are overlapped in the upper right figure.

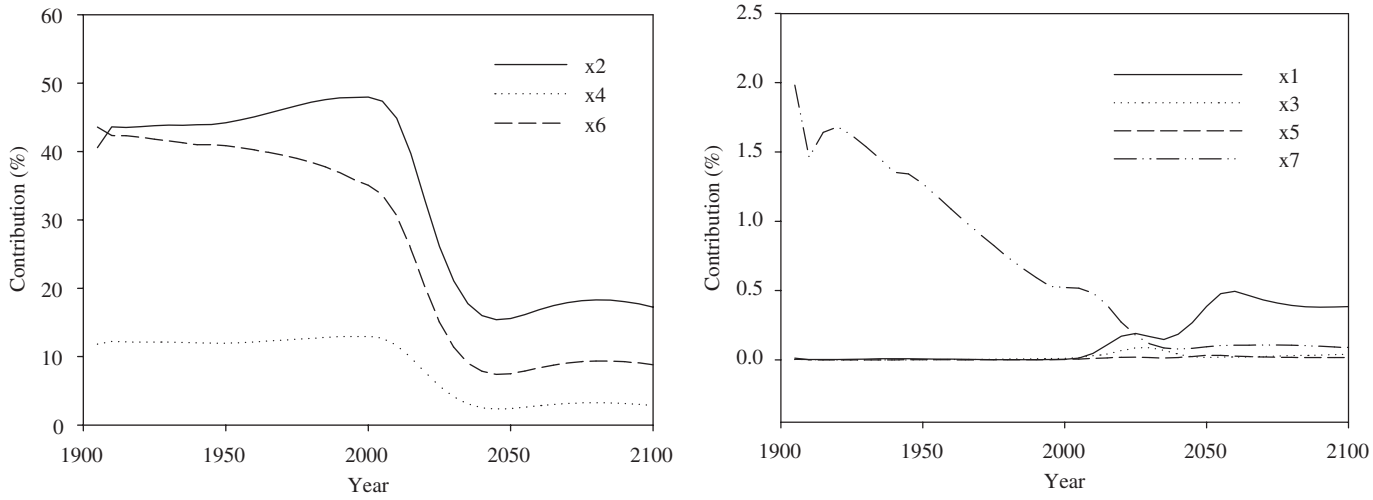

Fig. 5. Sensitivity indices of independent individual parameters for World3 model using FAST.

Table 3
Simulated rank correlation matrix for test case three

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1.0000	0.0656	0.0789	0.0392	0.0864	0.0247	0.0559
x_2	0.0656	1.0000	0.6086	0.0508	0.0609	0.0417	0.0568
x_3	0.0789	0.6086	1.0000	0.0666	0.0751	0.0686	0.0760
x_4	0.0392	0.0508	0.0666	1.0000	0.4264	0.0442	0.0406
x_5	0.0864	0.0609	0.0751	0.4264	1.0000	0.0627	0.0477
x_6	0.0247	0.0417	0.0686	0.0442	0.0627	1.0000	0.0488
x_7	0.0559	0.0568	0.0760	0.0406	0.0477	0.0488	1.0000

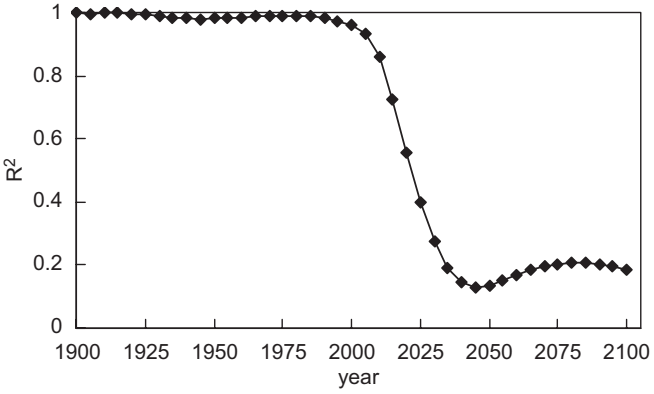

Fig. 6. Multiple correlation coefficient for linear regression in test case three.

close to zero. This suggests that x_3 itself is not important. Only the correlated variations between x_2 and x_3 have an effect. This can be validated by sensitivity analysis with no parameter correlation considered. We use the FAST method to calculate the first-order sensitivity index, while assuming all the parameters are independent. Results show that the independent contribution of x_3 is close to zero (Fig. 5).

For x_4, the uncorrelated contribution dominates. The correlated contribution is very small, although there is correlation between x_4 and x_5. This is because x_5 has a relatively small contribution to the uncertainty in model output and the correlation between x_4 and x_5 does not have much of an effect on the sensitivity of x_4. Thus, for x_4, the correlated contribution is not important. We only need to care about the uncorrelated contribution of x_4. The total contribution is very close to that of a parameter that is independent (Fig. 5). The difference between the total contribution calculated by the correlation ratio method and the regression method may result from the non-linearity of parameter effect.

For x_6, there is relatively small correlated contribution compared to the uncorrelated contribution. This is in agreement with our parameter settings, since there is no correlation between x_6 and other parameters specified. The non-zero correlated contribution may be due to the numerical error in the correlation calculation when using the Latin hypercube sampling (Table 3).

For all parameters except for x_4, the discrepancy between the two methods increases when the multiple correlation coefficients are less than 0.5, which is around simulation year 2025 (Fig. 6). The total contribution calculated from the regression method is much less than that from the correlation ratio method, especially for x_2 (Fig. 4). For x_4, there are greater differences between the regression-based method and the correlation ratio method before year 2025, which indicate that there is a strong non-linear effect of x_4 from the beginning. However, the difference decreases even though the multiple correlation coefficients are less than 0.5. This indicates that non-linearity effect decreases as time goes on.

4. Discussion

4.1. Causality and interpretation

In practice, the correlated variations result from unknown mechanisms among parameters. Correlation is a relatively simple and empirical way to incorporate the unknown mechanism. However, because we do not know the real mechanism, it would be difficult to accurately interpret what the correlated and uncorrelated contributions represent. Following factor analysis [46], a simple way to interpret the correlated and uncorrelated contribution is that the correlated contribution result from a common independent latent factor and the uncorrelated contributions result from an independent latent factor specific to the parameter of interest.

For example, for a two-parameter model $y = f(x_1, x_2)$, the correlated contribution of x_1 or x_2 is contributed by a common latent factor t_3 (Fig. 7). The uncorrelated contribution by x_1 is contributed by latent factor t_1. The uncorrelated contribution by x_2 is contributed by another latent factor t_2.

The latent factor accounting for correlated/uncorrelated contribution could be the parameter itself. For example, for model $y = f(x_1, x_2)$, if the correlation between x_1 and x_2 is 1 (causative relationship is $x_1 \rightarrow x_2$), then the uncorrelated contribution will be 0. We would like to focus on the correlated variations between x_1 and x_2 in the model. However, in this case, the latent factor t_3 accounting for the correlated contribution will be x_1 itself. If we do not know the casualty initially, then t_3 could be x_1 or x_2. For model $y = f(x_1, x_2)$, if the correlation between x_1 and x_2 is 0, then the correlated contribution will be 0. We would like to

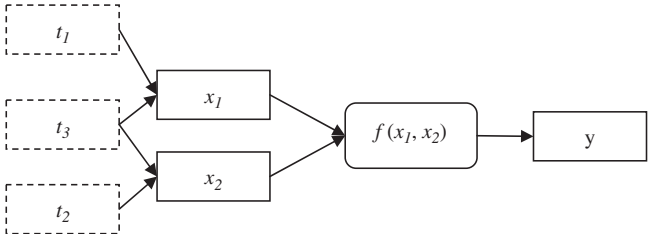

Fig. 7. Causality diagram for model $f(x_1, x_2)$.

focus on x_1 and x_2 themselves in the model. In this case, the latent factor t_1 will be x_1 itself and t_2 will be x_2. For model $y = f(x_1, x_2)$, if x_1 is partially caused by x_2 and t_1 is not important, then the uncorrelated contribution of x_1 is close to 0 and both the correlated and uncorrelated contribution of x_2 is high. In this case, both t_2 and t_3 could be x_2.

Overall, we should be very careful when interpreting the sensitivity analysis results. Basic knowledge about the underlying mechanism responsible for the correlation would help to interpret the results.

4.2. Uncertainty decomposition

Several uncertainty decomposition methods for linear models with correlated input have been developed. Bedford [34] proposed applying a Gram–Schmidt orthogonalization on the correlated parameter input to get the orthogonal and zero mean parameter space $g = g_1 \dots g_i \dots g_k$. For such a situation, we can apply a linear regression over g,

$$y = \beta_0 + \sum_{i=1}^{K} \beta_i g_i + e. \tag{28}$$

Since $g_1, \dots, g_i, \dots, g_k$ are mutually uncorrelated, the variance of y can be decomposed as follows:

$$\mathrm{var}(y) = \sum_{i=1}^{K} \beta_i^2 \mathrm{var}(g_i) + \mathrm{var}(e). \tag{29}$$

For a regression with zero mean and orthogonal input, the least-square estimation of β_i is as follows:

$$\hat{\beta}_i = \mathrm{var}(g_i)^{-1} \mathrm{cov}(y, g_i). \tag{30}$$

The variance contributed by parameter x_i is

$$\begin{aligned} V_i &= \hat{\beta}_i^2 \mathrm{var}(g_i) \\ &= (\mathrm{var}(g_i)^{-1} \mathrm{cov}(y, g_i))^2 \mathrm{var}(g_i) \\ &= \frac{\mathrm{cov}(y, g_i)^2}{\mathrm{var}(g_i)}. \end{aligned} \tag{31}$$

The variance decomposition in Bedford's method is similar to the sequential sum of square decomposition used in linear regression [47],

$$\begin{aligned} \mathrm{SSM_m} = {} & \mathrm{SS}(\beta_1|\beta_0) \\ & + \mathrm{SS}(\beta_2|\beta_0, \beta_1) \\ & + \mathrm{SS}(\beta_3|\beta_0, \beta_1, \beta_2) \\ & + \dots \\ & + (\beta_k|\beta_0, \beta_1, \beta_2, \dots, \beta_{k-1}), \end{aligned} \tag{32}$$

where $\mathrm{SSM_m}$ is the corrected total sum of square of the regression model and $\mathrm{SS}(\beta_i|\beta_1, \dots, \beta_j)$ $(j \neq i)$ is the sum of square accounted for by adding the predictor x_i to the linear regression model already containing an intercept and $(x_1, \dots, x_j)$ $(j \neq i)$. Since the later added parameter has part of its correlated contribution accounted for by other already added parameters, the uncertainty contribution by the later added parameters will be underestimated. Thus, the variance contribution is dependent on the order when a specific parameter is entered into the orthogonalization in Bedford's method or the linear model used in the sequential sum of square decomposition.

Our proposed method decomposes the variance of the output into partial variances contributed by the variations of a parameter uncorrelated and correlated with all other parameters. For the total contribution by a parameter, the contribution is only based on regression with a single predictor. For the uncorrelated contribution, it is based on the regression over the estimated residuals. Thus, our method does not depend on the order of the parameters. Compared to Bedford's method or sequential sum of square decomposition, it is more consistent for comparison between studies.

For first-order uncertainty analysis (FOUA) [17], the model output can be approximated by

$$y \simeq f(x^{(0)}) + \sum_{i=1}^{K} (x_i - x_i^{(0)}) \frac{\partial y}{\partial x_i}, \tag{33}$$

where $x^{(0)} = (x_1^{(0)} \dots x_K^{(0)})$ are the central parameter values. Thus, the local variance/variance decomposition equation can be obtained by

$$\mathrm{var}(y) = \sum_{i=1}^{K} \sum_{j=1}^{K} \left(\mathrm{cov}(x_i, x_j) \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right). \tag{34}$$

If we use the fitted regression coefficients in place of the respective local sensitivity terms, $\partial y / \partial x_i$, we get

$$\mathrm{var}(y) = \sum_{i=1}^{K} \sum_{j=1}^{K} \hat{\beta}_i \hat{\beta}_j \mathrm{cov}(x_i, x_j). \tag{35}$$

We need to point out that the variance decomposition in Eq. (35) is different from our regression-based method. The correlated contribution in FOUA refers to contribution by the correlation. However, in our method, the correlated contribution refers to contributions by correlated variations. Consequently, in FOUA, the correlated contribution for x_i and x_j only includes the partial variance of $\hat{\beta}_i \hat{\beta}_j \mathrm{cov}(x_i, x_j)$. However, in our method, the correlated contribution includes both $\hat{\beta}_i \hat{\beta}_j \mathrm{cov}(x_i, x_j)$ and parts of $\hat{\beta}_i^2 \mathrm{var}(x_i)$ and $\hat{\beta}_j^2 \mathrm{var}(x_j)$ due to the correlated variations between x_i and x_j. Similarly, in FOUA, the uncorrelated contribution for x_i is $\hat{\beta}_i^2 \mathrm{var}(x_i)$. However, if there is correlation between x_i and x_j, part of the variability in x_i may come from a common latent factor representing an unknown mechanism. Thus, in our method, only part of $\hat{\beta}_i^2 \mathrm{var}(x_i)$ contributes to the uncorrelated contribution.

4.3. Non-linearity

The proposed regression-based method for uncertainty and sensitivity analysis relies on the assumption that the parameter effects are linear and the linear regression based

on Eq. (1) is a valid approximation of the model. A measure of goodness of the approximation is the multiple correlation coefficient. However, because Eq. (1) does not take into consideration the interactions among parameters, it is difficult to tell from the multiple correlation coefficient when the linear regression is valid. Based on the results from test case three, we would recommend the regression-based method when the multiple correlation coefficient is larger than 0.5. Generally, in the case when the ranges of the parameters are narrow, we would expect a linear relationship between parameter and model output and the linear regression to be a reasonable approximation. However, if the ranges of the parameters are large, the non-linear relationship may be strong and the goodness of fit of the linear equation should be tested. For example, for test case two, increasing the range for x_3 would result in a poorer performance of the regression-based analysis.

In the case when the regression has a very low multiple correlation coefficient, non-linear transformation may be used to improve the multiple correlation coefficient. For response with a power form, the Box–Cox transformation may be suitable [41]. For monotonic response, the rank transform may improve the regression [48]. However, during the transformation, the variance decomposition relationship may be changed. There is a tendency that a parameter with a high contribution will have much higher contribution, and a parameter with low contribution will have a much lower contribution after the transformation [49]. For example, if we apply the log-transformation for test case two, we can see that the total contribution of x_1 based on regression-based method is overestimated compared to the results from the correlation ratio method (Fig. 8), while the total contribution of x_2 is underestimated. So we should be very careful when making non-linear transformations.

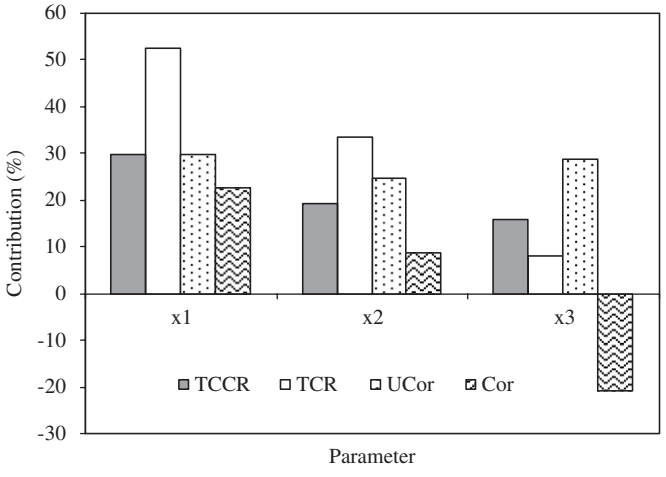

Fig. 8. Variance contribution by the regression-based method with log-transformed output for test case two. TCCR is the total contribution calculated by correlation ratio method on the non-transformed data for reference; TCR is the total contribution calculated by regression-based method based on the log-transformed output; Ucor is the uncorrelated contribution by regression-based method on the log-transformed output; and Cor is the correlated contribution by regression-based method on the log-transformed output.

Another alternative in the case of non-linear effects is the correlation ratio method. The correlation ratio method is always valid since it is a non-parametric method and is valid even where there are strong non-linear effects [31,33]. However, it requires much more computer runs than the regression-based method since it needs a replicated Latin hypercube sample. In addition, in order to adjust for bias, the correlation method needs thousands of replications for models with parameters having very small contributions (e.g. less than 0.1%). This can be impractical, even for simple models.

In real uncertainty and sensitivity analysis, we would recommend using a scatter matrix to check if there are strong non-linear effects based on a small Latin hypercube sample. If strong non-linear effects are present, the correlation ratio method is preferred. Otherwise, the regression-based method is preferred.

Acknowledgements

United States Department of Agriculture McIntire-Stennis funds were used to support this study. We thank the three anonymous reviewers for their very helpful comments.

References

[1] Iman RL, Johnson ME, Schroeder TA. Assessing hurricane effects. Part 1. Sensitivity analysis. Reliab Eng Syst Saf 2002;78(2):131–45.
[2] Iman RL, Johnson ME, Schroeder TA. Assessing hurricane effects. Part 2. Uncertainty analysis. Reliab Eng Syst Saf 2002;78(2):147–55.
[3] Helton JC, Davis FJ, Johnson JD. A comparison of uncertainty and sensitivity analysis results obtained with random and Latin hypercube sampling. Reliab Eng Syst Saf 2005;89(3):305–30.
[4] Borgonovo E, Apostolakis GE, Tarantola S, Saltelli A. Comparison of global sensitivity analysis techniques and importance measures in PSA. Reliab Eng Syst Saf 2003;79:175–85.
[5] Iman RL, Helton JC. An investigation of uncertainty and sensitivity analysis techniques for computer models. Risk Anal 1988;8(1):71–90.
[6] Helton JC. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. Reliab Eng Syst Saf 1993;42(2–3):327–67.
[7] Saltelli A, Ratto M, Tarantola S, Campolongo F. Sensitivity analysis for chemical models. Chem Rev 2005;105(7):2811–26.
[8] Saltelli A, Chan K, Scott M. Sensitivity analysis. Probability and statistics series. West Sussex: Wiley; 2000.
[9] Saltelli A, Marivoet J. Non-parametric statistics in sensitivity analysis for model output: a comparison of selected techniques. Reliab Eng Syst Saf 1990;28(2):229–53.
[10] Helton JC, Davis FJ. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliab Eng Syst Saf 2003;81(1):23–69.
[11] Helton JC, Johnson JD, Sallaberry CJ, Storlie CB. Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliab Eng Syst Saf 2006;91(10–11):1175–209.
[12] Cacuci DG, Ionescu-Bujor M. A comparative review of sensitivity and uncertainty analysis of large-scale systems-II: statistical methods. Nucl Sci Eng 2004;147(3):204–17.
[13] Frey HC, Patil SR. Identification and review of sensitivity analysis methods. Risk Anal 2002;22(3):553–78.

[14] Ionescu-Bujor M, Cacuci DG. A comparative review of sensitivity and uncertainty analysis of large-scale systems-I: deterministic methods. Nucl Sci Eng 2004;147(3):189–2003.
[15] Kleijnen JPC, Helton JC. Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: review and comparison of techniques. Reliab Eng Syst Saf 1999;65(2):147–85.
[16] Helton JC, Davis FJ. Illustration of sampling-based methods for uncertainty and sensitivity analysis. Risk Anal 2002;22(3):591–622.
[17] Helton JC. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. Reliab Eng Syst Saf 1993;42(2–3):327–67.
[18] McRae GJ, Tilden JW, Seinfeld JH. Global sensitivity analysis-a computational implementation of the Fourier amplitude sensitivity test (FAST). Comput Chem Eng 1982;6(1):15–25.
[19] Koda M, McRae GJ, Seinfeld JH. Automatic sensitivity analysis of kinetic mechanisms. Int J Chem Kinet 1979;11(4):427–44.
[20] Cukier RI, Schaibly JH, Shuler KE. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. III. Analysis of the approximations. J Chem Phys 1975;63(3):1140–9.
[21] Cukier RI, Levine HB, Shuler KE. Nonlinear sensitivity analysis of multiparameter model systems. J Computat Phys 1978;26(1):1–42.
[22] Cukier RI, Levine HB, Shuler KE. Nonlinear sensitivity analysis of multiparameter model systems. J Phys Chem 1977;81(25):2365–6.
[23] Schaibly JH, Shuler KE. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. II. Applications. J Chem Phys 1973;59(8):3879–88.
[24] Cukier RI, Fortuin CM, Shuler KE, Petschek AG, Schaibly JH. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I. Theory. J Chem Phys 1973;59(8):3873–8.
[25] Saltelli A, Andres TH, Homma T. Sensitivity analysis of model output: performance of the iterated fractional factorial design method. Comput Stat Data Anal 1995;20(4):387–407.
[26] Henderson-Sellers B, Henderson-Sellers A. Sensitivity evaluation of environmental models using fractional factorial experimentation. Ecol Modell 1996;86(2-3):291–5.
[27] Cryer SA, Havens PL. Regional sensitivity analysis using a fractional factorial method for the USDA model GLEAMS. Environ Modell Software 1999;14(6):613–24.
[28] Beres DL, Hawkins DM. Plackett–Burman techniques for sensitivity analysis of many-parametered models. Ecol Modell 2001;141(1–3): 171–83.
[29] Morris MD. Factorial sampling plans for preliminary computational experiments. Technometrics 1991;33(2):161–74.
[30] Sobol' IM. Sensitivity estimates for nonlinear mathematical models. Math Model Comput Exp 1993;1(4):407–14.
[31] McKay MD. Nonparametric variance-based methods of assessing uncertainty importance. Reliab Eng Syst Saf 1997;57(3):267–79.
[32] Fang S, Gertner GZ, Anderson A. Estimation of sensitivity coefficients of nonlinear model input parameters which have a multinormal distribution. Comput Phys Commun 2004;157(1):9–16.
[33] Saltelli A, Ratto M, Tarantola S. Model-free importance indicators for dependent input. In: Proceedings of SAMO 2001, third international symposium on sensitivity analysis of model output, Madrid. 2001.
[34] Bedford T. Sensitivity indices for (Tree)-dependent variables. In: Proceedings of the second international symposium on sensitivity analysis of model output, Venice, Italy. 1998.
[35] Helton JC, Johnson JD, Rollstin JA, Shiver AW, Sprung JL. Uncertainty and sensitivity analysis of chronic exposure results with the MACCS reactor accident consequence model. Reliab Eng Syst Saf 1995;50(2):137–77.
[36] Helton JC, Johnson JD, Shiver AW, Sprung JL. Uncertainty and sensitivity analysis of early exposure results with the MACCS reactor accident consequence model. Reliab Eng Syst Saf 1995;48(2): 91–127.
[37] Helton JC, Johnson JD, Rollstin JA, Shiver AW, Sprung JL. Uncertainty and sensitivity analysis of food pathway results with the MACCS reactor accident consequence model. Reliab Eng Syst Saf 1995;49(2):109–44.
[38] Iman RL, Shortencarier MJ, Jonson JD. A FORTRAN 77 and user's guide for calculation of partial correlation function and standardized coefficient. In: NUREG/CR-4122, SAND85-0044. Albuquerque, NM: Sandia National Laboratories; 1985.
[39] Saltelli A, Tarantola S, Campolongo F, Ratto M. Sensitivity analysis in practice: a guide to assessing scientific models. West Sussex: Wiley; 2004 (pp. 132–6).
[40] McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 1979;21:239–45.
[41] Weisberg S. Applied linear regression. 3rd ed. New York: Wiley; 2005. pp. 47–69.
[42] Iman RL, Conover WJ. A distribution-free approach to inducing rank correlation among input variables. Commun Stat Simul Comput 1982;11(3):311–34.
[43] Christensen R. Plane answers to complex questions: the theory of linear models. 3rd ed. New York: Springer; 2002. p. 51–2.
[44] Iman RL, Davenport JM. Rank correlation plots for use with correlated Input variables. Commun Stat Simul Comput 1982;11(3):335–60.
[45] Meadows DH, Meadows DL, Randers J. Beyond the limits. Post Mills, Vermont: Chelsea Green Publishing Company; 1992.
[46] Rencher AC. Methods of multivariate analysis. 2nd ed. New York: Wiley, Inc.; 2002. p. 408–50.
[47] Schabenberger O, Pierce FJ. Comtemporary statistical models. Boca Raton, FL, USA: CRC Press LLC; 2002. p. 98–101.
[48] Iman RL, Conover WJ. The use of the rank transform in regression. Technometrics 1979;21(4):499–509.
[49] Saltelli A, Sobol' IM. About the use of rank transformation in sensitivity analysis of model output. Reliab Eng Syst Saf 1997;50(3): 225–39.